MULTIMODAL EMG-EEG BIOSIGNAL FUSION IN UPPER-LIMB GESTURE CLASSIFICATION

Michael George Pritchard MEng MIET

Doctor of Philosophy

ASTON UNIVERSITY
March 2024

Abstract

Multimodal EMG-EEG Biosignal Fusion in Upper-Limb Gesture Classification
Michael George Pritchard MEng MIET
Doctor of Philosophy
2024

Upper-limb gesture identification is an important problem in the advancement of robotic prostheses. Prevailing research into classifying electromyographic (EMG) muscular data or electroencephalographic (EEG) brain data for this purpose is often limited in granularity of gestures classified, the extent to which generalisation is demonstrated, and methodological rigour.

This work proposes three architectures for multimodal fusion of EMG & EEG data in gesture classification, including techniques grounded in literature precedent and a novel "Hierarchical" strategy. Classification systems of these architectures are designed via Combined Algorithm Selection & Hyperparameter Optimisation (CASH) to ensure comparisons between the approaches are unbiased; likely this methodology's first application to the biosignal classification domain. All architectures are demonstrated suitable for use in a same-hand multi-gesture classification problem less separable than seen in much Brain-Computer-Interface research.

Fusion of EMG & EEG is shown to provide significantly higher ($p<0.05$) subject-independent classification accuracy (73.4%) than an equivalent single-mode EMG model, when tested on unseen individuals' data. Subject-independent single-mode EEG classification achieved accuracies (51.9%) competitive with those reached by many subject-specific systems in the literature on similar, or more separable, problems. The efficacy of CASH optimisation as a means of determining modelling choices — over inferring such decisions from literature — is also evidenced.

A desire to minimise the burden placed on potential prosthesis users motivates investigation of cross-subject and cross-session classification. Strategies for minimising per-session calibration, including through transfer learning, are explored. Results demonstrate that less session-specific data is needed to adapt a model pre-trained on an individual's previous-session data than would be needed to train a session-specific classifier to a similar accuracy (85%). Domain transfer using data collected from other individuals as the basis for adaptation is proven capable of accuracies (83%) nearing those of the subject-specific approach, laying the groundwork for future developments in low-calibration gesture classification systems.

Keywords:
Biosignal Fusion, Hybrid Brain Computer Interface (hBCI), Gesture Classification, Robotic Prostheses, Machine Learning, Multimodal Classification, Cross-Subject Learning, Inter-Session Calibration, Data Fusion, Hand Gesture Recognition

Acknowledgements

With thanks to my supervisors, without whose continual support, guidance, and friendship I would not have developed into the researcher I — at last — feel able to say I have become.

Thanks to my friends and colleagues at Aston Students' Union, who have given me a means to contribute to the community I call home, and to the Aston University Music Society & Choir, who have given me a space to escape from it.

Thanks to my mentor, and others who have supported in fighting for a more accessible PhD experience — and to the wider disability community, particularly my fellow disabled Postgraduate Researchers and those elsewhere in academia. The struggle for access and justice is not yet won, but it will be.

Finally, to my friends, family, and loved ones. I need not mention names, you know who you are. You have kept me sane, picked me up when things have gone wrong, celebrated with me when they have gone well, and been a safe haven in times of stress. Thank you; I love you.

Collaboration Acknowledgements

Experiments were run on the Aston University Engineering & Physical Sciences (EPS) Machine Learning Server, funded by the EPSRC Core Equipment Fund, Grant EP/V036106/1.

Appendix C presents extracts from the paper "Synthetic Biological Signals Machine-Generated by GPT-2 Improve the Classification of EEG and EMG Through Data Augmentation" [1], published February 2021 in *IEEE Robotics and Automation Letters* Volume 6, Issue 2, Pages 3498–3504 ©2021 IEEE.
This work was a collaboration between myself and Dr. Jordan J Bird (as co-first authors), and Prof. Aniko Ekárt & Drs. Antonio Fratini and Diego Faria who provided supervision and gave feedback on the manuscript prior to corrections and submission to IEEE Robotics and Automation Letters.
Dr. Bird's PhD research interests focused more significantly on Electroencephalography than Electromyography and the focus of his contribution to the paper was in this area. The EEG experiments were included in his PhD thesis and thus do not appear among the excerpts of the paper presented herein.
Select passages, marked in bold for convenience, discuss aspects of the work which applied to both the EEG and EMG experiments; the reader is advised that these hence appear in both Dr. Bird's thesis and in the extract in Appendix C.
Signed: Michael George Pritchard.
Signed: Jordan James Bird.

Ethical Review

The use of secondary data in this work has been reviewed by the Aston University College of Engineering and Physical Sciences Research Ethics Committee and given a favourable ethical opinion [ID EPS21031].

Content Warning

Anatomical diagrams are used in this work to describe the musculature of the human hand and forearm, and the structure of the human brain. The images used are detailed drawn illustrations, some colourised, however no photographic images of human dissection or medical operation are present in this work at any point. Nevertheless said illustrations may be uncomfortable to view for readers sensitive to such material. Readers are advised that the images in question appear in Figures [2.1, 2.2, 2.3, 4.2, 5.18, and 5.19] on pages [7, 8, 8, 40, 104, and 105] of the thesis.

[1] Notice included in accordance with "Frequently Asked Questions Regarding IEEE Permissions, IEEE, April 2013", accessed October 26th 2023, available: `https://web.archive.org/web/20231026161252/https://www.ieee.org/content/dam/ieee-org/ieee/web/org/pubs/permissions_faq.pdf` [archive copy]

Contents

Acronyms & Definitions

Term	Definition
Bioelectric Signal	*(also **biosignal**)* A measurable, time-variant electrical signal present in the human body such as EMG, EEG, ECG, and EOG.
Congenital Limb Difference	A limb difference present from birth, as opposed to one acquired through accident or surgical amputation.
Contralateral Hemisphere	The hemisphere of the brain laterally (left–right) opposite to a given movement or body part.
Electrocorticography	*(also occasionally **intracranial Electroencephalography**, iEEG)* The measurement of electrical activity in the human brain via electrodes surgically placed on its surface.
Electromyography	The process of measuring electrical activity in the skeletal muscles of the human body.
Surface Electromyography	The specific non-invasive form of electromyography in which measurement electrodes are placed on the skin rather than implanted.
(scalp) Electroencephalography	The process of measuring electrical activity in the human brain via electrodes placed on the scalp.
Sensorimotor Area	The region of the brain responsible for sensory and motor functions, primarily comprising the Motor Cortex and the Primary Somatosensory Cortex.

Table of Definitions

Acronym	Meaning
ARVIS	Aston Robotics, Vision, and Intelligent Systems Laboratory
ASL	American Sign Language
BCI	Brain-Computer Interface
CASH	Combined Algorithm Selection And Hyperparameter Optimization
CSP	Common Spatial Pattern
DOF	Degree(s) of Freedom
ECG	Electrocardiography *(also Electrocardiographic, etc)*
ECoG	Electrocorticography *(also Electrocorticographic, etc)*
EEG	Electroencephalography *(also Electroencephalographic, etc)*
(s)EMG	(surface) Electromyography *(also Electromyographic, etc)*
EOG	Electrooculography *(also Electrooculographic, etc)*
ERD	Event-Related Desynchronisation
ERP	Event-Related Potential
ErrP	Error-Related Potential
ERS	Event-Related Synchronisation
FD	Frequency Domain
FES	Functional Electrical Stimulation
FFT	Fast Fourier Transform
GAN	Generative Adversarial Network
GPT-2	Generative Pre-trained Transformer 2
HCI	Human-Computer Interaction
IMU	Inertial Measurement Unit
(K)MI	(Kinaesthetic) Motor Imagery
KNN	K- Nearest Neighbours
LDA	Linear Discriminant Analysis
LH(S)	Left Hand (Side)
MAV	Mean Absolute Value
ME	Motor Execution
MLP	Multilayer Perceptron
MUAP	Motor Unit Action Potential
(G)NB	(Gaussian) Naïve Bayes
PSD	Power Spectral Density
QDA	Quadratic Discriminant Analysis
RBF	Radial Basis Function
RF	Random Forest
RH(S)	Right Hand (Side)
RMS	Root Mean Square
sEMG	Surface Electromyography
SFS	Sequential Forward Selection
SVM	Support Vector Machine
TD	Time Domain
TMS	Transcranial Magnetic Stimulation

Table of Acronyms

Introduction

1.1 Context & Motivation

At least 1 in 190 — more than 1% — of the U.S. population are believed to have experienced the loss of a limb [2], a proportion believed similar worldwide. Much of society and the built environment is designed for an imagined "typical" individual and inaccessible to those falling outside this narrow definition, thus for many amputees (among other disabled people) day-to-day life is made more difficult. The term "Activities of Daily Living" (ADLs) is used to describe a number of tasks thought necessary for individuals to live independently [3]. Though used with unfortunate frequency used to imply a "deficit" on behalf of those disabled individuals who cannot perform them or an inherent lesser quality-of-life [4], a framing which this work categorically rejects, ADLs can be a convenient shorthand for the types of activities with which an amputee may struggle without appropriate support.

The choice of language here and throughout the work is deliberate. The term "amputee" is used in this work to refer specifically to those who have experienced limb loss, through surgical or traumatic amputation. Those born with a limb difference, a condition known as "amelia", are at times described as "congenital amputees". Such individuals tend not only to present physiological or neuroanatomical differences compared to those who have undergone amputation (they do not, for example, typically experience phantom limb sensations [5]), but are frequently less likely to desire use of prosthesis, having been accustomed to their limb difference since birth and often not conceiving of it as something in need of "fixing", instead considering themselves as a group with distinct needs and priorities. Further, some literature in contrasting amputees with non-amputees describes individuals of the latter group with terms such as "intact", "healthy", or even "normal". As a disabled individual, the author of this thesis is distinctly uncomfortable with such a portrayal implying — intentionally or otherwise — disabled people as somehow incomplete, unhealthy, or abnormal. Used instead are terms such as "non-amputee", "individual(s) without limb differences" (notwithstanding the aforementioned distinction between those with congenital and acquired limb differences), and "able-bodied"[1].

For many upper-limb amputees, support comes in the form of prosthetic limbs: accessibility devices worn by an amputee which replicate in some way the functionality of a biological limb. While the earliest functional artificial limbs date back centuries to the devices of sixteenth-century surgeon Ambroise Paré [6], it is only within the last hundred years that more dextrous control has become possible, with the advent

[1] The latter it should be stressed is meant solely as a shorthand for "not disabled in ways relevant to the specific topic discussed", and not intended to imply erasure of those who are multiply disabled or who, while not being amputees, would nonetheless not consider themselves "able-bodied".

of electrically powered prostheses. Dr. Reinhold Reiter in 1948 pioneered the use of naturally occurring electrical activity in the muscles known as Electromyography (EMG) as a means of prosthesis control, with the *Elektrokunsthand* [7]. Such myoelectric prostheses have become prevalent in the decades since, however despite advancements in their ergonomics [8] and aesthetics [9], the sophistication of their control mechanisms has remained largely static for some time.

1.1.1 Prevailing approaches

In the United Kingdom, the National Health Service has begun offering EMG-controlled prosthetic arms to amputees as standard, where previously such devices were available only to military veterans & other patients provided with cosmetic or mechanical limbs [10]. These are at least in part facilitated by OpenBionics, whose flagship Hero Arm is generally recognised as state-of-the-art among affordable, accessible, commercially-available robotic prostheses but is nevertheless limited. While the Hero Arm is capable of executing a range of different gestures and grip patterns [11], these are not available to an amputee simultaneously. Instead they are achieved through mode-cycling, with a number of individual operational modes available each restricted in dexterity to the opening and closing of a single pre-defined gesture through a Direct Control mechanism [11].

Direct Control refers to a common prosthesis control paradigm, described as the “clinical standard of care” [12], wherein the amplitude of an individual’s measured electromyographic signal acts as the control signal for actuation. Activity over a threshold by one muscle results in robotic actuation in one direction, such as opening a hand, and activity by an opposing muscle triggers the opposite actuation, such as closing it [13,14]. The speed of actuation is in some cases made a function of the level of muscle force applied, thus the scheme is sometimes alternatively named Proportional Control [12]. Pattern Recognition systems — those employing machine learning algorithms to classify between defined gestures from measured biological data, and actuating a prosthesis accordingly — have been the focus of extensive research but rarely seen deployment in the “real world”. This is despite much research indicating their suitability. Wurth & Hargrove [12] found a simultaneous multiclass pattern recognition system to provide better control in a cursor-steering task than both a “sequential” (mode-cycling) pattern recognition approach and a conventional proportional control scheme. Kuiken et al. [15] also found mode-switching systems to be inconvenient for users and that they would often alternate operational modes unintentionally [15]. Their participants however, who had lived experience of Direct Control prostheses, noted difficulty in learning to consistently perform multiple distinct movements as required by the Pattern Recognition approach [15]; research by Resnik et al. [16] has likewise found Direct Control to be preferred on some metrics, though the potential confounding factor of familiarity should again be noted.

While there is clearly further research to be done on gesture classification for prosthesis control applications, it should be acknowledged that this approach is not entirely without commercial precedent. Esper Bionics offer a robotic hand apparently capable of a wide range of gestures — to which users can add additional custom movement classes [17] — and purport an adaptive cloud-based gesture recognition algorithm [18], though provide only sparse technical details.

1.1.2 Why prostheses? A note on alternatives

There are, of course, alternatives to robotic prostheses. The need for more naturalistic control of replacement limbs is such that much research has been done into the transplanting of biological arms and hands from human donors to amputees: Dubernard et al. achieved the first bilateral arm transplant in 2000, with positive results [19]. The surgical approach is naturally not a sustainable solution at scale. Even disregarding the obvious challenges in sourcing suitable donor limbs, ongoing personalised treatment is needed post-surgery to manage the body's attempts to reject donor skin, and even with such immunotherapy further medical complications are common [20].

Some work has also been done on the development of bionic limbs — prostheses enabling direct sensory feedback by a two-way surgical interface [21]. Skeletal implants have allowed amputees to feel the presence, firmness, and force applied on an object [22]. Though not bidirectional, another surgical approach which has seen recent success involved grafting muscle fibers to peripheral nerves in an amputee's residual limb to amplify the natural bioelectric signals, enabling intuitive multi-gesture prosthesis control [23]. This latter example illustrates a key limitation of such advanced surgically-installed prostheses however: the implants required specialist adjustment within a year of operation. While an impressive duration among surgical approaches, this nevertheless presents an inconvenience to users beyond the already significant undertaking of the initial surgical implantation — and the time, cost, and specialist expertise required for their installation & maintenance naturally make such approaches difficult to deploy at scale.

1.2 Prosthesis User Needs

Among research on prostheses much is unfortunately said about, rather than by, the amputees using them. Data on the needs and opinions of prosthesis users is frankly sparse. Biddiss et al. in reviewing literature of the late 20th century found that mechanical (or "body-powered") arms were rejected or abandoned by amputees at a higher rate than EMG-based ones, by 26% to 23% in adults but by a much wider margin (and at higher rejection rates overall) among children of 45% to 35% [24]. Their later primary research likewise found body-powered arms to be rejected by many more (50%) of their respondents than myoelectric (39%). This survey also illuminated desirable properties in prosthesis design, with reductions in weight and cost routinely among the highest priorities. Improvements in dexterity were a key priority among users of EMG prostheses [25]; though as discussed above, such individuals would be most likely to have experienced Direct Control systems than Pattern Recognition. Kyberd et al. found that among frequent users of upper-limb prostheses, the type used (between cosmetic, body-powered, and EMG) was not correlated with amputees' experiences of encountering issues with their devices, but rather with the level of dexterity required for their work. The level of limb loss also had a significant effect — those with below-elbow ("transradial") amputations were more likely to use myoelectric prosthesis than above-elbow ("transhumeral") amputees [26]. This is most likely due to the former group's likelihood of retaining some control over residual forearm muscles key to dexterous hand movements, as outlined in Chapter 2.

Engdahl et al. explored the opinions of upper-limb amputees — both those who used prostheses and those who didn't — regarding the importance of certain levels of functionality and the interest in potential control mechanisms [27]. EMG systems appealed to 83% of respondents, but surgical approaches saw much less interest, at approximately 65% for strategies involving muscle surgery and 39% for the use of a brain implant. Perhaps unsurprising given the extensive surgery amputees will have already undergone, qualitative responses regularly indicated surgery as a significant deterrent. Even advanced levels of hypothetical functionality had only a slight impact on systems' attractiveness; for example those not drawn by a cortical implant which would provide only basic hand movements remained largely uninterested in one capable of enabling tactile sensation. In fact, the higher levels of functionality were not perceived as being of great importance; an ability to perform basic grasps was regarded most important. Cordella et al. likewise found grasp capabilities a particular need of myoelectric prosthesis users. Their review defined the most important gesture functionalities as the lateral, pinch, hook, spherical, cylindrical, and centralised grips (all noted as key to performing ADLs), along with both a flattened hand and a more natural "neutral" pose [28]. It is worth noting though the common focus on "necessity" and "required" functionality among such work — Engdahl et al.'s survey for example discussed *importance* not *desirability*. It is not hard to imagine that asking the latter question may offer different results & provide for an interesting comparison.

Researchers' understanding of prosthesis users' experience evidently needs to be developed further in many areas. For example, errors in a multiclass system can present in many different ways; a datapoint belonging to a given class could be misclassified as any one of the remaining classes. Misclassifications could thus result in a number of different types of incorrect actuation, but it stands to reason they would not all be of equal severity. Rather, they could plausibly be assumed context- and task- dependent: a prosthetic hand erroneously releasing a cup of hot liquid would likely be more problematic for a user than one which sporadically moved while intended to be at rest. Unfortunately this topic is under-reported by the literature; such specificity does not seem to have been a particular focus of any of the sparse research on the needs and desires of prosthesis users. In the absence of such data, examples more nuanced than this extreme illustrative case cannot be reasonably guessed by those without lived experience of prosthesis use. As with so many fields of study, especially those involving marginalised communities, there clearly remains much to be done in furthering direct user-involvement in prosthesis research. To enable systems to be developed better suited to prosthesis users' needs, it is paramount that their views on acceptable rates and categories of errors be centred in future work in the domain. Bringing patients into the design loop could have many other beneficial impacts. Patient-first design can result in a lower likelihood of needing to recall products and a greater chance of their being successful (both in terms of patient outcomes and commercially) [29]. Of course, patients are not unanimous and to do such user-centred design justice a sizeable enough dataset would need be recruited — this is ultimately the central reason it is considered out-of-scope here. It is wholly acknowledged that such steps would be necessary for the findings of this research to be translated to the "real world", and no implication to the contrary is intended.

1.3 Research Foci & Contributions

It is clear from the discussion above that further research is needed on the development of suitable control strategies for prosthetic limbs — that is, on suitable methods for using biological signals in classification of gestures. Of course, as revealed by studies on prosthesis users' needs this is not *all* that is needed to advance the field, and is not even necessarily viewed as the singular most pressing priority. It is not the intention of this work to suggest its findings are the only or the most urgent improvement to be made in the field of upper-limb prostheses; as Kyberd et al. state: "*Every part of the process of fitting a prosthesis can be improved, which will have an effect for some of the population who use their devices regularly. There is, however, no single factor that would bring greater improvement to all users*" [26]. Much important work is being done in the domain by both academia and industry on comfort, usability, cost, and various other factors. Nevertheless, gesture classification is a vitally important component in the progression towards higher-quality robotic upper-limb prostheses, and is where this research contributes.

In particular, research on amputees' needs highlights the need for greater distinction between grasping hand gestures, while both overcoming the limitations of conventional myoelectric approaches but avoiding the need for surgically-implanted sensors. This thesis hence explores the simultaneous use of electromyography with electroencephalography, a non-invasive technique for measuring brain activity, for classification of multiple types of right-handed grasp motion (as outlined in Chapter 2, much can be gleaned regarding an individual's intended movements from the electrical signals in the brain's motor cortex).

The experiments in Chapter 5 form the core of this research, proposing and evaluating three architectures for such multimodal classification: "early" fusion at the feature-level, "late" fusion at the decision-level, and a novel "hierarchical" strategy which synthesises the two. Its rigour in minimising the risks of bias and data leakage through careful application of techniques such as Combined Algorithm Selection & Hyperparameter optimisation to this problem set it apart from many works in the domain [30, 31], as discussed in Chapter 3. The fusion architectures' viability is demonstrated and validated by their generalisation to data of novel subjects.

The barriers of cost and convenience faced by prosthesis users then motivate the investigations of Chapters 6 and 7. The former assesses the extent to which EMG & EEG data gathered from a wider population could supplement classification models tailored to individual subjects, and whether such cross-subject learning could lessen the data collection burden on new users of a system. Chapter 7 explores similar principles in considering a "real world"-aligned paradigm of gesture classification on the basis of data local to individual recording sessions. In finding strategies to reduce the level of necessary calibration, it paves the way for future work to develop multimodal gesture recognition systems more robust to changes in data distributions, and thus eventually to more reliable and dexterous multi-gesture robotic prostheses.

Background

In its exploration of biosignal-based classifiers a number of biological and anatomical concepts are discussed throughout this thesis. This chapter introduces some of the key relevant concepts and phenomena. Additionally an introduction is given to some of the techniques and considerations in the measurement of bioelectric signals and their application to gesture recognition systems. This is intended as a high-level overview of the topics at hand and is simplified in areas; a deep understanding of the underlying biological mechanisms at play in voluntary motor control is not strictly necessary to follow the work.

2.1 Biological Principles

2.1.1 Model of voluntary upper-extremity movement

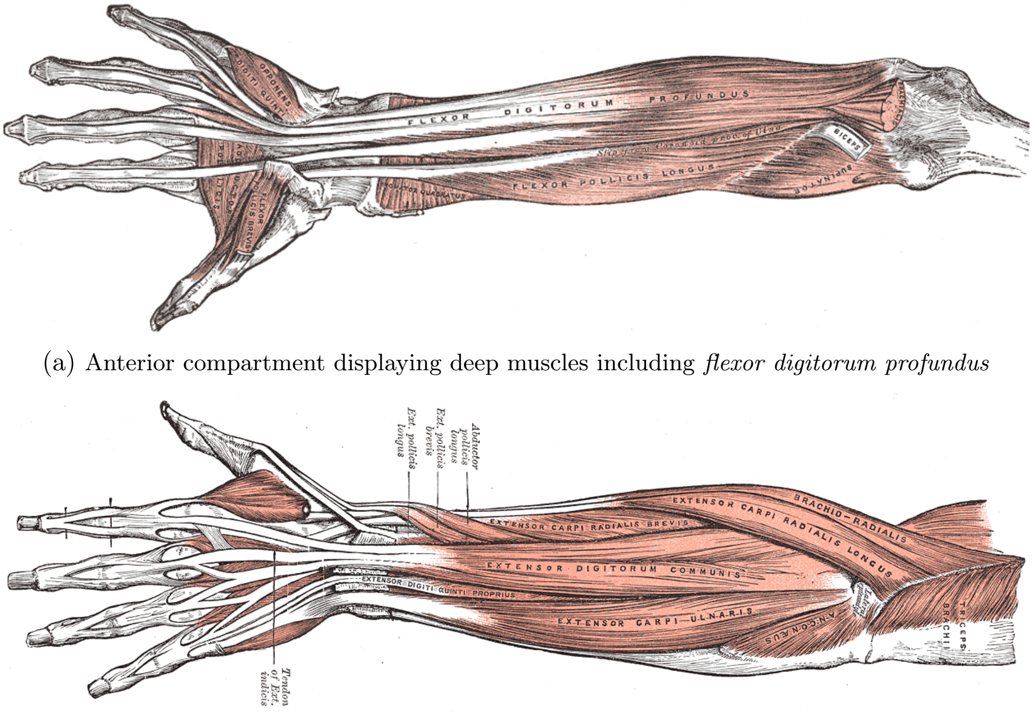
Voluntary movement of skeletal muscles in the body, or "motor activity", is initiated in the brain. Cells known as "motor neurons" are comprised of a cell body (also called the *perikaryon* or *soma*), situated in the brain, and an *axon*, which connects them via the spinal cord to the muscle fibres they control[1]; the neuron and the muscle fibres it innovates are together referred to as a Motor Unit. When an organism moves, an electrical impulse is transmitted by the relevant motor neurons, travelling down their axons to reach the innervated muscle fibres and actuating them. Multiple motor units moving together can cause a muscle to contract or expand, and thus move a connected body part.

The fingers of the human hand are not wholly actuated by muscles located within the hand or fingers themselves. Rather, the *phalanx* finger bones are connected via long tendons to a number of muscles housed largely in the forearm. A movement of a hinge joint such as those in the fingers typically involves the contraction of one muscle or group thereof, which by shortening "pulls" the joint to the desired position, and the expansion of an opposing muscle whose controlled lengthening stabilises the movement and allows the new position to be maintained. These two muscles are described as the "agonist-antagonist pair" for a given movement. In a curling of the index finger for example the *flexor digitorum profundus*, located in the anterior compartment of the forearm (the "underside"[2]), acts as the agonist, while the *extensor digitorum*

[1] This is a simplified model for ease of comprehension; an understanding of the particulars of synaptic transmission pathways is not required for this work.

[2] "Anterior" strictly means "towards the front", and "posterior" the opposite. By convention, in the context of the human body these terms are relative to the "standard anatomical position". This resembles a typical standing posture but with the palms of the hands turned to face *forwards*. The underside of the forearm is hence considered "anterior" in anatomical terms, regardless of the position at which an individual's arm is held.

communis in the posterior compartment (the "top" side) is the antagonist [32]. The arrangement of these forearm muscles can be seen in Figures 2.1a and 2.1b.



(a) Anterior compartment displaying deep muscles including *flexor digitorum profundus*

(b) Posterior surface displaying superficial (surface) muscles including *extensor digitorum communis*

Figure 2.1: Anatomical diagrams of musculature of the left forearm. Public domain work as taken from Henry Gray's *Anatomy of the Human Body* [32].

One resulting property of this anatomical structure is that an absence of the fingers or hand does not necessarily imply an absence of the muscles which would typically control them. That is, following a traumatic or surgical amputation of the hand, an individual may retain some ability to voluntarily control those muscles, dependent on various factors including the degree of forearm remaining post-amputation; from such muscle movements it can be possible to intuit the intended movement of the absent hand.

2.1.2 Neural Physiology

The *cerebrum* is the largest part of the human brain and is broadly divided into a number of geographical regions, called lobes, which perform different functions and serve different purposes. The frontal lobe is the section of the brain responsible for most voluntary activity including movement, decision-making, planning, executive and higher cognitive functions. Physical activity in particular is controlled in the main by the Primary Motor Cortex, where the aforementioned Motor Neurons are located. This is situated at the very posterior of the lobe immediately in front of the central sulcus (the large transverse groove separating the anterior and posterior sections of the brain), as depicted in Figure 2.2.

On the other side of the central sulcus is the parietal lobe, largely responsible for measuring and in-

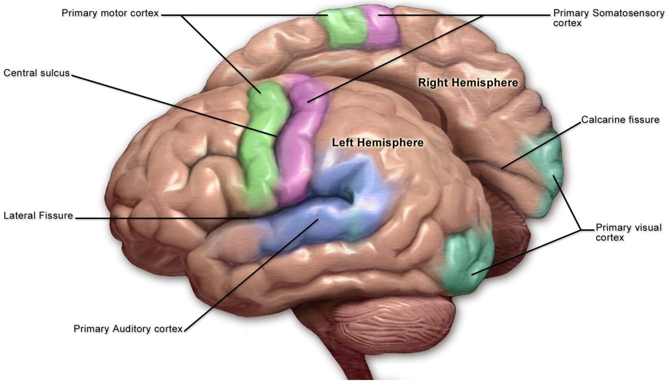


Figure 2.2: Location of the Motor and Somatosensory Cortices, highlighted green and purple respectively [33], used under CC-BY 3.0 [`https://creativecommons.org/licenses/by/3.0/deed.en`]

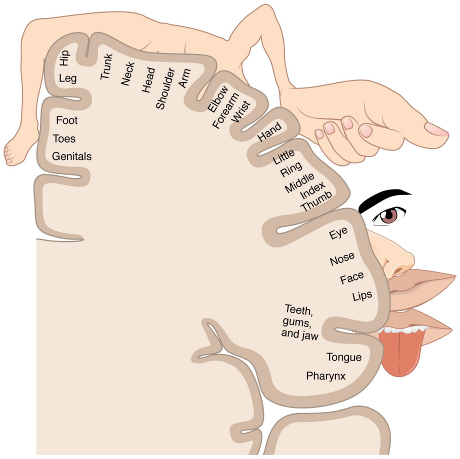


Figure 2.3: Cross-sectional view of the Motor Cortex, with "cortical homunculus" illustration of its somatotopic mapping [34], used under CC-BY 4.0 [`https://creativecommons.org/licenses/by/4.0/`]

terpreting sensory information. At the very anterior of the parietal lobe lies the primary somatosensory cortex. This cortex monitors tactile information, i.e. the sense of touch, and runs parallel to the motor cortex as seen in in Figure 2.2; the two cortices together are referred to as the brain's "sensorimotor area". The cerebrum is divided longitudinally by the longitudinal fissure into distinct left and right hemispheres, with certain brain functions being lateralised to one hemisphere or the other. Many functions related to the body, including motor control, are typically thought of as "contralateral"; the left hemisphere controls and processes information regarding the right side of the body, and vice versa.

The arrangement of the motor cortex itself is heavily somatotopic; distinct locations on the cortex are associated with the control of different parts of the body as seen in Figure 2.3. The somatosensory cortex is similarly somatotopically arranged; the brain location responsible for moving a particular body part is hence typically in close physical proximity to the location responsible for sensing it.

2.2 Bioelectric Activity

2.2.1 Muscular Signals

2.2.1.1 The nature of muscular bioelectrical activity

The electrical impulse in a Motor Unit when it is actuated is known as a Motor Unit Action Potential (MUAP) [35]. Such a signal is measurable and has a distinct characteristic pattern formed from the superposition of individual Action Potentials[3] in each of the muscle fibres in the Motor Unit, as illustrated in Figure 2.4. Identifying every MUAP in an organism could thus provide a complete picture of their muscles' movements.

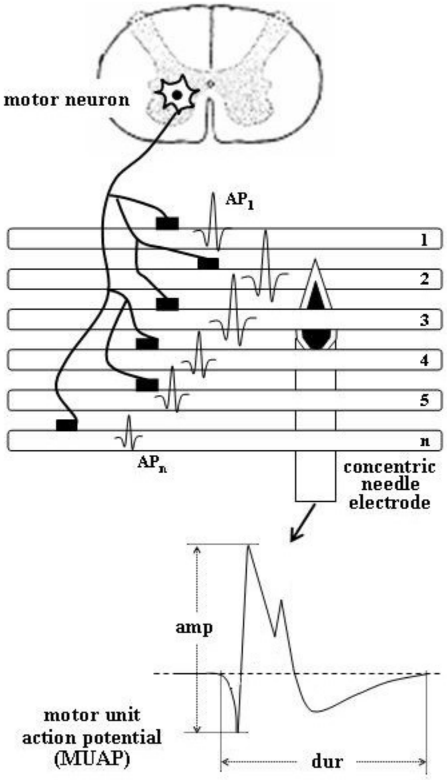

motor neuron
AP₁
1
2
3
4
5
n
APₙ
concentric needle electrode
amp
motor unit action potential (MUAP)
dur

Figure 2.4: Illustration of Action Potentials in n muscle fibres of a single Motor Unit, and the Motor Unit Action Potential observed by their measurement. ©2012 Ignacio Rodríguez-Carreño, Luis Gila-Useros, and Armando Malanda-Trigueros. Adapted from [36]; originally published under CC-BY 3.0 license. Available from: doi.org/10.5772/50265.

Measuring individual MUAPs in isolation however is challenging, as the highly localised nature of the signal requires a sensor of very high spatial fidelity or the use of sophisticated data processing techniques to reverse-engineer them from coarser data. For many applications though it is ultimately unnecessary. While the properties of a MUAP can indeed be of medical relevance — it has for example been characterised as distinct in individuals who have experience a stroke [35], and found to be affected differently by neuropathic & myopathic conditions [37] — for characterising or identifying movements, the activity of a whole muscle

[3] The word "potential" here refers to its meaning in electronics, as in "electric potential difference" commonly referred to by its unit as "voltage".

is of greater interest than any particular clusters of the fibres constituting it, and thus a less spatially granular observation can suffice. Such a measurement captures a superposition of MUAPs from a number of geographically local Motor Units, forming what is commonly referred to as an Electromyographic (EMG) signal [38, 39]. EMG data, though inherently coarser than MUAP-level measurements, can provide rich information regarding the level, timings, and characteristics of physical movement. By assessing the properties of the measured electrical signals, the nature and extent of a muscle's contraction or expansion can be estimated. At the simplest level, the amplitude of EMG is indicative of the degree of muscle activity — a movement of greater force results in recruitment of more motor units, thus more MUAPs are present to superpose into a stronger signal. The degree of activity in different muscles, such as the agonist-antagonist pair at a joint, at a given point in time can hence be predicted and the joint's movement inferred.

2.2.1.2 The measurement of muscular bioelectrical activity

Electromyography of a high spatial resolution can be carried out by the use of intramuscular "needle" electrodes surgically implanted into the body [40,41], referred to in some cases as Implantable Myoelectric Sensors (IMES) [42] but more generally as Implanted or Intramuscular EMG (iEMG). This approach can give very high-fidelity data, and can indeed enable decomposition of the recorded signal into individual MUAPs, differentiable by Motor Unit according to specific characteristics of each MUAP [37]. An invasive approach such as this however carries significant challenges which preclude it from seeing widespread use. The burden on a user is high; although the medical risk is minimal the likelihood for discomfort and the general inconvenience of further surgical procedures may well be too great to be considered worthwhile. This is particularly relevant considering amputees will have already undergone extensive surgery; Engdahl et al.'s 2015 survey of upper-limb amputees found approximately 40% of respondents neither "likely" nor "very likely" to consider any of three surgical prosthesis control techniques, even if they enabled fine motor control and the sensation of touch in the limb [27]. Even notwithstanding users' aversion to the technique, the process of inserting electrodes naturally requires the expertise and time of a trained specialist, and further support may be required for postoperative care, presenting financial barriers. Needle electrodes are also not always designed to be used for more than a single session before removal, and those that are may not be suitable for prolonged use; Waris et al. in 2018 found iEMG electrodes which remained in participants' muscles to notably degrade in performance over the course of just a week, and that this degradation was particularly significant for those amputees with a smaller residual limb [43].

Due to such obstacles the more common approach is Surface Electromyography (sEMG[4]), wherein the muscular signals are instead measured by electrodes placed on the surface of the skin. Such electrodes can take the form of single-use adhesive pads containing a conductive electrolyte gel such as in the commonplace "Ag/AgCl" silver chloride electrodes [44], dry metal electrodes for greater convenience such as those stainless steel electrodes used in the now-discontinued Thalmic Labs "Myo" armband [45, 46] which has seen much use in research [47–50] and in prosthesis applications [51], and even conductive textiles designed for low-

[4] Referred to frequently as just "EMG" throughout the work by convention, other than where specifically discussing the contrast between sEMG and other types of EMG such as in this section.

cost applications [52]. While the benefits of sEMG as a noninvasive method are self-evident, it also carries challenges by comparison to the iEMG approach. Various factors can degrade the quality of the measured signal, due primarily to the fact that the measurement electrodes are significantly further from the generating source (the muscles) of the signals they are attempting to record. That gap is filled nonuniformly with different biological materials such as bone, fatty tissue, other muscles, and the skin itself — all of which have varying electrical properties and can attenuate and add noise to an EMG signal. Additionally, activity in muscles located near to the target muscle an electrode seeks to measure, including those situated *between* the electrode and the target muscle, can be picked up in a recorded signal as electrical interference or "crosstalk".

The proper placement of EMG electrodes is thus important to obtaining high quality EMG data. The prevailing recommendation as outlined by Zipp in 1982 is for electrodes to be placed in most cases over the thickest part of a muscle (its "muscle belly") [53]. Recent work such as that of Takala & Toivonen [54] has reinforced the importance of these standards with regard to the forearm muscles described in 2.1.1, and extended them to recommend a "through-forearm setting" wherein electrodes are placed over both the agonist and antagonist muscles of a given joint to be measured simultaneously.

The intensity of a muscular contraction also affects the nature of a measured EMG signal. In broad terms the degree of muscle force applied during a movement has a linear correlation with EMG amplitude (though this simplification does not account for the unique properties of different muscles nor that a given joint movement may involve actuation of a number of muscles simultaneously [55]). This presents a challenge to gesture-recognition systems relying solely on Electromyography, as one cannot guarantee that a user will perform a motion with the same degree of muscular force in each attempt. Those developing such systems, or indeed the individuals contributing muscular data to them, may plausibly be inclined to perform gestures at or near their maximum exertion (a "Maximal Voluntary Contraction") under the naïve belief this would best represent an "ideal"representation of the gesture. However a classifier trained on gestures performed at a given level of force is not always able to recognise stronger or weaker variants of those same gestures. Such variation can pose challenges in the design and training of EMG-based systems, particularly considering that the level of muscle force applied is itself frequently used to determine the actuation speed for a device such as a robotic prosthesis [56] (as in the Proportional Control approach mentioned in 1.1.1).

Another difficulty with EMG-based systems is muscular co-contraction. As discussed above, a typical movement of a joint involves the contraction of the agonist muscle and the simultaneous extension of its antagonist. Co-contraction however occurs when both the agonist and antagonist muscles of a given joint contract at the same time. This is not inherently a problematic phenomenon and in fact frequently occurs in natural movements such as to freeze or lock a joint, or to regulate its movement for stability by stiffening a limb under strain, among other functions.In amputees however, the newly arranged biomechanics following surgery can cause changes in co-contraction patterns or even involuntary co-contractions of muscles. Seyedali et al. [57] found that among trans-tibial (below-knee) amputees, co-contraction was greater in their residual (amputated) limb than the intact limb while walking using passive (unpowered) prosthetic feet. Interestingly this applied to both the knee joint which was above the amputation site, and the ankle musculature despite the amputation of the foot & ankle itself[5]. They also noted that the degree of co-contraction may vary with a movement's speed or force, posing further challenges to systems reliant solely on electromyographic data.

2.2.2 Neural Signals

2.2.2.1 Cortical activity and motor-control relevant phenomena

Electrical activity in the brain is typically oscillatory in nature, driven either by repeated patterns of activity within individual neuronal cells, or through the transmission of electrical signals *between* networks of cells in a "loop" connecting different parts of the brain, such as the thalamocortical connections between the cerebral cortex and the thalamus [58] or the cerebello-thalamo-cortical network formed by the motor & premotor cortices, the thalamus, and the cerebellum [59, 60], with neighbouring neurons' osicllations superposing into a measureable electrical signal in either case.

These oscillations can occur at varying frequencies; a measured neural oscillation, or "brainwave", will contain a range of frequency components. Much as different physical locations of the brain are responsible for different functions, neural signals at different frequencies are variously associated with different activity. Measured brainwaves are commonly discretised into a number of frequency bands, with their relative bandpowers & the dominance of any given component taken as an indicator of the type of neural activity occurring. The Alpha wave for example is defined as approximately 8 – 12 hertz[6]; a high level of activity in the Alpha region (colloquially described as "high Alpha waves") often presents most strongly when an individual's eyes are closed and can indicate a state of wakeful relaxation[7]. Table 2.1 describes the five brainwaves as they are commonly framed in discussion of neural activity, their approximate corresponding frequency bands, and indicative (though non-exhaustive) examples of activity associated with high relative bandpowers in each.

Brainwave	Frequency Band	Associated Activities
Delta	0.5 - 4 Hz	Deep sleep [66, 67], learning learning [68], decision-making [69]
Theta	4 - 8 Hz	Memory retrieval, Spatial navigation [61]
Alpha	8 - 12 Hz	Wakeful relaxation (especially with closed eyes), some sleep stages [63]
Beta	12 - 35 Hz	Waking consciousness, continuance of current cognitive or motor activity [70]
Gamma	35 - 150 Hz	Various including Sensory Perception & Processing [71]

Table 2.1: Brainwave bands and their implications

[5] Akin to the finger-joint musculature being primarily located in the forearm as described in 2.1.1, many ankle-controlling muscles are situated in the lower leg & so would remain, in varying degrees, following trans-tibial amputation.

[6] Whilst there is much scientific consensus as to the approximate frequency band of each brainwave, there is inconsistency among literature regarding the specific cut-off points between each band. The Alpha wave for example is variously reported as 8-13 Hz [61], 8-12.25 Hz [62], and 8-12 Hz [63], among other definitions. It has been suggested [64] that the most discriminative brainwave cutoff frequencies may vary between individuals, and may even not be wholly consistent within any given individual.

[7] Understanding of the Alpha, and indeed all brainwaves, has and continues to evolve over time and is far from complete. High Alpha waves were previously thought to indicate an idle, nonspecific brain state but are since understood to present during a range of neural activities [65]. A deep knowledge of all possible functions of the various brainwaves is not required to understand this work as a whole & more specific details of relevance will be introduced where they are discussed.

In the context of motor activity however another signal known as the Mu (or μ) rhythm is often of particular interest. While the body is at rest, individual neurons within the motor cortex generate electrical impulses, or “fire”, synchronously. They do so approximately 10–12 times a second, i.e. they take the form of a 10–12Hz oscillation — though as with other brainwaves the exact frequency range is debated, and has been suggested to present also in the Beta band [64] or even to manifest nearer 20Hz in the motor area and 10Hz in the somatosensory [72, 73]. These oscillatory electrical impulses, regardless of their precise frequency bounds, act as point sources originating in each neuron, such that they superpose into a measurable signal in the order of microvolts: the Mu wave.

When an individual performs a physical movement, the motor neurons related to the Motor Units required to elicit that movement fire in accordance with the intended timings of the movement, rather than at their natural resting rate. This results in fewer neurons oscillating synchronously, and thus the superposed signal observably attenuates, a phenomenon known as Mu desynchronisation, or Mu suppression. When performing a motor action with one side of the body this desynchronisation presents contralaterally, and is accompanied by a event-related synchronisation (i.e. an increase in amplitude of the μ) in the corresponding motor neurons in the ipsilateral hemisphere. Considering the aforementioned somatotopic nature of the motor cortex (Figure 2.3), the location of such Mu desynchronisation can be taken as indicative of which specific motor unit(s)' associated motor neurons are active, and thus the nature of the resultant muscular movement. Such precise spatial localisation of the Mu's attenuation is difficult without invasive measures as discussed below, but the differing patterns of desynchronisation elicited by different movements will nevertheless naturally result in different effects on the measurable superposed Mu wave. Though complex and potentially subtle in nature, machine learning can be employed to distinguish these and thus estimate an individual's physical movements from their measured neural activity.

Crucially to many Brain-Computer-Interface applications, the process of imagining the physical sensations of moving, Kinaesthetic Motor Imagery (KMI), elicits patterns of electrical activity in the brain which are remarkably similar to those induced when a genuine movement (Motor Execution) is made [64], and the extent of this similarity is correlated with the subjective vividness of the imagery [74]. Though KMI-induced electrical activity can be more difficult to accurately classify than that arising from ME [75], this means that even where an individual has undergone amputation of a body part, through measuring neural activity while KMI is performed, their *intended* movement of that body part may still be identifiable. KMI has additionally been shown to elicit measurable electromyographic activity in the muscles corresponding to the imagined movement. This, while much weaker than the EMG signals that would be present during Motor Execution, was similar in nature — and was found to be correlated in magnitude to the imagined physical effort of the movement task, akin to the aforementioned relationship between EMG magnitude and muscle force during genuine physical movement [76].

2.2.2.2 The measurement of brain activity

Electrical activity of individual motor neurons can be measured through intracortical microelectrodes inserted into the brain's motor cortex. Perhaps unsurprisingly this is a technique not widely adopted; the majority of its application in research is in studies on monkeys & other non-humans. Even setting aside the financial cost and medical risk of the brain surgery required to install such devices, their instability over long-term use diminishes their suitability [77]; studies *in vivo* have seen fewer than 40% of samples work function properly for prolonged periods [78]. This degradation in quality is thought to be in part due to the damage caused during their insertion but also to continued interactions between cortical tissue and the electrodes while they are in place; the brain's natural reactive response to foreign bodies results in a "sheath" somewhat akin to scar tissue being formed around the needles, which affects the quality of the electrical signals they can record [79].

Electrocorticography (ECoG) is another invasive neuroimaging technique which makes use of electrode arrays placed directly on the surface of brain, rather than within it. Being less severely invasive than the intracortical implantation of electrodes [77] ECoG is more common in humans, though still decidedly rare. Research studies which use this technique routinely recruit participants in whom ECoG electrodes had already needed implanting, for monitoring health conditions such as certain types of epilepsy [80,81]. Although ECoG electrodes are only placed on the outside of the brain they nevertheless require intracranial surgery for their implantation, and they can damage surrounding tissue [79], so for obvious ethical reasons this technique is not widely carried out on subjects for whom it is not medically necessary.

Given this, and indeed the strong aversion among prosthesis users to invasive systems even were such technologies able to improve prosthesis dexterity [27], noninvasive measurements of neural activity are significantly preferred. Electroencephalography (EEG) is a noninvasive method of measuring the brain's electrical activity through use of electrodes placed on the scalp[8]), first devised by Hans Berger in 1929 [82]. While other noninvasive neuroimaging techniques exist and do see use in research, such as functional Magnetic Resonance Imaging (fMRI) and Magnetoencephalography (MEG), these often require immobile, extremely expensive equipment. EEG devices are not only often much cheaper but can be made wireless [83], powered by battery and transmitting data via Bluetooth or similar protocols, thus allowing for portability.

The most widely recognised standard for placement of EEG electrodes originates with the International Federation of Clinical Physiology's 10–20 system pioneered by Herbert Jasper [84], so named because the distances between the electrode sites are variously 10% and 20% of the skull's length. This has since been extended by adding electrodes midway between those of the 10–20, this higher-density arrangement being thus named the 10–10 or "10%" system [85]. As is implied by the labelling system of the original 10–20 (certain numbers were "left out" to be later "filled in"), Jasper anticipated such an extension [86]; it is perhaps for this reason that despite the resultant imprecision of its title these additional electrode locations have been adopted back in to the International 10–20 system [87][9]. While originally utilising 81 electrodes, implementations of the extended 10–20 system typically use 64 or 32, or at times some other quantity where there is particular focus on a specific brain region or where cost reductions are a motivating factor. A further extension of yet higher electrode density, the 10–5 system [88], also sees occasional use. This has been critiqued however as being of such high density that the location of any given electrode, when actually positioned upon the head,

[8] It is thus also occasionally called "scalp EEG", in contexts where ECoG is referred to as "intracranial EEG", but this is not convention; simply "EEG" will be used throughout this thesis.

[9] Properly this version is called the "*Extended* 10–20 system", but such a distinction is rarely drawn in literature due to the extended system's overwhelming prevalence.

is often within standard deviations of its neighbours' intended locations [89].

The use of EEG over invasive measures, while clearly preferable as discussed above, can nevertheless present challenges. One intrinsic challenge in the measurement of EEG is that of volume conduction. Just as in sEMG, EEG electrodes are situated far from the sources of the signals they are used to measure, and in travelling this distance signals encounter various materials of differing electrical properties which can distort, attenuate, and noise them [90]. The issue of volume conduction has even been found to impact the temporal resolution of EEG [91], as the time signals take to travel from their originating sources to the scalp electrodes can obfuscate the exact time at which a neural activity itself occurred, which can be problematic when such precise timings are of interest to researchers. In addition, there is somewhat of a topological mismatch; neurons in the brain are distributed in a three-dimensional space, but an array of EEG electrodes is framed largely as a two-dimensional surface [92] (as exemplified by Figure 4.2 below). A signals being measured at a given EEG electrode hence may not always necessarily indicate the location of the source from which it originates [92, 93]; up to half of the measured neural activity at a given scalp electrode has been found in simulation to be crosstalk [90] originating from interfering sources outside a 3cm radius [93, 94]. A number of source localisation techniques have been established for solving the "inverse problem" [95] by essentially projecting or "beamforming" [96, 97] recorded scalp EEG data into a dimensional space better representing the brain's structure [98]. Despite this, many studies opt not to perform such projections and nevertheless see success in classifying EEG data.

Another challenge in the application of brain data — regardless of neuroimaging technique — to contexts such as prosthesis control is the neural rearrangement which can take place in the brains of amputees. In short, the motor and somatosensory cortices of amputees can reorganise themselves post-amputation. Residual limb muscles located immediately nearby the site of amputation can end up with motor cortex representations which expand or move into the region previously associated with the amputated body part [99, 100]. This may be due to the cortex's somatotopy: remaining muscles situated near the amputation site are generally controlled by motor neurons nearby those which controlled the removed or "deafferented" muscles themselves, however in some cases the hand-controlling region has after amputation been remapped to control of the lip [101] despite these appendages' non-adjacency in the cortex (Figure 2.3). It should be noted that phenomenon presents only among those with a surgical or traumatic amputation and not those with a congenital limb difference, and is thought related to the phenomenon of the "phantom limb", which the latter group do not experience [5]. It may even be associated with the degree of phantom limb pain (PLP) an amputee experiences, though research is yet inconclusive on this topic [99, 102] nor the extent to which prosthesis use may exacerbate PLP, alleviate it through accelerating remapping, or have a no effect but simply noncausal association, as those experiencing PLP are often less likely to use their residual limb with or without a prosthesis [103]. In fact much of the nature and mechanism of cortical remapping is not yet fully known; some results have even found the cortical map of the residual limb to be no different to that of the homologous intact limb, suggesting that such a shift did not occur at all [104]. While the specific implications that neural reorganisation may have on prosthesis control, such as any potential long-term drift in amputee's EEG data, are well beyond the scope of this work, it is certainly a factor which highlights the limitations which can arise in systems using neural data alone for gesture recognition.

Literature Review

3.1 Approaches to multimodal gesture classification

The term "Hybrid Brain-Computer Interface" was coined by Pfurtscheller et al. in 2010 [105] to refer primarily to systems comprising two or more electroencephalography-based BCIs working in tandem, but also encompassing systems which use data of multiple sensing modalities. While Pfurtscheller et al. gave only the heart rate and eye gaze as examples of such other data sources, a number of studies since have sought to leverage multiple different bioelectric signals (often neural and muscular activity, but at times others or even non-biosignal data [106,107]) for the classification of human physical activity. The approaches by which these multimodal data are used in a hybrid system however vary significantly among the literature [108]. To review the range of fusion approaches among biosignal gesture classification literature, a search was conducted of the Scopus and IEEE Xplore databases as sources of information. Inclusion criteria for considering a publication were the availability of a full-text article in English, a publication date between 2010 & 2022, and the following boolean search operator being met among the publication's Title, Abstract, or Keywords:

(((("EMG" OR "electromyo") AND ("EEG" OR "electroencephalo*")) OR "biosignal") AND ("fusion" OR "multimodal")).*

Studies describe gesture classification with various similar phrases (e.g. "movement intention") which could not be exhaustively anticipated, therefore instead of restricting the search by mandating such terms appear, articles with titles including the terms "emotion" or "sleep" were excluded as not relevant. After excluding duplicates, the titles & abstracts of identified publications were manually reviewed for relevancy. Additional articles discovered organically from other sources, or falling outside the inclusion criteria, were considered if judged to be of particular relevance. This highlighted a number of schools of fusion algorithm, which Table 3.1 summarises.

Approach	Synchronicity of data modalities	Purpose for which modalities are used
Sequential (Gated)	Upstream component using one modality activates an **independent** downstream component, which **subsequently** performs classification using other modality.	Upstream component typically uses simple decision-making (e.g. activity threshold) and **does not contribute** to the actual class discrimination.
Sequential (Cascaded)	Upstream component using one modality performs coarse classification. According to the upstream decision, one of multiple **independent** downstream components using the other modality **subsequently** performs fine classification.	**Both** modalities used to classify, though system's final class decision is output by downstream component. Can thus be used to construct multi-class system from binary component classifiers.
Decoupled (Error-Correction)	Upstream component using one modality performs classification. **Independent** downstream component uses other modality to identify errors in upstream classification.	Downstream error detection can be used to prompt re-attempted classification or to abort action taken in response to upstream decision - but **does not itself contribute** to classification.
Decoupled (Components)	**Independent** components using different data modalities **simultaneously** perform **separate** tasks.	Tasks may both involve classification but are typically **not directly related**, e.g. classifying movement at different joints.
Decoupled (Mode-Switching)	**Independent** components using different data modalities can be selected between, to perform equivalent or separate tasks. Only **one** component performs classification **at a time**, though a simultaneous "monitor" may identify a "trigger" command to cycle between the modes.	Components may provide **alternate mechanisms** for performing the same classification task. Alternatively, different components can be used to classify different possible gestures, expanding total range of classifiable gestures at expense of simultaneanity.
Joint (Early)	**Single** component uses **both** data modalities **together** to perform classification task.	Both modalities contribute **directly** to classification.
Joint (Late)	Ensemble of **independent** components each using different data modalities **simultaneously** perform a classification task.	Both modalities used to perform **equivalent** classification tasks. Components' outputs are combined.

Table 3.1: Overview of categories of biosignal fusion architecture precedented among literature for gesture classification

3.1.1 Sequential contribution

Some, such as Rocon et al. [107], utilise the signals on a sequential basis. Rocon et al. constructed a hierarchical system involving EEG, EMG, and Inertial Measurement Unit (IMU) data to characterise tremors in 12 patients. When their EEG-based Brain-Computer-Interface (BCI) detected activity in the motor cortex that was classified by bayesian classifiers as indicating "movement intention", a 128-channel EMG sensor was used to detect the onset of physical movement by measuring the width of peaks in the EMG signal. At movement onset, the IMU data was subsequently used to measure the kinematic characteristics of the movement, and filter voluntary movement components from those caused by the tremor. While such data can be of great use in the development of tremor suppression strategies with technologies such as Functional Electrical Stimulation (FES), the reduction of EEG & EMG information to be used solely as indicators of movement intention and onset respectively is of limited use to the classification of gestures for control purposes.

A sequential paradigm of multimodality such as this could be described as “gated”, in that it is characterised by some modalities of data contributing to the system when and only when certain characteristics have been observed in other modalities.

Sarasola-Sanz et al. used a similar gated strategy to enable control of a robotic upper-limb exoskeleton by a stroke patient [109]. A Linear Discriminant Analysis model continuously classified EEG data as representing either attempted movement or a lack thereof; where movement intention was identified, the exoskeleton would actuate, with its speed & direction partially determined by the EMG data. As Sarasola-Sanz et al. describe, this reduces the risk of an unintended movement by the robot — which is certainly a desirable property, particularly for an exoskeletal device wherein actuation in opposition to the movement of the body part within could undoubtedly be hazardous. It is again limited is usefulness however; while the patient’s EMG affected the kinematics of the exoskeleton’s movement, this was ultimately through modifications to a predetermined target, and only primarily enabled opening and closing of the hand.

Khan and Khan. [110] reported joint use of EMG & EEG in the real-time control of a robotic hand by both able-bodied individuals and amputees. A measure of subject’s “concentration” levels was computed from their Beta-band neural activity, collected with NeuroSky’s commercial MindWave EEG device, and the robotic hand actuated only when this focus exceeded a given threshold. While the users’ EMG activity did affect the nature of these robotic movements, it was used only to determine the number of fingers actuated — three to pick up an object requiring less grip strength, and five for one requiring more — on the basis of EMG amplitude. This simplistic control mechanism appears motivated by the resulting reductions in sensor costs and computational requirements, both meritworthy goals in their own right. However it precludes the approach from being suitable for the identification of multiple distinct gestures for a more naturalistic control of the device.

A similar use of EMG data to determine the force or intensity of a movement following its detection via EEG can be seen in in the work of Du et al. [111], in the context of lower-limb activity. Here, the presence and direction of a leg movement were identified from a subject’s EEG data with a neural network. Only if this model predicted forwards motion, and a fibre-optic sensor corroborated this, did an EMG-based Bayesian classifier attempt to distinguish between walking and running.

There have been some more sophisticated variations on this strategy which do draw more fully on the information captured by Electromyographic and Electroencephalographic signals by using both in machine learning algorithms. These extend the “gated” system architecture to a “cascaded” design. This can be thought of as a tree-like structure, with branch nodes closer to the root relying on one data type to place datapoints in broad categories and, dependent on these decisions, those nodes “further down” a path from root to leaf discriminating between specific classes with data of the other modality.

Hooda et al. [112] proposed a cascaded system for lower-limb movement classification. Their system incorporated a bagged Decision Tree predicting the presence or absence of movement from EEG data, and in the case of movement a Support Vector Machine using EMG to distinguish between plantar flexion (“downwards") and dorsiflexion ("lifting") of the right or left foot. This strategy was more accurate than the use of either data modality individually, and outperformed the use of a single SVM to classify both EMG & EEG data combined at the signal level before feature extraction.

Ozdenizci et al. [113] were able to apply a similar approach to classify upper-limb gestures, with the addition of a layer of joint decision-making. For a given hand movement, EEG data alone were used to predict it as belonging to the right or left hand. EMG and EEG were then together used to identify this movement as an opening of the hand or as belonging among one of four grasp types, and these grasping motions themselves were subsequently discriminated between on the sole basis of EMG data.

3.1.2 Decoupled contribution

In certain studies which do in fact use multiple sensor modalities in parallel, these data still serve separate roles, being used for wholly decoupled tasks.

3.1.2.1 Error-correcting

One approach of this nature is the use of EEG as a supplementary measure to monitor or error-correct an EMG classification system. Kiguchi et al. developed an assistive exoskeletal robotic arm, which provided task-specific assistance to users through EMG-based movement detection and task identification via a stereo camera [114]. Error correction was facilitated first through EMG; when a user resisted the robot's movement, their work against it would require sudden high levels of muscular activity. Kiguchi et al. found that additionally monitoring users' brain activity via EEG could help detect this resistance more reliably. Förster et al. similarly identified the benefit of using EEG for error-correction, through the observation of signals known as error-related potentials (ErrP) [115]. These patterns of neural activity naturally present when an individual observes an unexpected error, and were hence observable in participants when an EMG-based classifier incorrectly identified the gesture they were performing. The use of ErrP signals as an indicator for a system to automatically re-attempt classification in efforts to correct it improved the classification accuracy by almost 10%.

In an interesting inversion of the paradigm, Riccio et al. implemented EMG-based error-control for an EEG "speller" BCI [116]. The P300 speller BCI is an established methodology that enables letter-by-letter communication by sequentially highlighting letters on a screen and observing occurrence of the "P300" brain signal which presents itself in a subject when their "target" stimulus appears among successive non-target visual stimuli [117] — in the case of the speller, after their intended letter is highlighted [118, 119] — after a delay of approximately 300ms when measured by EEG, thus giving it its name. Riccio et al.'s system involved an implementation of a P300 speller in which the "backspace" was controlled by a single-channel EMG sensor (placed on a bespoke muscle site in accordance with a user's residual voluntary movement capability) rather than being an ordinary option among the speller system's input characters. They found that in six healthy subjects, and one individual with tetraplegia (below-neck paralysis) and dysarthria (motor-related speech difficulties), the EMG-backspace not only significantly improved the error rate of a spelling task, but reduced both the time taken to complete it and the burden on users as measured by the NASA Task Load Index [120].

3.1.2.2 Discrete movement components

Kiguchi et al. later proposed [121] (though did not implement) an alternative system design involving the use of an EEG-based neural network to estimate the angle of a subject's arm during pronation or supination of the wrist in tandem with an earlier system which could enable control of a robotic elbow joint by EMG measurement of the *biceps* and *triceps brachii* (the upper-arm muscles which move a biological elbow joint). Ruhunage et al. [122] enabled control of a robotic limb by a similar strategy. In their work the elbow joint was likewise actuated in response to movement of the bicep and tricep as identified from EMG data. Their EEG component was responsible for the opening and closing of the hand, though rather than any classification of actual or imagined movements from motor cortex data, Ruhunage et al. used the Steady-State Visually-Evoked Potential (SSVEP). The SSVEP is a well-known pattern of brain activity often described as a "resonance phenomenon", wherein upon observing an optical stimulus which oscillates at a frequency in the order of a few tens of hertz, areas such as the visual cortex of a subject's brain will exhibit electrical potentials of the same frequency [123, 124]. By mounting a blinking LED of a known frequency onto the prosthetic hand, a measurable cortical potential of that frequency could be induced in a subject when they looked at it. Ruhunage et al. simply filtered the EEG data gathered at only one electrode (other than the reference) for the LED's 6Hz frequency, and commanded the hand to close when this signal's bandpower exceeded a threshold. This method is certainly admirable for its simplicity but ultimately does not offer a great range of dexterity to the user. It is also unlikely to be robust in a "real-world" setting; under experimental conditions the presence of external stimuli can be carefully controlled, but in a user's day-to-day environment they might encounter any number of visual stimuli that happened to oscillate at approximately 6Hz, which could pose a severe risk of unintended actuation by the hand.

While seemingly limiting in the range of possible control gestures enabled, a "divison of labour" of the constituent parts of a limb movement between sensing modalities such as is employed in these studies could provide an interesting means by which to potentially make systems more robust to the complexity of compound movements. This is an important topic of research unto itself, particularly if gesture recognition systems are to be reliable enough for everyday activities. In a task as ostensibly simple as for example throwing a ball, an individual is likely to perform the same hand gesture at a range of elbow angles. Whether the decoupling of data modalities in this way is in fact an appropriate solution to this challenge however remains to be found.

3.1.2.3 Mode-switching

Zhang et al.'s 2019 study [125] presents an interesting example of an approach which while technically multimodal in its use of EMG, EEG, and Electrooculography (EOG) in the identification of nine distinct gestures, does not in fact "fuse" the data *per se* by using them in any form of simultaneous classification. Rather, Zhang et al. implement a mode-alternating strategy. Two rapid eye blinks, detected via EOG, cycle the system between three states each defined by the data modality used: an EEG mode distinguishing between Kinaesthetic Motor Imagery of the right or left hand, an EOG mode identifying an eye movement looking to the right or left, and an EMG mode classifying five hand gestures with the "Myo" armband [46][1]. While each of these nine total input gestures were mapped to distinct behaviours for a robotic hand to perform, the mode-switching approach prevents them all from being simultaneously accessible by a user. Zhang et al. make effort to mitigate this usability issue by using the control actions easiest to perform to the robotic gestures deemed most important or used most frequently, but this is undermined by the resultant unintuitiveness of the mapping — with, for example, KMI of the right hand resulting in a "close fist" command to the robot, but a closing of the user's actual fist, classified in the EMG mode, resulting in a "ball pinch" (spherical grasp).

3.1.3 Joint contribution

That Ozdenizci et al.'s study [113] described above was one of few which classified between distinct same-hand gestures, and did so by incorporating a small degree of simultaneous use of the multimodal biosignals for gesture identification, suggests such parallel use to be more informative than the division of tasks between data types. Indeed, simultaneous usage of EMG & EEG data is a more common paradigm among prior attempts at fusion in this domain.

3.1.3.1 Early Fusion

One school of approach to the simultaneous use of multimodal data in classification problems is to combine the data in some way prior to learning from it, referred to as "early" fusion [127][2].

This is often described among literature as "data-level" or occasionally "signal-level" fusion and in certain cases the multimodal biosignals are indeed combined at such a stage. While Gordleeva et al. [128]'s work in distinguishing movement of the right or left foot from a resting limb was carried out primarily in the EMG and EEG domains separately, i.e. on a "unimodal" rather than multimodal basis, it did also trial multimodal classification on an offline basis. (Weak unimodal EEG performance led them to collapse their problem to a two-class one for the multimodal case, identifying only the presence of a movement in either foot regardless of which, rather than seeking to classify between them.) Their early fusion approach involved the processing of both EMG and EEG data jointly with a Common Spatial Pattern filter algorithm, an established technique in the field of EEG classification (discussed further in 4.2.5.1 below), providing the resultant processed features to a single Linear Discriminant Analysis classifier. This was found only to achieve accuracies between those reached by their unimodal EMG and EEG classifiers on an equivalent two-class problem. Aly et al. [129] used a Convolutional Neural Network to classify the EMG & EEG dataset gathered by Li et al. [130] of four transhumeral (above-elbow) amputees attempting four gestures: opening and closing of the hand, and pronation and supination of the wrist. These signals were jointly provided as raw time-series data to the CNN, in effect allowing its first hidden convolution layer to identify informative characteristics from the

[1] A commercial EMG device produced previously by Thalmic Labs Inc. [45] but since discontinued [126].

[2] A variation of early fusion wherein data are first transformed to share a common space is described by Pawłowski et al. [127] as "Sketch", but appears more applicable when dealing with data modalities vastly different in nature, such as text and images. While some biosignal fusion research does involve projecting data into shared spaces, given that EMG & EEG whilst distinct are both multivariate time-series measurements of electrical potential, such manipulations are not considered to be in line with the intended definition of the "Sketch" paradigm. They are hence for simplicity included among other early fusion strategies here.

merged data, with no distinction drawn as to the origin of each data modality.

Another strategy for signal-level combination of data is the direct analysis of the functional relationship between EMG & EEG activity. Corticomuscular coherence (CMC), an established measure of the synchronicity between brain & muscle activity [131, 132], has been used to identify the onset of movement [133] and to discriminate between movements of different muscle groups, such as the left and right hand [134], and has been used by Lou et al. [135] to classify between pairs of finger movements with greater accuracy than a model based solely on EEG. Other measures such as the Correlation between Band-limited Power Time-courses (CBPT) have also been proposed which, in contrast to CMC, are able to consider phenomena which occur synchronously in EMG and EEG but at different frequencies in each. Measures such as CMC and CBTP are primarily relevant however to the context of stroke rehabilitation applications, wherein the interaction between the motor cortex and the muscles is of specific interest in monitoring patients' development [134, 136]. The application of this approach to the context of prosthesis control, wherein multiclass problems of particular interest, is less established.

The more common form of early fusion however is for such combination to actually take place not with raw biosignal data, but after informative features have been extracted from them; thus more accurately described as fusing at the "feature-level".

Aly et al.'s earlier work [137], again using Li et al. [130]'s dataset of 64-channel EEG and 32-channel EMG, extracted time-domain, frequency-domain, and entropy-related features from each signal and additionally included coefficients of autoregressive models describing said signals' behaviour [138]. These features, in various combinations, were provided to a single machine learning model for the classification of gestures. While their reported classification accuracy is undoubtedly impressive, the extreme width of the feature arrays used poses potential risks both of overfit and excessive computational load. Also, that this approach was compared neither against alternative fusion strategies nor against a non-fusion (i.e. unimodal) system makes it difficult to ascertain the extent to which the early fusion itself contributed to the system's performance.

Li et al. themselves provided time-domain features of both EMG & EEG data to a single LDA classifier [130]. Unlike many others, they did trial broadly equivalent unimodal systems for comparison, along with various different combinations of EMG & EEG channels with a view to optimising the number of electrodes required. They found the multimodal system to classify gestures more accurately than models using either EMG or EEG data alone, and that this was the case even when the fused system used only 10 maximally informative channels each of EMG and EEG recordings. Al-Quraishi et al. [139] similarly extracted time-domain features from EMG and EEG, but performed Discriminant Correlational Analysis [140] to transform these into a single combined feature vector which was provided to their classifiers. While they found this feature-fusion system consistently more accurate than a single-mode EEG model, including where the EMG signal quality was decreased by muscle fatigue induced through physical exercise, they did not verify whether a system relying on EMG alone could have reached similar accuracies or been similarly robust to the fatigue.

Tryon et al. [141] supplied both EMG & EEG data corresponding to elbow movements of varying speed and force to a single Support Vector Machine (SVM), finding it able to classify between movement and rest with equivalent mean accuracy to a unimodal SVM trained solely on EMG, but with a slightly lower spread —

they concluded fusion be capable of enable greater robustness to variations in such kinematic characteristics. Subsequent expansions of this work in Tryon & Trejos [142] and Tryon et al. [143] found respectively that an SVM provided with fused EMG & EEG data was as also capable of predicting the muscle force as an EMG-SVM, and that feature-level fusion could also be performed by the use of a Convolutional Neural Network to classify both spectographic (frequency domain) and signal images generated from the EMG & EEG data and combined in various ways.

3.1.3.2 Late Fusion

The alternative approach known as "late" or "decision-level" fusion [127] involves the processing & synchronous classification of each data modality with wholly separate models. The predictions made by that ensemble of models for a given datapoint are combined in some way to determine the system's overall decision. The various strategies for the fusing of these can be broadly categorised into two groups: those which implement some form of static rule for combining the modalities' predictions, and those do so with an additional machine learning algorithm "stacked" onto the end of the classification process.

In binary systems, rule-based fusion algorithms can take forms as simple as Boolean operators [144]. Gordleeva et al. [128], in their aforementioned offline multimodal identification of the presence of foot movement from a resting state, also trialled a system which predicted movement when a datapoint was classified as such by either the EEG-LDA "OR" the EMG-LDA, and one which did so only when movement was predicted in both EEG "AND" EMG domains. While the accuracies of these strategies were similar to that of their data-level fusion approach, and likewise were well below that of a classifier modelled on EMG data alone, the OR rule was able to achieve the greatest True Positive Rate, and the AND rule the lowest False Positive Rate. Tryon et al. [141]'s work on multimodal "move-vs-rest" classification of elbow joint movement also trialled Boolean OR & AND operators for decision-level fusion, finding neither to be the most accurate of the various attempted fusion methods. In particular the OR rule's inherent susceptibility to the decisions of the EEG classifier, itself the weaker of the two data modalities, was noted as a likely cause of its poor classification accuracy — though their results indicate an apparent side-effect of the OR rule's accuracy being less variant over different levels of muscular force than most of the other fusion algorithms.

Naturally, Boolean rules do not readily lend themselves to multiclass problems. In such cases, classifier predictions corresponding to different data modalities are better combined if output in the form of classwise probability estimates, providing richer data than simply a single class label. Probability distributions can be fused with similarly simple rules; Cui et al. [145] implemented the "Max" rule, classifying each datapoint with the class label assigned the single highest probability among all the ensemble's classifiers, but found it consistently the weakest fusion method among all two-modal combinations of EEG, EMG, and Mechanomyogram (MMG) data and the second weakest in a three-modal system.

More common among biosignal literature implementing rule-based decision fusion is for a "fused" probability distribution to be calculated mathematically through some method of averaging those produced by the independent data modalities' classifiers. Leeb et al. in 2010 fused EMG & EEG data in the classification between movement of the right and left hands by computing the arithmetic mean of the two models' predicted probability distributions, in one of the seminal works in the field of biosignal fusion [146]. They found this to significantly outperform single-modality models, and the incorporation of EEG data in this way was also thought to partially compensate for the impact on classification accuracy of muscular fatigue — simulated by attenuating the amplitude of the EMG signals by various degrees — though as the effect of such fatigue on their unimodal EMG model was not investigated this cannot be claimed with certainty. The various works of Tryon et al. mentioned above also explored such averaging as a method of decision-level fusion in both move-vs-rest classification problems [141] and identification of the weight being lifted by a subject (and thus the degree of muscle force being applied) [142]. They trialled the mean as in Leeb et al. [146] alongside other static weightings, with variations biased at a ratio of 3 to 1 both in favour of EMG and of EEG. While no fusion approach at the feature- or decision-level outperformed unimodal EMG classification in detecting the presence of movement [141], their most accurate late fusion strategy (the equally-weighted average) did significantly outperform an EMG system at the $\alpha = 0.05$ level, albeit by a small margin of approximately 2.5%, in the weight classification task. [142].

There is also some precedent for the distributions output by models corresponding to different data sources being combined in ways informed by estimates of the reliability of their respective unimodal classifiers. Leeb at al.'s aforementioned study included a Bayesian method of fusion based on computing the two modalities' classwise predictive performance over the training data, finding this to reach accuracies competitive with those of their equal-weightings averaging [146]. Pritchard et al. [147] trialled a Weighted Average method wherein modalities' weights were determined by the accuracies achieved by corresponding unimodal classifiers in predicting a subset of testing data. Wang et al. [148] determined weights by a similar method but incorporated a penalty factor for datatypes determined more noisy, successfully classifying between four stages of a grasp-and-lift motion.

Cui et al. [145] provide one of very few comprehensive prior investigations involving comparison between a range of biosignal fusion strategies. Such explorations have also been carried out by works such as those of Tryon et al. [141,142], but Cui et al.'s is distinct from the latter in their classification of movements of wholly different types, rather than variations in speed or force of the same fundamental motion. Specifically, through various combinations of EEG, EMG, and MMG data they sought to distinguish between walking, cycling, and repeated stepping up to and down from a raised surface. As well as attempting the aforementioned "Max" rule, a Weighted Average with equal weighting distribution (i.e. the mean), and in the three-modal system only a Majority Voting strategy, Cui et al.'s work is a rare example among multimodal biosignal gesture classification literature of a "stacked" meta-model being used for late fusion. Each of the candidate classification algorithms tested for their system's constituent EMG, EEG, and MMG models were also trialled as options for this meta-classifier, which attempted to predict system-level classwise probability distributions on the basis of the probability distributions output by those constituent models. This classification-based fusion strategy outperformed their rule-based methods (described above) near-universally, and were frequently more accurate than unimodal systems. The Support Vector Machine (using a radial basis function kernel, and applying Platt's method [149] to compute probabilistic outputs) proved repeatedly the most highly-performing choice of meta-model.

3.2 Similarity of gestures in biosignal classification

Another limitation of many works in the biosignal literature is in the selection of gestures between which they seek to classify.

3.2.1 Among Fusion studies

As noted, a number of works using multimodal data do not attempt multi-gesture classification at all. Gordleeva et al. [128] reduced the specificity of their classification problem when using EMG & EEG together, on the grounds of their unimodal EEG system exhibiting poor multi-class performance. Hooda [112]'s classification of foot movements initially attempted to classify between flexion and extension of the left and right ankles in parallel, but later merged their gestures into a three-class system: movement of any kind in the left foot, the same in the right foot, and a rest class. While not entirely trivial such a reduced problem is ultimately much less edifying. The highly somatotopic structure of the motor cortex, and the nature of motor control as being primarily driven by the contralateral hemisphere of the brain (as covered in 2.1.2), means this classification task can essentially be reduced to one of identifying which area of the brain — i.e. which EEG electrode channel — presented a given pattern of neural activity.

Even Ozdenizci et al.'s work [113], one of few fusion studies which does classify between a range of task-relevant hand gestures, draws on EEG primarily in the earlier stages of their tree-like decision making process, to discriminate between movement of the right and left hands while leaving the identification of the movement's nature to the EMG component of the system. This is as mentioned again a more straightforward task for an EEG classifier than the actual gesture recognition stage, but more relevantly it would be an entirely trivial one for an EMG model, which could detect the presence or absence of movement generically in each arm by as simple a mechanism as an amplitude threshold. Likewise limited are the various works [129, 130, 137] drawing on the amputee dataset gathered by Li et al. [130]. Although these do indeed use information carried by the EEG data to contribute to discrimination between gestures of the same limb, the granularity of these gestures is quite coarse; as noted above the four defined movement classes are the closing & opening of the hand and pronation & supination (rotational movements) of the wrist. While self-evidently gestures of relevance to the topic of prosthesis control — indeed, many commercial prostheses do not provide biosignal-controlled wrist rotation — wrist movements and finger movements rely largely on distinct muscle groups. Grasping motions of the hand are mainly carried out by the various *flexor* and *extensor digitorum* muscles as discussed in 2.1.1, while the wrist is largely rotated by the *flexor carpi ulnaris* and *radialis* in tandem with their respective *extensor*s. These muscles are also located in the forearm, and their activity certainly less immediately distinguishable through EMG than, for example, movement of the right hand is from that of the left, but with sufficiently dense, properly placed EMG electrodes are not easily confused for the finger-controlling muscles[3]. More crucially however than than the potential lesser challenge of the classification problem in such cases is the resultant limitation in the system's capabilities, offering an imagined user only

[3] Notwithstanding complexities arising from compound movements, nor the influence of muscles which play a dual role such as the *flexor digitorum superficialis* which is used mainly in finger flexion but can assist some wrist motions.

one type of grasping gesture at any given time despite the range of hand shapes necessary for everyday activities.

Kurzynski's 2013 work [150] should be noted as a rarity among multimodal studies in its choice of task-relevant, similar, same-hand gestures. By taking a Weighted Average of three separate classifiers which used EMG, EEG, and Mechanomyogram (MMG, the measurement of acoustic vibrations generated by muscle movement with a microphone placed on the skin) data respectively, Kurzynski was able to classify between six distinct grasp gestures. While the classification accuracy is not reported, Kurzynski notes that the fusion approach offered performance statistically significantly greater than that of unimodal classifiers in over 75% of the trials.

3.2.2 Among Unimodal studies

In discussing Zhang et al.'s mode-switching system [125] above (3.1.2.3) it was noted that unlike many others this approach did enable control of nine distinct robotic gestures — but that this came at severe cost to usability, with the relationship between gesture inputs and robotic actuation being unintuitive. Such concerns are far from limited to Zhang et al.'s work, and in fact represent a trend across much of the biosignal literature.

With regard to EMG systems, Kim et al. [151] discuss the Repeatability and Separability Indices of gestures, metrics of the within-class and between-class variability (based on similarity in measured signals — they acknowledge that incorporating anatomical information regarding the muscle groups involved in different movements & the human musculoskeletal structure more broadly would be a meritworthy extension of their work). Their study suggests that minimising the former while maximising the latter ought to be the basis upon which input gestures are chosen. While in an abstract sense this is indeed a desirable property of a system, it ought to be recalled that Kim et al. largely discuss EMG-based gesture recognition in the context of everyday "consumer" human-computer interaction: their example applications include interfacing with software such as Microsoft PowerPoint and Google Earth. In such situations the relationship between a gesture input and its resultant command will inherently be loose; while facets of the design language may aid intuitivity, such as emulating the "pinch zoom" gesture common in touchscreen devices, the notion of naturalistic control does not apply. It is only a minor concern if choosing gestures chosen on the basis of maximal seperability results in them being non-representational.

In the context of prosthesis control however, determining input gestures solely on this basis could in fact be a significant hindrance to the quality of users' experience, if the selection of highly distinct gestures results in an unintuitive mapping of input to action. An accessibility device ought to minimise the demand it places on its user, lest it become itself a source of inaccessibility; properties such as the range of recognisable input gestures ought to be based primarily on users' needs, capabilities (given the variation in dexterity of control over residual limb muscles among amputees), and comfort. This means the framing of the problem ought to be inverted from the angle by which Kim et al. approach it. If intuitive control is desirable, rather than simply choosing gestures which are easily separable, more pressing is to find classification systems which can accurately identify gestures of lower intrinsic separability. Of course, the feasibility of accurate classification will nevertheless be a factor in gesture selection, and separability & repeatability are naturally beneficial to enabling this. Kim et al.'s framework may well provide an interesting way by which system designers can assess the feasibility of a potential suite of gestures. However it also serves to highlight the extent to which unintuitiveness in gesture recognition systems is problematic even among those solely using EMG — which are typically thought of as capable of much finer granularity in gesture definition than other approaches.

This issue is however much more prevalent among EEG-based systems, with many such studies in the literature classifying between imagined or executed movements which are not only highly dissimilar from one another, but also from the resultant system action they induce. Many works discriminate between broad categories of movement by different parts of the body, such as arm movement versus leg movement [58]. While the direction & speed of movements have been sometimes identified from EEG for applications such as remote control of a quad-copter [58], such systems also often lack intuitiveness; one common control scheme maps the simultaneous KMI of both right and left hands to the quad-copter's z-axis ascent [152–154]. This is perhaps unsurprising — not only are more physically distant appendages likely to present more geographically separable patterns of neural activity due to the motor cortex's somatotopy, but they also may be easier for participants to engage with. Many EEG studies situate themselves in contexts of stroke rehabilitation, paralysis, and similar cases wherein individuals' capacity for voluntary motor control is limited. Even if the participants of these studies are themselves able-bodied, this context motivates the choice of gesture classes. Similarly, in keeping with such intended applications the gestures themselves are often performed as Kinaesthetic Motor Imagery rather than genuine muscular movement. KMI is to some degree a learned skill [155–157]; novice users of EEG-BCIs are not always able to imagine the physical sensations of movement, as opposed to simply a conceptual notion of it, in such a way as consistently elicits motor cortex activity like *mu* desynchronisation. Given this, it is certainly plausible to expect highly distinct gestures, such as those of wholly separate muscle groups, to be less arduous for a subject to perform KMI of than those which differ more subtly. Such limitations exacerbate the issue of gesture selection seen in fusion research. By failing to investigate the potential usefulness of EEG in classifying between multiple same-hand gestures, the use of such gestures in multimodal biosignal studies is demotivated not through an established understanding of systems' incapability, but through a lack of confidence arising from the sparsity of evidence.

Nevertheless, some works have indeed attempted to classify between similar gestures of the same appendage with EEG data. Ofner et al. [75] used Linear Discriminant Analysis classifiers to identify six types of arm movement (hand opening & closing and wrist pronation & supination as in [130] among others, and also flexion and extension of the elbow) from EEG signals filtered to a very low frequency range of 0.3–3Hz. In the six-class "move-vs-move" problem they achieved a peak mean accuracy across subjects of 42%, and in classifying an aggregated "movement" class from a resting state a peak of 81%; exceeding the chance level in both cases by a statistically significant margin. As may be expected considering the motor cortex's somatotopic structure, errors in the multiclass system presented frequently in the form of movements being misclassified as their "opposites". This was most evident among hand movements. Finger flexion and extension while often confused were more likely to be classified as an hand movement of some form than as either a wrist or an elbow movement; the latter two were more often confused for one another than for movements of the hand. As seen in Figure 2.3, the motor cortex region associated with finger movements is both larger than those associated with the elbow and wrist and more distant from them than they are from each other. Not assessed in Ofner et al.'s work was the ability of a model to discriminate between multiple gestures and a rest class simultaneously, or the nature of errors in one which attempted to do so. Jochumsen et al. likewise used LDA models but in the classification between palmar, pinch, and lateral grasps of the right hand, gestures of much higher similarity. Curiously, as with Ofner et al. above they also did not include a rest class among their multiclass problem [158], though attempted most other conceivable tasks: move-vs-rest with an aggregated "movement" class, separate move-vs-rest tasks for each of the three gestures, pairwise gesture-vs-gesture classification, and a three-class problem between all grasp types. In the latter of these, the most interesting for applications such as prosthesis control where dexterity is desirable, accuracies averaging 63% were achieved with the use of frequency-domain EEG information. Spectral EEG features were also found useful by Xiao et al. [159] and subsequently Xiao, Liao, et al. [160] in the classification of individual finger movements. In the former, Xiao et al. reached a mean accuracy of 45% across six subjects in identification of the specific finger being moved, i.e. a five-class problem, with a Support Vector Machine [159]. The latter co-authored study classified instead between pairs of finger movements (e.g. "thumb-vs-index"), again with SVM models, achieving a mean accuracy of 77% across all subjects and movement pairs. Alazrai et al. [161] were able to go further by implementing a two-layer system of SVMs, which identify first the finger being moved, and subsequently used that finger's corresponding specialist SVM to discriminate between types of movement, such as flexion or extension. While their reported accuracies of each SVM are impressive they do not discuss the system's overall accuracy, considering that a motion classified as being of the wrong finger at the first layer will inherently be incapable of being accurately classified at the second.

Some other works have explored the classification of different grasping movements — common motions involved in various Activities of Daily Living — though rarely reaching accuracies notably greater than those of Jochumsen et al. [158]. Agashe et al. trialled the use of Multiple Kernel Learning models in classifying five distinct gestures defined by the hand shapes used to hold five everyday objects of varying size and shape: three whole-hand grips (a drinks can, a CD, and a screwdriver) and two precision-pinch grips (a coin and a bank card). Mean accuracy across these five grips was only 40% but varied according to grip type: the True Positive rate among whole-hand grips averaged nearly 50% whilst for the precision-pinch grips it was below 30%, suggesting the greater number of motor units recruited for the gestures which used more muscle groups may have lead their associated patterns of EEG activity to be more distinct. Their proposed Joint Angle Decoding technique however, involving the direct relating of measured EEG to angular velocities of the finger joints, was able to be used in real-time by an amputee to perform a reach-and-grasp task, though only enabling two grip types (a cylindrical whole-hand grasp for a bottle, and a lateral precision-pinch for a bank card). Iturrate et al. also focused on the distinction between whole-hand and precision-pinch grasps [162] of the right hand. Finding EEG electrode channel C3 of the 10–20 System [84] to be maximally informative in their early experiments, unsurprising given its placement over the hand-relevant part of the motor cortex in the left hemisphere (see Figures 2.3 & 4.2), they were ably to use LDA models to classify between these grasps using just eight channels of EEG from the contralateral motor cortex at an impressive mean accuracy of 76% over 10 subjects. Schwarz et al. [163] achieved similar results with EEG-LDAs, achieving a mean of

74% accuracy on the most separable pair of gestures among a whole-hand cylindrical grasp (for a cup), a precision-pinch (a needle), a lateral pinch (a key), and a fourth class of a reaching motion with no grasp, though accuracy fell to 66% when considering these gestures together as a multi-class problem. They did however find, akin to Iturrate et al., that systems with as few as 15 selected relevant EEG channels could be considered usable. An interesting alternative means of expanding the range of identifiable grasp types can be seen in the work of Mohseni Salehi et al. [164]. Here four task-relevant gestures were defined by the movement components comprising them: whether the four fingers (taken together) were flexed or extended, and whether the thumb was abducted or adducted. While less granular than the individual finger flexions classified by Alazrai et al. [161], they were likewise classified with a cascaded system, discriminating first between movements of the right and left hands, then identifying the flexion or extension of the fingers, and lastly the position of the thumb. Their reported accuracy of 64.5% across five participants demonstrates that while such an approach shows promise, there remain strides to be made in multi-class same-hand gesture recognition.

It ought to be acknowledged that the higher-resolution measurements of neural activity provided by electrocorticography (ECoG) have been used to successfully discriminate between similar hand gestures [58]. The higher spatial density of ECoG electrodes than EEG allow for better distinction between nearby motor neurons; by extension studies using intracortical electrodes — rarely carried out with human subjects — can achieve still greater fidelity, even at the level of individual neurons [165]. Schalk et al. [77], Miller et al. [166], and Kubanek et al. [80] among others have used ECoG in the identification of individual finger movements, and Pistohl et al. [167,168] were able to classify between a precision-pinch and whole-hand grasp types, including when such grasps were combined with an arm reaching motion. Despite these efforts ECoG, being an invasive procedure is as discussed in 2.2.2.1 less suitable for widespread use, and indeed research such as that of Engdahl et al. [27] makes clear that even if it enabled more dexterous, naturalistic prosthesis control, amputees are significantly deterred by the surgery required for such an approach.

3.3 Other limitations

While the limited depth of multimodal approaches explored, and the unintuitive & constrained range of gesture classes defined, are perhaps the most crucial limitations of much research in the field of biosignal-based gesture classification and the areas to which the research in this thesis makes its primary contributions, this work seeks also to improve upon two further aspects which it would be remiss not to mention.

3.3.1 Cross-subject generalisation

Among biosignal gesture classification research most studies,including the vast majority of works on multimodal classification [128,129,141,145,147,169], focus their attention on subject-specific systems — or even restrict their data universe to that of one recording of the bioelectric signals, a “single trial” [93,170]. This individualised nature of systems may be defended as being aligned with the typical paradigm of prosthesis acquisition, wherein amputees generally receive close supervision and support from prosthetists and other specialists to aid in the fitting of and adaptation to the prosthesis. In the case of a prosthesis controlled by gesture recognition, the selection and training of classification algorithms would naturally be a part of this process. Categorically, no attempt is made here to imply that such dedicated clinical care should not be given, or that a reduction in the support provided would be desirable. However if reducing the subject-dependence of gesture classification systems could enable such processes to be streamlined, through better foreknowledge of suitable models or a reduced need for subject-specific training data, this would certainly merit exploration.

While the fundamental underlying principles discussed in Chapter 2 are consistent, the particular characteristics of biosignal data vary between individuals, and even between data collected from the same individual at different points in time. This can be viewed as a domain shift; when attempting to classify on a cross-subject basis, i.e. with training data and testing data being gathered from different subjects, these datasets may be of different distributions [171]. Works which approach this problem have done so by various means. One strategy which allows for true subject-independence — generalisation of a trained model to wholly novel individuals — is to trial systems on a "Leave-One-Participant-Out" basis. This is in effect a specialised variation of k-fold cross-validation, wherein data are split into k subsets and, for each k in turn, a model trained on all folds except for k and used to predict k. Rather than a random split however as is typical of k-fold cross-validation, the data are split by individual: data of all subjects except for k used for training, and the model tested on k's data. Lu et al. [172] were able to use a Naïve Bayes classifier to distinguish four gestures from EMG data on this basis, at a mean accuracy of 89% across 20 subjects[4]. These gestures were an opening & closing of the hand and flexion & extension of the wrist. The discussion in 3.2 highlighted the need for further research on the classification of gestures more subtly distinct than these; this serves as a reminder that the limitations of prior research discussed in this chapter cannot be fully addressed in isolation but in fact interplay. Benalcázar et al. [173] used a k-Nearest Neighbours model to classify a similar suite of gestures, with the addition of a thumb-to-index-finger "pinch" grasp and a rest class. They reached only 54% subject-independent accuracy over 10 participants, the more complex classification task likely contributing to their poorer performance than Lu et al.'s system. Castellini et al. [174] performed cross-subject classification on a one-to-one basis, rather than a leave-one-participant-out, but interestingly reported similar mean accuracies when classifying between much more similar gestures — three types of grasp — both when subjects' arms were still (52%) and when the grasps were combined with other arbitrary motions of the forearm (54%). Fazli et al. [64] reported the first subject-independent zero-calibration EEG-based classifier, discriminating between Kinaesthetic Motor Imagery of the left and right hands with an ensemble of LDA models using a similar leave-subject-out cross-validation scheme. Their baselines with no subject-specific learning achieved an accuracy of 71% & their proposed strategy for subject-dependent adaptation reached up to 73% accuracy when tested on data from novel individuals. While undoubtedly significant achievements, much as with the aforementioned work of Lu et al. this should be considered in the context of the challenges in similar-gesture classification outlined in 3.2 above. The use of a "Leave-One-Participant-Out" strategy for exploring the cross-subject classification ability of various systems will be explored further in Chapter 5

As referenced with regard to Fazli et al.'s work, some studies have also trialled strategies for calibrating

[4] This is calculated from only those gestures classified by the EMG component of their system, so is marginally below the reported headline accuracy in [172]

or otherwise adapting a base "generic" gesture classification system on a subject-specific basis. Works such as those of Du et al. [175] & Ketykó et al. [171] explored such domain adaptation in the classification of EMG data with Convolutional Neural Networks (CNN), the former notably adopting an interesting method of continual "online" model adaptation. Transfer learning strategies explored among EEG-based works have often focused on attempts to generalise the Common Spatial Pattern, a popular signal processing technique discussed further in 4.2.5.1 below which is highly subject-specific in nature. Various strategies have been trialled to regularise these [176,177], devise them in a way which considers other subjects' EEG data [178,179], or model them for multiple subjects simultaneously [180, 181], in efforts to reduce the degree of necessary subject-specific calibration [182]. Other approaches have looked instead at the featurespace of subjects' data; Joadder et al. assessed the similarity of features' distributions between subjects to seek those most useful in Leave-One-Participant-Out classification [183], while Azab et al. [184] likewise assessed featurespace similarity but instead to identify for each novel subject the most similar individuals in their dataset, from whom learned model parameters could be transferred. In the EMG domain, Gonzales-Huisa et al. [185] considered subjects' featurespaces not to *assess* similarity but to explore style transfer techniques with which their data could be projected such that they *were* sufficiently aligned for effective cross-subject classification. Methods for cross-subject transfer learning as a means by which to potentially reduce the subject dependence of gesture classification systems while retaining the benefits of subject-specific learning through calibration are the focus of Chapter 6, and the precedent among literature of various techniques will be discussed further therein.

It should also be noted that very few works explore classification between distinct sessions even of the same individual. Here, the term "session" is used to mean a single occassion upon which sensors were fitted to a participant and their data recorded [171]; typically separate "sessions" would take place on different days but this is not strictly a factor.

3.3.2 Validity & Data Leakage

Poor practice in data handling, a misuse or lack of statistical analysis, and other related methodological issues are pervasive in Machine Learning research, having been described as fuelling a reproducibility crisis [186], and studies applying Machine Learning to biomedical contexts such as biosignal classification are unfortunately no exception [187].

Lotte et al. [30, 152] describe that many BCI studies are undermined by the presence, or at least appearance, of bias. Choices of models and their hyperparameters are often not justified adequately (or at all) by researchers, calling into question the validity of the claims & comparisons made between systems. Due to this lack of transparency, the rigour of methodologies cannot be assessed and the possibility that such choices were made manually by "cherry-picking" results to find those which lead to high testing accuracies cannot be ignored.

This is ultimately a form of data leakage. Leakage is typically conceived of as a duplication of datapoints between training and testing datasets leading to artificially high classification accuracies, but while this is clearly a potential source of leakage it is not the only one. In discussing the prevalence of such issues among studies on brain data, Hosseini, Powell, et al. [31][5] outline the risks posed by over-optimisation. Among studies that provide some justification for their hyperparameter choices (which not all do [143]), it is not uncommon for it to be stated that values were found through k-fold cross-validation or some other optimisation procedure, with few further details being offered [75, 160, 161], including a number of the already relatively scarce research on multimodal biosignal fusion [137,139,145]. In other cases it is even explicitly demonstrated that such selection was done on the basis of optimising test set performance [189]. While in such cases it is reasonable to expect that care was taken to avoid direct leakage by ensuring separation train and test data in the final model, an inherent leakage of information arises from the learning within that optimisation process. The model configuration used for testing to generate the headline classification accuracies reported is already known to be suitable for classifying the test data, giving it an unfair advantage and potentially leading to an inflation in the accuracies reported, by comparison to those it would achieve if provided genuinely novel data. Abreu et al.'s work on identifying sign language gestures from electromyographic data demonstrates exactly this, achieving vastly higher accuracies in offline cross-validation than could be achieved when classifying novel data [47].

It is worth highlighting explicitly here the way by which this is problematic. While generally speaking any machine learning model is typically expected to be capable of classifying “new” data when it is deployed, given that as discussed above (3.3.1) many studies limit their horizons to the data gathered in a single “trial” it may be tempting to dismiss this as a non-problem. However, even if a specific model is only intended to be used in the context of a given data sample in a specific study, the findings of that research — matters such as the suitability of the methods trialled for the given classification problem — are most useful if applicable beyond it. This is of course not say that research which does not go far enough on this point is wholly without merit or incapable of being influential and informing the direction of further research in the field. Likewise it is not even the intention to suggest that all studies which are not comprehensively outline how they have avoided such data leakage are compromised — such opacity could arise inadvertently through any number of unintentional oversights — but as Lotte et al. state it is not possible to categorically rule out bias in such cases. It is as such nevertheless a notable weakness of many works and one which the research presented in this thesis makes particular effort in its attempts to avoid.

Hosseini, Powell, et al. discuss various ways by which this problem can be mitigated. One strategy is the replacement of cross-validation with nested cross-validation — for each outer fold k, performing a further cross-validation on the non-k folds for optimisation purposes while k remains untouched, and only then testing on k. This has some precedent among biosignal literature; Pistohl et al. for example in classifying ECoG data optimise hyperparameters with a nested cross-validation over each 9-fold training set in their larger 10-fold cross-validation routine [168]. Hooda et al. take similar care to highlight that their 10-fold cross-validation is carried out only using the 80% of their data designated for training in their 80/20 train/test split [112]. Garrett et al. go somewhat further, dividing their five trials in every possible combination of a 3:1:1 ratio, with three used for training, one used as the optimisation target, and the last an unseen test dataset. This use of unseen data is ultimately the key recommendation of Hosseini, Powell, et al. [31]; they suggest that a

[5] Co-first authors, both named here in keeping with the principle of moving towards a better acknowledgement the collaborative nature of research [188].

portion of data be reserved throughout the entire experimental process, to be used only for final validation. This, as described in 4.2.3 below, is exactly what this work does — and in fact goes beyond. Here, not only are some data reserved but whole subjects are held out from experimentation. This not only allows for greater confidence in the validity of findings but for an indication of their generalisability to novel subjects, which is of clear importance in the seeking of gesture classification approaches which could be deployed for contexts such as prosthesis control at scale.

It should also be noted that such reservation of data need take place — as it does in this work — throughout *all* stages of the modelling pipeline, even those prior to optimisation. Whelan et al. [190] highlight how performing procedures such as the selection of informative features or of prospective modelling candidates on the basis of a whole dataset undermines any subsequent attempt at separation, stating that: "*restricting analyses to regions of interest that were determined in an initial analysis that included all participants will render invalid the subsequent cross-validation*" [190]. This is particularly pertinent to the context of biosignal data given the popularity of data-driven techniques such as the Common Spatial Pattern. The CSP, discussed further in 4.2.5.1 below, involves a supervised identification of suitable transformations for EEG data. If such transformations are learned on the basis of all available data, any subsequent divisions of the transformed data will not offer genuine separation of the information in the training % testing sets [191]. Even simpler, commonplace techniques such as the data normalisation can, if care is not taken to consider the level of information which ought to be available to a model and signals or features are instead normalised over the whole datastream as in [143], risk causing leakage in this way; 5.3.2.1 outlines the measures taken to mitigate that risk in this work.

The matter of temporal leakage should also be highlighted here. In the kinds of contexts relevant to prosthesis control, gesture information is not typically thought of as temporally correlated: an individual's hand gesture at a given point in time is not necessarily likely to be predictive over the next gesture they perform (without further contextual information as to their broader activity). Biosignal data are non-stationary however and time-correlated within narrow snapshots of time; it has been demonstrated that a model fit to EEG data of a given time t_0 may be able to predict data of a subsequent time unit t_1 not on the basis of any task-discriminant properties but simply on that temporal association [192]. This means that even when care is taken in the handling of differently assigned divisions of data, that data splitting process can itself be a source of leakage. Biosignal studies frequently do not specify the granularity at which they split data. Though some like Lu et al. [172] do specify that data were split according to "repeats" — that is, individuals "trials" or "performances" of a given gesture — in many more cases this is not made clear. If data are split randomly at the level of individual dataset instances, and time-adjacent samples are thus distributed between training, testing, and validation sets, this temporal leakage will undermine the separation of those sets; they will not be sufficiently independent [193]. As with the other matters of validity and leakage discussed here, the experiments in this work take particular care to avoid this pitfall, as is described in 5.2.3.

Dataset, Processing, and Feature Extraction

4.1 Overview

The experimental work of this thesis comprises three lines of investigation, each building on the previous, into the classification of gestures from multimodal biosignal data. This chapter serves to lay the foundation for those investigations by introducing the facets of the experimental conditions which are consistent between them — primarily the secondary data itself which is used, the steps taken for its processing, and the feature extraction procedure applied. Such topics are fields of research unto themselves and undoubtedly of great relevance to the quality of a gesture classification system, but are not the foci of this work's investigation and are hence controlled throughout it.

4.1.1 Selection of Dataset

As 3.2 describes, much biosignal research is hampered by the selection of gesture classes on the basis of high separability with little regard for their intuitiveness, with many studies not even going as far as to classify gestures of the same limb. In a context such as prosthesis control however a system's dexterity and ease-of-use are of great significance, as evidenced by the higher rates of abandonment of body-powered "hook" prostheses, inherently limited in their dexterity & unintuitive in their control, than of robotic ones [25]. Evidently there is much to be gained from further research into multimodal classification of naturalistic same-hand gestures of the categories most useful to prosthesis users, such as grasp variations [28], thus a multimodal dataset containing such gestures was sought.

Section 3.3.1 discussed the infrequency with which much of the biosignal gesture classification literature explores cross-subject classification. This is undoubtedly of significant interest to the field; progress towards achieving subject-independence may lead to gesture identification systems which are better able to classify data of naïve users, lowering the "barrier to entry" and thus potentially enabling such systems to be more readily used by those who need them. The stability of gesture classification models across subjects and the extent to which systems can benefit from cross-subject data is the focus of Chapter 6's investigation and thus the dataset's sample size was a characteristic of particular interest for reasons beyond simply the presence of more potential replicates.

Likewise, a classification system's generalisation ability across sessions of the *same* subject is obviously of great import in the context of prostheses. It would be distinctly unwieldy for a system to require complete retraining with new data on any given occasion upon which an individual wished to use it. A great deal of biosignal research does not however obtain data from the same participants across multiple sessions — more often, a single data collection session is treated as the entire "data universe" of the experiment [147, 184]. To enable the experiments on cross-session classification presented in Chapter 7, the retention of participants for multiple data recording sessions was also a desirable criterion in the selection of a dataset.

Some domain-standard single-modality biosignal datasets do possess some of these qualities. Among the BCI Competition sets of neurological data [194] are variously EEG motor imagery data (albeit not of distinct same-hand gesture classes) collected from nine subjects on each of two days (set 2a), ECoG data corresponding to individual finger movements of the same hand (set 4), and Magnetoencephalographic (MEG) data of two subjects performing four distinct movements of the right wrist (set 3). The NinaPro electromyographic datasets meanwhile comprise mainly EMG, often accompanied by kinematic data, from typically multiple tens of individuals performing up to 52 classes of hand movement, including some data gathered from transradial amputees [195].

With multimodal classification being much less precedented however, fewer multimodal biosignal datasets are available [196], and certainly none as well-established as the aforementioned BCI Competition or NinaPro. Li et al.'s 2017 dataset [130], utilised by a number of others in the field such as Aly et al. [129, 137] as noted in 3.1.3.1, captures data from four transhumeral amputees, each in a single session. While five same-limb gestures are included (closing and opening of the hand, pronation and supination of the wrist, and a rest class), these are coarser in nature than the kind of task-relevant same-hand gestures this work seeks to discriminate between. Pritchard et al. [147] — prior work of the author of this thesis — collected a small quantity of primary data, but as a proof-of-concept only coupled EMG and EEG activity which were simultaneous but otherwise unrelated. Tryon et al. [141]'s EMG & EEG data collected from 18 subjects is another dataset which has seen some re-use [142, 143]. The subjects' movements however are not of sufficient relevance to the problems of interest in this work — rather than distinct gestures, classes were defined by the speed and force of flexion and extension motions at the elbow joint. Luciw et al.'s WAY-EEG-GAL dataset [197], while capturing hand-grasping and releasing movements performed by its 12 subjects, similarly primarily investigated not a range of gestures but the nature of the single "grasp-and-lift" motion, by modifying the mass and surface friction of the stimulus participants were tasked to pick up.

The multimodal dataset ultimately identified as suitable for use in this research is that published by Jeong et al. in 2020 [198, 199][1]. This meets the various criteria described: both EMG & EEG data were collected from twenty-five individuals performing a range of gestures of the right hand on each of three separate occasions, and is described further in 4.2 below.

4.1.1.1 Regarding the use of subjects without limb differences

It is acknowledged here that despite the application of gesture recognition in robotic prosthesis control providing the key context in which much of this work is discussed, and motivating various decisions made throughout it, the data used is collected not from amputees but from able-bodied individuals.

[1] [198] is Jeong et al.'s paper presenting their work; the dataset itself is available at [199].

This is in part a pragmatic necessity — as discussed, the availability of multimodal same-hand gesture data is somewhat sparse, and this applies considerably moreso for data gathered from amputees. It is also however a conscious choice. Put simply, amputees are not a monolithic group and a range of factors which vary from individual to individual can cause significant heterogeneity among biosignal data collected from this group, plausibly moreso than among those without limb differences for whom many such factors do not apply. Scheme & Englehart note that differences both in scar tissue and in the geometry of the muscles remaining in am amputee's residual limb can affect the nature of their electromyographic activity [14]. While such factors will undoubtedly be partially dependent on the exact amputation site and the degree of residual limb remaining, it seems unlikely that even those who have undergone surgically similar amputations would for example develop scar tissue in exactly similar ways. The electromyogram is not the only bioelectric signal which may be unpredictably modified by amputation — as noted in 2.2.2.1, amputees frequently experience "cortical remapping" [5]. This reorganisation of the structure of the sensorimotor cortex (described in 2.2.2.2)is not wholly consistent in nature across individuals [101, 103, 104] and will naturally have an impact on the electroencephalographic signals measured.

Given the exploratory nature of this research, it was felt that the potential for greater variance among amputee data — which would be likely to be particularly impactful considering the smaller sample sizes typical of amputee datasets — would present too great an uncertainty to the experimental work. The author looks forward with much anticipation to seeing how future research can translate the experiments and findings presented in this thesis to work directly involving amputee participants.

Despite this, Scheme & Englehart's work [14] further found that while the classification accuracy of amputees' gestures from EMG data trended lower than of able-bodied participants, when comparing possible classification algorithms, the *ranking* of candidate models' performances was largely consistent across both groups. This clearly suggests that findings regarding suitable modelling choices and system architectures for biosignal gesture classification can generalise between able-bodied individuals and amputees. Of course the extent and reliability of such generalisation is not fully known and will undoubtedly be a topic of great interest to future research.

Such work may even find it possible to manipulate the data of able-bodied individuals like the subjects in Jeong et al.'s dataset [199] to artificially resemble that of amputees more closely. Research such as that of Campbell et al. [13] has successfully characterised some of the key population-level differences between amputees' and non-amputees' recorded EMG. While clearly the acquisition of further multimodal amputee datasets is paramount for the field, in the absence of this data there may be merit to making use of such transformations as, in a sense, a form of style transfer — though the exploration & application of such techniques is decidedly beyond the scope of this work.

4.2 Dataset description

4.2.1 Overview

The biosignal data used in this research were collected by Jeong et al. in 2020 [198]. The dataset comprises Electromyographic, Electroencephalographic, and Electrooculographic signals recorded from 25 participants, all right-handed and all inexperienced with Brain-Computer-Interfaces, performing both real movements (Motor Execution, ME) and Kinaesthetic Motor Imagery (KMI) of a range of upper-limb activities.

Participants gave informed consent to the study and for their data to be shared anonymously for reuse in future work such as this, and its methodologies were "*approved by Institutional Review Board (IRB) at Korea University (1040548-KU-IRB-17-181-A-2)*" [198]. The use of secondary data in this work is consistent with the dataset's Terms of Use[2] and has been reviewed by the Aston University College of Engineering and Physical Sciences Research Ethics Committee & given a favourable ethical opinion [ID EPS21031].

In this work the data of interest are the Motor Execution of Hand Grasping tasks, wherein subjects were instructed to pick up one of three common objects with their right hand and in doing so performed one of the following three grasp types:

- Cylindrical Grasp to pick up a glass cup, wherein the thumb and fingers are flexed while the thumb is abducted, i.e. the thumb and palm are in opposition;
- Lateral Grasp to pick up a credit card - style card, wherein the thumb is flexed following flexion of the fingers such that side opposition [200] is formed between the adducted thumb and index finger;
- Spherical Grasp to pick up a cricket ball, wherein the thumb is abducted and flexed, and the fingers splayed out and flexed to different degrees.

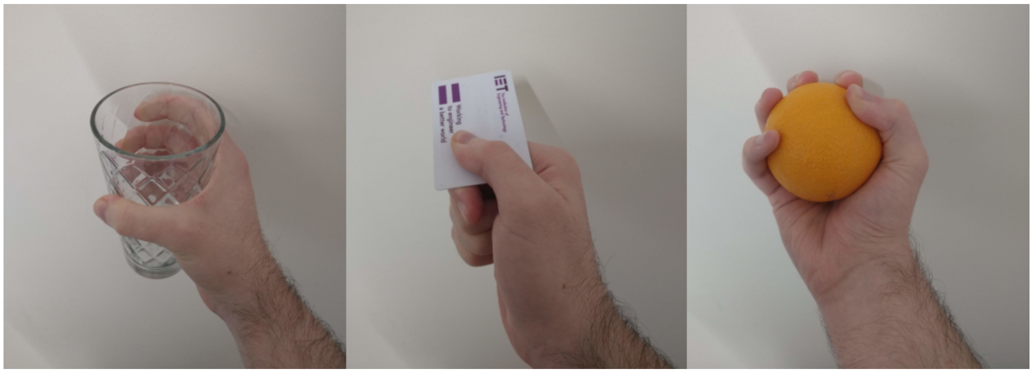


Figure 4.1: Illustration of the three grasp types (Left-to-right: Cylindrical, Lateral, and Spherical)

[2] `http://gigadb.org/site/term`

4.2.2 Procedure

In each task, participants were first shown a visual cue indicating a gesture to be performed, and given three seconds of preparation time. A second visual prompt was given instructing them to perform the gesture, which they did for four seconds, after which a final visual cue prompted them to put down the object and return their hand to a neutral resting position, which was maintained for a further four seconds before the next cue indicating the next gesture to be performed. In total they performed each of these three grasping gestures 50 times, in a randomised order, during the data collection procedure.

Each of the 25 subjects took part in three such data collection sessions, each separated by a one week interval. A total of 450 gesture performances, 150 of each grasp type, were thus collected from each participant.

The same procedure was followed in the recording of KMI data, with the distinction that rather than physically performing the commanded movement, subjects instead imagined the sensation of doing so. While KMI has been found in some research to induce trace EMG activity [76] such signals are weak in comparison to the EMG signals associated with genuine movement; given that ordinary EMG activity from Motor Execution tasks are available in Jeong et al.'s dataset, the inclusion of KMI would add little value to this work's investigation of multimodality and it is hence unused here. Future research may well focus more specifically on the needs of those amputees with severely limited control over their remaining muscles near the site of amputation, and so may find the EMG & EEG data from the KMI tasks to be of interest for this purpose.

On the same days as the above the subjects additionally performed wrist rotation tasks, involving pronation and supination of the wrist ("palm down" and "palm up" orientation, respectively), and arm reaching tasks, wherein the arm was moved up, down, left, right, forward, or back relative to the body hence incorporating movement at the shoulder and elbow joints. As with the Hand Grasping tasks, both Motor Execution and Kinaesthetic Motor Imagery of these activities were carries out. Such gestures are not within the scope of this work however subsequent studies may take interest in the inclusion of such movements, or of exploring more complex, naturalistic movements composed of multiple types of gestures; for example a task involving both a hand grasping and arm reaching component.

Between each ME or KMI task subjects were provided rest breaks, and longer breaks were allowed where subjects self-reported any physical or mental fatigue or other discomfort, in efforts to reduce any impact of fatigue, particularly muscular fatigue, on the recorded signals. During these rest breaks, the impedance of measurement electrodes were checked and appropriate steps taken, such as the injection of additional electrolyte gel, to ensure it was kept $< 15k\Omega$ to maintain signal quality.

4.2.3 Holdout Data

Section 3.3.2 discussed the issues of validity prevalent among much biosignal research due to cross-validation leakage & a failure to verify findings on unseen data. The experiments of Chapters 5, 6, & 7 explore many approaches for designing multimodal gesture classification and draw comparisons between them. As outlined below in 5.2.1, an optimisation procedure is used to determine classification systems' configurations in an unbiased way, free from "cherry-picking" in the selection of models and hyperparameter values [152]. This process however results in an intrinsic "baking-in" of information learned about the data used to optimise.

Total Participants	25
Gender	
Male	15 (60%)
Female	10 (40%)
Age (mean)	27.8
Development Set	20
Gender	
Male	12 (60%)
Female	8 (40%)
Age (mean)	27.8
Holdout Set	5
Gender	
Male	3 (60%)
Female	2 (40%)
Age (mean)	27.8

Table 4.1: Demographic composition of dataset splits

The resultant systems would have an inherent artificial advantage when classifying; systems would have been "pre-selected" on their suitability for classifying datapoints seen during optimisation, even if those instances were not present among the final systems' training datasets. Claims regarding systems' relative superiority made on the basis of such tests, without validation on unseen data, would risk generalising no further than the dataset used for optimisation, and their apparent classification accuracies risk being unduly inflated.

Therefore, to enable fair and valid comparisons between systems, five participants (20%) of the dataset were reserved throughout experimentation and used exclusively for verification, hereafter referred to as the "Holdout" Dataset, and the remaining 20 treated as the "Development" Dataset upon which modelling decisions would be made.

It should be stressed that the term "hold-out" is used throughout this work to mean "data excluded from all parts of analysis, modelling, and testing, not being accessed until such time as they are used to verify observations or test specific hypotheses" as outlined. This is described by Hosseini, Powell, et al. — whose 2020 paper "*I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data*" [31] illustrates the principle (though not the specific structure of the experiments in this work) in its 5th Figure — as a "*Lock-Box*" dataset on the grounds of "hold-out data" being defined inconsistently among the literature. Notwithstanding this concern however, the reserved data is referred to as "hold-out" in this work on the assumption it will be a more familiar framing to the reader.

Held out were participants 1, 6, 11, 16, and 21. These were chosen to preserve a consistent mean age (27.8 years) and proportion of female participants (40%) across both Development and Holdout Sets, as presented in Table 4.1, in efforts to control for any potential influence these factors may have on the generalisation ability of a system. For transparency & to enable replicability, the reader is advised that **these Holdout subjects continue to be identified as participants 1, 6, 11, 16, and 21 throughout the work**, rather than being relabelled e.g. as subjects 1 - 5 in any way.

4.2.4 Data Collected

Biosignal data were recorded from 70 Ag/AgCl electrodes, all sampled at 2500Hz by the same BrainAmp digital signal amplifier and passed through a 60Hz notch filter to reduce power-line interference. Three additional electrodes provided a ground and two reference signals. Of these 60 electrodes were used to collect Electroencephalographic data, arranged in accordance with the International 10-20 system and its 10-10 extension [84, 89]; the ground electrode was situated at Fpz and the primary reference at FCz. While a 64-channel array is conventional, Jeong et al. excluded electrodes FT9, FT10, TP9, and TP10. These four were repositioned to instead measure Electrooculographic activity, at three sites around the right eye and one at the left eye, for the purposes of artefact removal.

In this work, the EEG data collected by Jeong et al. were trimmed to 20 channels situated near brain regions relevant to the planning, execution, and sensation of movements (see 2.2.2.1). Reducing the number of required EEG channels could enable a gesture recognition system to be deployed at lower cost, a clear benefit to prosthesis users & a factor routinely reported as important to them [25,27]. While the minimisation of EEG channel count & the optimal selection of channels is a field of research unto itself [112, 201], precedent from works such as Schwarz et al. [163] & Iturrate et al. [162] (who as described in 3.2 were able to discriminate grasps with as few as 15 and 8 motor-relevant EEG electrodes respectively) among others [145, 146, 158, 201–203] demonstrates the viability of reducing a system's total EEG channel number in a way informed by established knowledge of the motor cortex structure, as is done here. Electrodes FC1-6, Cz-6, and CPz-6 were thus retained as highlighted in Figure 4.2.

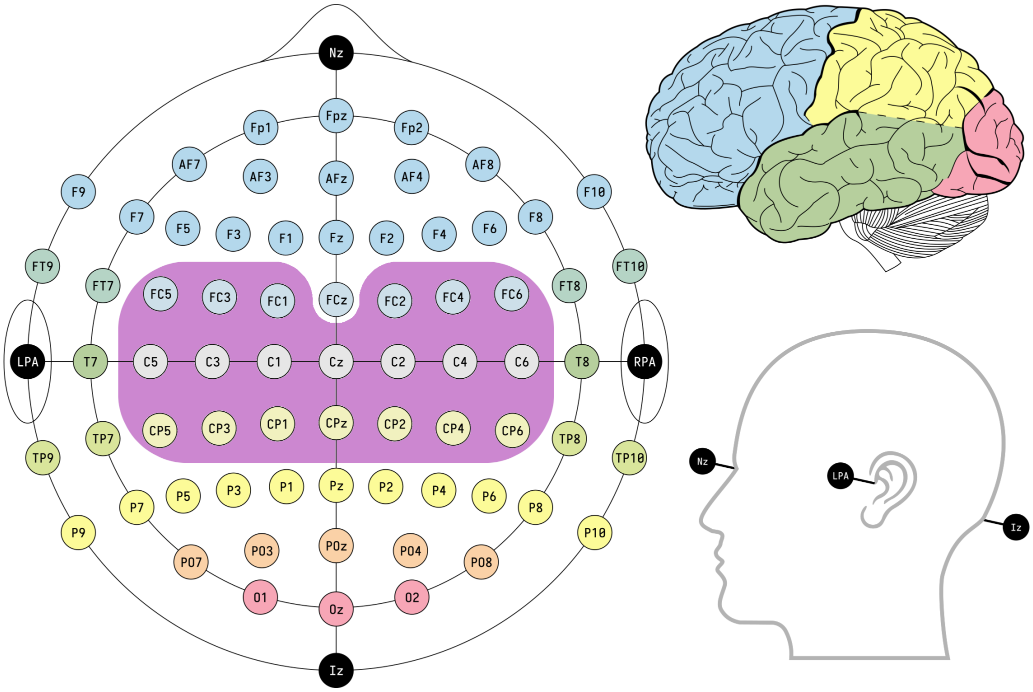


Figure 4.2: Placement of EEG electrodes used according to International 10-10 system [87], shaded purple. Adapted from [204]; originally published under CC0 1.0 [`https://creativecommons.org/publicdomain/zero/1.0/deed.en`].

Six electrodes were positioned over muscles of the right forearm, namely the *extensor digitorum*, *extensor carpi ulnaris*, *flexor carpi radialis*, *flexor carpi ulnaris*, *triceps brachii*, and *biceps brachii* (shown in Figure 4.3 as EMG1 - EMG6 respectively), to record EMG data. EEG electrodes Fpz and FCz remained the ground and reference respectively for EMG recordings, though a final electrode was placed at the elbow to provide an alternative reference. In Jeong et al.'s work the EMG signals were not used for modelling, rather to demonstrate the presence of muscular activity during Motor Execution tasks and its absence during KMI.

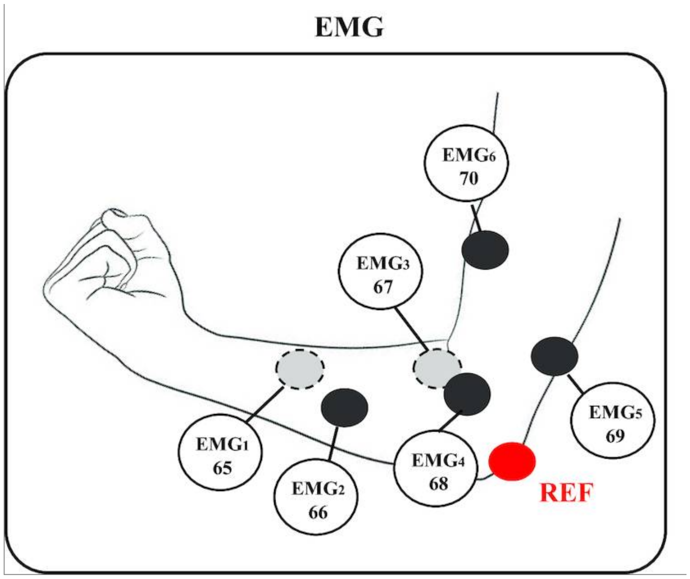


Figure 4.3: Sites of EMG electrodes used. © Jeong et al. 2020. Published by Oxford University Press GigaScience in [198] under CC-BY 4.0 [`https://creativecommons.org/licenses/by/4.0/deed.en`].

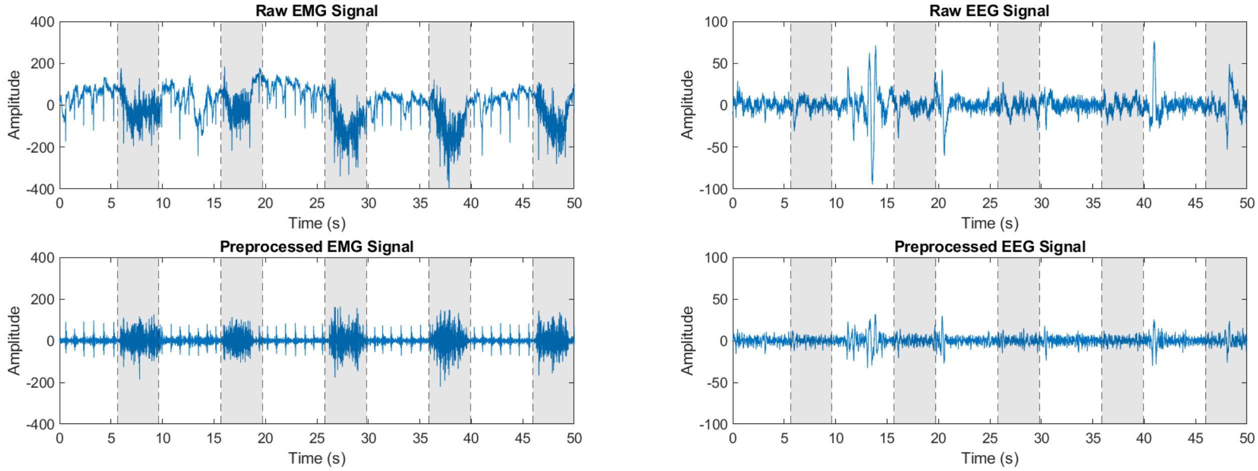
4.2.5 Data Preprocessing

Data were preprocessed in Mathworks MATLAB R2020a [205] with a script adapted from those provided by Jeong et al. [199], and the version of the Berlin Brain-Computer Interface (BBCI) Toolbox provided in their repository. For each recording, EEG signals were bandpass filtered from 2 - 30 Hz with a 4th-order Butterworth filter. EMG data were filtered from 10 - 500 Hz using a 5th-order Butterworth bandpass filter. Being an Infinite Impulse Response filter, a Butterworth filter introduces a phase delay which can differ at different frequencies. While MATLAB offers zero-phase IIR filtering via its *filtfilt()* function, this is implemented by forward-backward filtering of a waveform and thus requires the entire signal be available [206]. That would be achievable in the offline experiments of this work, but not in the kind of real-time gesture control contexts which motivate it. Filtering is thus instead performed with MATLAB's *filter()* function; the phase shift is accepted as an unavoidable by-product, which a developed classifier would need to handle.[3] Figures 4.4a and 4.4b provide illustrative examples of raw and pre-processed EMG & EEG data respectively.

As these figures also show, no specific measures were taken to identify and remove eyeblink artefacts in

[3] For completeness, the effect on EEG-based classification (the modality more significantly encoded in the frequency domain) of replacing this filter with a zero-phase equivalent was investigated and found to be negligible, outlined in Appendix B.

the EEG data. Conventional methods for this involve decomposing signals into Independent Components, eliminating those ICs similar to measured Electrooculography (EOG) signals, and projecting the remaining components back into the scalp space to reconstruct the EEG data [207], and Jeong et al. provide EOG recordings for this purpose [198]. Though much research into eyeblink artifact elimination has been conducted there is not yet a domain-established adaptation of the technique for real-time processing[4]. Those strategies proposed though often successful vary in effectiveness and efficiency [208, 209], and many increase the computational load or can introduce additional delay to the processing of EEG signals, even in the order of hundreds of milliseconds [210]. Given these limitations, and that there is no reason to anticipate eyeblinks as being class-dependent in the dataset used for these experiments, their removal was not a priority in this work. Figure 4.4b additionally demonstrates that while eyeblinks do remain present in the preprocessed EEG signal, due to their slow rhythm the 2–30 Hz bandpass filter suppresses them significantly.



(a) EMG (b) EEG

Figure 4.4: Representative examples of EMG (left) and EEG (right) signals before and after data pre-processing. Regions shaded grey correspond to periods of hand movement. NB: Pre-processed EMG seen here is before rectification.

In both EMG and EEG, individual gesture performances and rest periods were extracted from the recordings, each of a three second duration. While the gestures were as noted performed for four seconds, extracting from these a three-second epoch allows for some mitigation of the impact of any minor variability in subjects' reaction times or gesture durations between performances. This demarcated the 50 performances of each grasp type and approximately 150 distinct rest periods (each following a gesture performance). To ensure balance between classes, a pseudorandom sample[5] of 50 rest periods was taken. The resultant dataset thus comprised a total of 200 gesture performances, inclusive of rests, per participant for each session — 600 gesture performances in total per subject.

[4] Both this and zero-phase filtering could in theory be applied separately to each "window" of data processed in real-time. This would however add significant computation, and be limited in effectiveness — for a window of length T, the Rayleigh frequency $\frac{1}{T}$ dictates the lowest identifiable frequency component, and thus would limit the achievable lower cut-off of the bandpass filter.

[5] determined using a combination of the participant and session IDs of the recording as a seed, to ensure the same sample was taken of both EMG and EEG despite independent processing

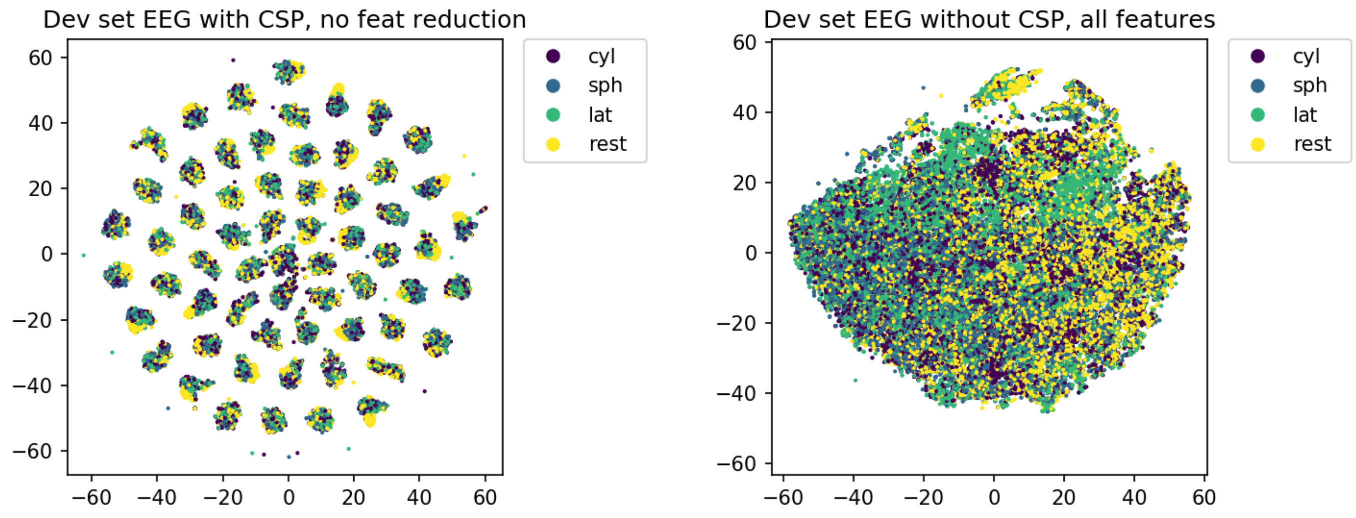
4.2.5.1 On Common Spatial Patterns

A popular (though not ubiquitous) signal processing & feature extraction technique among EEG literature is the use of the Common Spatial Pattern algorithm. This is a data-driven process by which spatial filters are learned and then used to project the raw EEG signals into components which are more discriminative with respect to a given condition — in the case of motor cortex studies, the gesture being performed. Often multiple CSP filters are used, selected to minimise within-class variability while maximising separation between classes [182]. This filtering aids in addressing the limitations of spatial resolution in typical EEG by identifying components of interest among the data which may not be easily detected otherwise, such as those electrical signals originating from point sources which do not align spatially with individual electrodes, or which are otherwise obfuscated by the volume conduction through the head [211]. This technique was used by Jeong et al. in their classification of the EEG data collected [198], and given its prevalence as a domain-standard technique was initially considered for use in this work. However, both exploratory investigations and theoretical reasoning suggested it unsuitable.

The CSP is a supervised data-driven method [152]; the projections it makes from EEG data and the transformations it learns to produce them are dependent on and unique to the characteristics of that data. Indeed it is generally applied not only on a single-subject basis [93], but within the context of a single continuous datastream from one recording session [170]. This is not itself an inherent weakness of the method; it is in line with the experimental paradigm of most EEG studies, which as noted in 3.3.1 often devise and test classification models on a subject-specific basis. The resultant dissimilarity between separate CSP projections would however be likely to risk amplifying cross-subject and cross-session variance in EEG data, impeding the ability of models to generalise across subjects, or even across data gathered from the same individual on multiple separate occasions. As put by Lotte et al., the CSP “[does] *not perform well with a large quantity of heterogeneous data recorded from other subjects or other sessions*” [152]. Given the specific interest this work takes in exploring the possibility of subject-independent and session-independent classification, applying CSP in the traditional manner would evidently not be suitable.

This unsuitability is evidenced by preliminary experimentation. Figure 4.5 visualises t-distributed stochastic neighbour embeddings (t-SNE), a dimensionality reduction technique which plots similar datapoints near each other & dissimilar ones farther away [212], of the EEG data both where CSP processing had been applied before feature extraction (outlined below) and where it had not. It should be noted that only EEG data of 20 subjects designated for model development are used here; as outlined in 4.2.3 the remaining subjects had been excluded from analysis at this stage to ensure modelling decisions were made without foreknowledge of the data used to verify results. Where CSP has been used, approximately 60 highly separable clusters can be seen. These are not present where features have been extracted from raw EEG. Recall that subjects each contributed 3 sessions to this dataset; across the 20 aforementioned subjects this equates to a total of 60 distinct data collection sessions. It certainly appears that process of Common Spatial Pattern projection did indeed introduce an artificial degree of separation between data originating from different recording sessions.

Further evidence for this separation can be seen by modelling & visualising t-SNEs on a per-subject basis rather than across all 20. Figure 4.6 present t-SNEs of EEG featuresets, both with and without the use of



(a) With CSP

(b) Without CSP

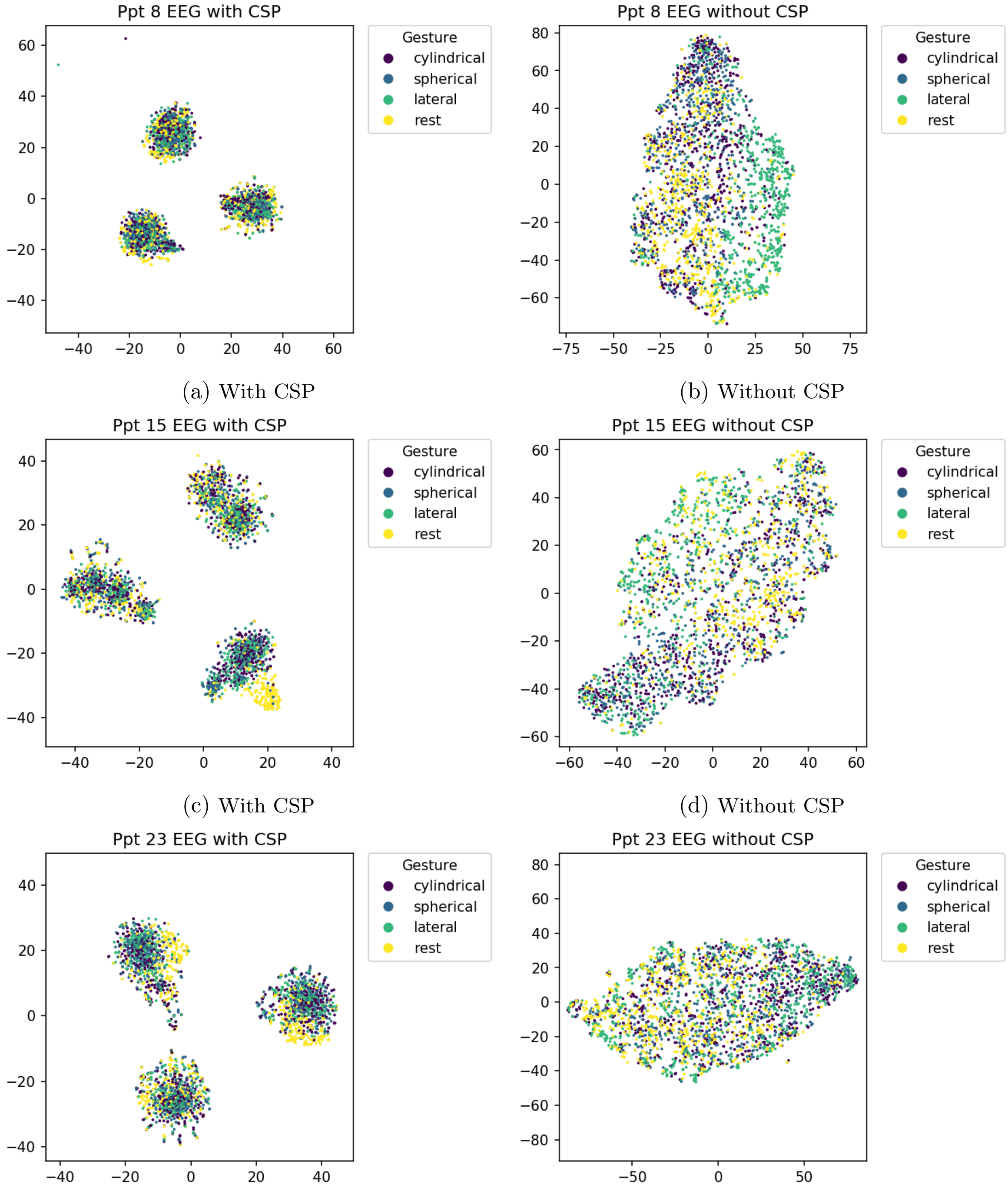
Figure 4.5: t-SNEs of EEG over all Development Set subjects

CSP, belonging to three subjects selected as illustrative examples. In each, the case where CSP has been applied can be seen to display a distinctly separable clustering of the data by the three recording sessions.

There has been some work, by Kang et al. [176], Lu et al. [177], and Guar et al. [178] among others, in despecifying Common Spatial Pattern filters to reduce the propensity to overfit. These approaches typically involve regularising the per-class covariance matrices estimated during the learning process towards predetermined matrices, such as the identity matrix (unit matrix) or a ”generic” matrix constructed from covariance matrices of other subjects' data. The latter of these has also been extended to involve pre-selection of the “other” subjects with whom to regularise the learning of the new subject's CSP filters for a more tailored variation [182]. Such techniques however are both less firmly established, and still reliant upon the availability of EEG recorded from the target subject or session. The CSP filters they find are somewhat less specialised than is conventional but are nevertheless data-driven; they are not universally generalisable to wholly novel data. The quantity of subject- or session-specific data required to effectively find even regularised CSPs ought not to be ignored. As with any other modelling process, to avoid a classifier's ostensible performance from being undermined by data leakage [191], the CSP projections applied to the data used to train it should not be found over the entirety of a given EEG recording. Rather, the spatial filters must be learned from training data (or a portion thereof) and those specific filters applied to otherwise unseen testing data[6]. This learning is likely to be a source of overfit in cases where only a small amount of such data is available, and thus to hinder systems' accuracies in investigations into cross-subject & cross-session classification, precluding regularisation of CSPs from being an appealing strategy here.

Given both the theoretical justification for its unsuitability and the preliminary evidence of a likely detrimental impact, Common Spatial Pattern filtering was not, despite being a well-precedented technique in BCI literature, applied to EEG data in this work.

[6] This is well illustrated in Figure 2 of [213], but is unfortunately not a universally applied principle in EEG studies.



(a) With CSP

(b) Without CSP

(c) With CSP

(d) Without CSP

(e) With CSP

(f) Without CSP

Figure 4.6: t-SNEs of EEG produced on a within-participant basis for subjects 8, 15, and 23

4.3 Feature Extraction

4.3.1 Time-Window Segmentation

While there is some precedent, at least in the EMG domain, for raw sensor readings to be used as input data for a classifier [214], bioelectric signals are typically understood to be stochastic in nature and instantaneous measurements of signal amplitude to carry minimal information [215]. The more common approach is to generate informative attributes for modelling by the extraction of statistical features from successive "windows" of the time-series data, describing the nature of a given signal over that time period [216]. Typically these windows are made to overlap by some percentage; with a period of x and an overlap of 50%, the first window w_0 would span from time t_0 to t_x and the second window w_1 from $t_{\frac{x}{2}}$ to $t_{\frac{3x}{2}}$. This is generally done for two purposes: to mitigate the loss of information which would otherwise be lost at the border between adjacent windows, and to improve the information rate of real-time systems by enabling a "new" window to be processed at, in the given example, a periodicity of $\frac{x}{2}$ rather than x [216, 217].

The appropriate length of such time windows, and the suitable degree of overlap between them, varies much among the biosignal literature. In EEG for example, choices have ranged from windows of 1 second with only a 50 millisecond offset (i.e. a 95% overlap) by Toriyama et al. [74], to 1 second with a 50% overlap in Bird et al. [218], to 250 milliseconds with a 150 millisecond overlap by Rimbert et al. [219], to 500 and 100 milliseconds with no overlap at all as in Gordleeva et al. and Yang et al. [220] respectively, among various others. While there appears little consensus, window dimensions are in part a task-dependent decision — this range of precedented choices is actually relatively narrow in the wider biosignal classification context; Candra et al. [221], for example deem windows of 3 – 12 seconds appropriate for EEG-based emotion recognition. Given Farrell et al. [222]'s finding of the optimal delay between control and actuation of a prosthesis to be in the order of 100ms[7], such extended durations would evidently be unsuitable for gesture recognition; all of the aforementioned window sizes used in motor cortex studies are by comparison vastly more viable.

Such variation is similarly found among EMG studies. Dolopikos et al. [223] use 1 second windows with a 500 millisecond overlap, Shahzaib et al. [224] non-overlapping windows of 200 milliseconds, Khushaba et al. [225] 150 millisecond windows with a 50 millisecond overlap, and Atzori et al. [226] used windows of 200 milliseconds offset only by 10 milliseconds, using parallelisation to extract features from multiple successive windows simultaneously to enable prompt availability for classification. Some studies have sought to specifically investigate the effect of the window length in EMG classification. Zardoshti-Kermani [227] found that the error rate in the classification of an above-elbow amputee's muscle force by a k-Nearest-Neighbours model dropped as window lengths were increased from 10 to 200 ms, and that this was the case for a range of different features trialled. Menon et al. [228] similarly found in the classification of gestures from able-bodied, transradial, and partial-hand amputees' EMG data, that the window length was negatively correlated with error rate when trialling windows from 10 to 550ms — but that the degree of overlap between adjacent temporal windows had no effect in any limb condition group. Smith et al. [229] meanwhile found

[7] It should be noted that [222] refers to a Proportional Control prosthesis (see Chapter 1) wherein a user may plausibly need a device to have an especially rapid response to enable them to adjust their muscle force in real-time.

that when trialling windows between 50 and 550 milliseconds in length, those 150ms or longer provided greater offline classification accuracy, but windows longer than 450ms resulted in lower completion rates of a real-time movement task by the participants. This again highlights the relationship between window length, prediction rate, and real-time suitability, though whether reducing the time-to-prediction by increasing the overlap between windows could mitigate here was not explored.

In this work, time-series data of each gesture performance were divided into windows of one second in length, with a 500 millisecond (50%) offset overlapping consecutive windows. The effects of window properties not being a focus of the current study, values were chosen which lie within the bounds of those precedented in literature. They have also specifically seen prior use in conjunction with the scripts[8] this work uses for feature extraction from both EMG [230] & EEG [218] data.

Using these scripts, a number of features (detailed in 4.3.2.2 below) were computed from each window of data. Features from a window were joined with those of the immediately consecutive window to form samples which each corresponded to 1.5 seconds of raw signal data. Thus four datapoints were extracted from each three-second gesture performance, each of which (except the first of the given performance) shared one window with its predecessor and so introduced 500ms of novel data. This is illustrated in Figure 4.7.

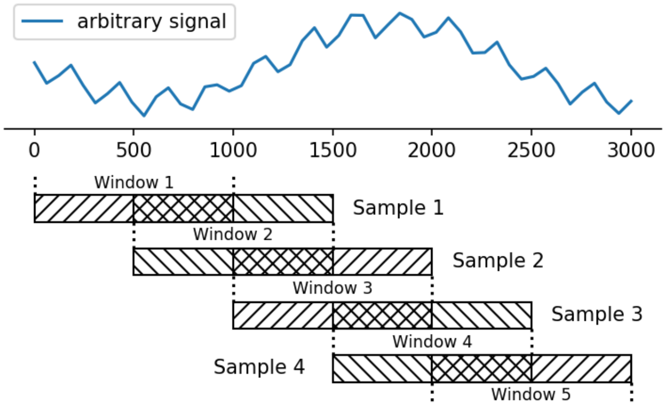


Figure 4.7: Illustrative sketch of time-window segmentation procedure.

In a deployment setting wherein live biosignal data were being measured for classification, this windowing would be performed sequentially in real-time. At a given time t_0, data from time $t_{-1.5}$ to $t_{-0.5}$ would form the first window from which features were extracted, and data from t_{-1} to t_0 the second; thus at each 500 millisecond interval (or more generally at each half-period interval of $\frac{T}{2}$, should an alternative window-length be used) a new datapoint would be generated for classification.

[8] Script adapted from that available at `https://github.com/fcampelo/EEG_Classification_`, itself adapted from that first pioneered in [218].

4.3.2 Feature Ensemble

4.3.2.1 Precedent

A number of popular feature choices for EMG data are outlined by Hudgins et al. [215]: Mean Absolute Value & Mean Absolute Value Slope are respectively measures of the signal's rectified amplitude and the rate of change of the same, Zero Crossing and Slope Sign Change both capture information related to signal frequency, and Waveform Length is a measure which also incorporates periodicity. While these specific features have themselves seen extensive use among EMG-related literature [49,50,109,112,130,145,224,231–235], alternative measures of the properties they capture have also proven popular such as for example the Root Mean Square as an amplitude measure [231,232,234,236] or the Wilson Amplitude (the number of times in a given window that two consecutive samples differ in amplitude by more than a set threshold) as a means of assessing the rate of amplitude change [50]. Other established feature choices include measures of signal amplitude's distribution over the given window such as the standard deviation [236, 237] or variance [50, 238], or even higher-order statistical moments [225, 239], and the extrema — the signal's maximum and minimum value — within the window [237, 238]. In some cases the assessment of the rate of change in amplitude has been extended to consider also the difference in standard deviation between subdivisions of a signal [189].

While many of these features are in the time-domain, frequency-domain EMG information is also used in classification tasks. The Zero Crossing and Slope Sign Change mentioned above have seen extensive use but other measures of spectral characteristics such as band powers of signal components at certain frequencies [240], the mean of frequencies weighted by their magnitudes ("spectral centre of gravity"), and the ratio between high and low frequency components [241]. Time-frequency-domain features including wavelet based decomposition [138, 242, 243] and the Short-Time Fourier Transform [38] have been found useful [244], and noted to be of potential benefit in the identification of the Motor Units recruited during a muscle movement when considered in conjunction with known properties of the relevant muscle fibers [245].

Many more features have seen successful use but are less frequently adopted; interesting examples include the Irregularity Factor, a ratio of the number of "ascending" zero-crossings to the number of positive "peaks" in the signal [225], and coefficients related to analysis of the "quefrequency" or "cepstrum" — the result of performing an inverse Fourier transform on the logarithm of frequency-domain data [241, 246].

While time-domain features dominate the EMG literature, EEG information by contrast is largely understood to be primarily encoded in the frequency domain and this is a much more common category of feature to be found among EEG studies [158–160]. As discussed in 2.2.2.1 the relative powers of EEG signal components at various established frequency bands are often taken to indicate the nature of neural activity; per 2.2.2.1 the power of the μ wave in the motor cortex & its event-related desynchronisation are typically of particular interest in classification of motor or motor imagery tasks [202, 247, 248].

Nevertheless a number of works have found time domain features of EEG signals to provide useful information for classification. Jochumsen et al. [158] used the mean amplitude of EEG signals along with spectral bandpowers in the classification of hand gestures, and Schwarz et al. [163] were able to achieve promising accuracy in the identification different grasping actions from only time-domain EEG features (primarily a moving average). Al-Quraishi et al. [139]'s classification of ankle movements likewise found success in the use of time-domain features including the aforementioned Mean Absolute Value and Root Mean Square extracted from both EEG and EMG data. Works such as that of Siuly et al. [249] have also evidenced the cross-correlation between EEG signals to be another informative property which can be of use in classifying motor imagery.

4.3.2.2 In this work

Given this precedent, this study thus sought to extract time-domain, frequency-domain, and correlation-based features from both EMG and EEG data. Key popular feature choices incorporated included measures of signals' amplitude and its rate of change, of distribution and extrema, and of their powers at certain frequency bands. The ensemble of features presented in Table 4.2, which has been found informative in previous work in the classification of both EEG [218] and EMG [223, 230], encompasses many of the features noted above as established in one or both domains. While not all the features comprising it are domain-standard for both data modalities, for simplicity (and given the featureset's aforementioned precedent), in this work all the listed features were extracted from each time window of both EMG and EEG data[9].

Following the formation of datapoints by the joining of features from consecutive epochs as described above, a small number of features which overlapped between those adjacent windows were then purged from the ensemble, namely:

- the mean, maximum, and minimum of the leading window's third quarter, which equate to those of the trailing window's first quarter;
- the mean, maximum, and minimum of the leading window's fourth quarter, which equate to those of the trailing window's second quarter;
- the forward difference in mean, maxima, and minima between the **initial** window's third and fourth quarters, which equate to the forward differences in the same between the **adjoining** window's first and second quarters.

The final dataset thus comprised EMG & EEG features extracted from 4 samples drawn from each of 4 gestures, each repeated 50 times in each of 3 sessions by each subject, i.e. $N_{samples} * N_{gestures} * N_{repetitions} * N_{sessions} = 2400$ datapoints per subject.

To enable synchronisation and later stratification of EMG and EEG datasets, each instance was additionally assigned identifying attributes corresponding to the participant number, recording session, repetition count of the gesture within that recording session, epoch start time, and epoch end time, which were all later removed before data were used for any modelling.

[9] Due to inconsistent literature definitions of the μ frequency cutoffs, and some work indicating a presence of μ activity in the low-Beta band as well as the Alpha with which it is typically thought to coincide, neural oscillations were binned into only the five most "established" bands rather than risk misleading implications of labelling any as the specifically motor-relevant μ.

Mean μ of all signals $y_1, y_2, y_3, ...y_n$	$\mu_n = \frac{1}{N}\sum_{i=1}^{N} y_{n_i}$
Standard deviation σ of all signals $y_1, y_2, y_3, ...y_n$	$\sigma_n = \sqrt{\sum_{i=1}^{N}(y_{n_i} - \mu_n)^2}$
Skewness γ of all signals	$\gamma_n = \frac{\sum_{i=1}^{N}\left(y_{n_i} - \mu_n\right)^3}{N(\sigma_n)^3}$
Kurtosis κ of all signals	$\kappa_n = \frac{\sum_{i=1}^{N}\left(y_{n_i} - \mu_n\right)^4}{N(\sigma_n)^4} - 3$
Maximum value of all signals	$y_{n_{max}} = \max(y_n)$
Minimum value of all signals	$y_{n_{min}} = \min(y_n)$
Backward difference in mean μ between first & second half-windows h_1, h_2 of all signals	$\nabla\mu_n = \mu_{h_2 n} - \mu_{h_1 n}$
Backward difference in standard deviation σ between half-windows of all signals	$\nabla\sigma_n = \sigma_{h_2 n} - \sigma_{h_1 n}$
Mean μ of each quarter-window $q_k \in \{q_1, q_2, q_3, q_4\}$ of all signals	$\mu_{q_k n} = \frac{1}{N}\sum_{i=1}^{N} y_{n_{q_k i}}$
Forward difference in mean μ of paired quarter-windows for all signals	$\Delta_{ab}\mu_n = \mu_{q_a n} - \mu_{q_b n}$
Backward difference in maximum value between half-windows	$\nabla y_{n_{max}} = \max(y_{h_2 n}) - \max(y_{h_1 n})$
Maximum value of each quarter-window q_k	$y_{q_k n_{max}} = \max(y_{q_k n})$
Forward difference in maxima of paired quarter-windows	$\Delta_{ab} y_{n_{max}} = \max(y_{q_a n}) - \max(y_{q_b n})$
Backward difference in minimum value between half-windows	$\nabla y_{n_{min}} = \min(y_{h_2 n}) - \min(y_{h_1 n})$
Minimum value of each quarter-window q_k	$y_{q_k n_{min}} = \min(y_{q_k n})$
Forward difference in minima of paired quarter-windows	$\Delta_{ab} y_{n_{min}} = \min(y_{q_a n}) - \min(y_{q_b n})$
Lower triangular elements of the covariance matrix of all signals	-
Eigenvalues of the covariance matrix	-
Lower triangular elements of matrix logarithm of the covariance matrix	-
Signal bandpowers corresponding to neural oscillations (computed by binning FFT components)	Cutoff Frequencies (Hz): Delta: $0.5 < f \leq 4$ Theta: $4 < f \leq 8$ Alpha: $8 < f \leq 12$ Beta: $12 < f \leq 35$ Gamma: $35 < f$

Table 4.2: Feature ensemble extracted from each window of raw EMG or EEG data

This feature extraction procedure was performed independently on the EMG & EEG datasets. This enabled the two data modalities to be treated wholly separately throughout the modelling pipeline until merged in accordance with the various fusion strategies outlined in 5.3.1, giving confidence that any impact of the multimodal approach was indeed due to said fusion. It additionally ensured any differences in magnitude between the EMG and EEG signals did not unduly influence any subsequent feature scaling process.

Further work however may find merit in investigating a joint feature extraction approach, enabling fusion at the signal-level (see 3.1.3.1); through the inclusion of the covariance matrix this would allow the measurement of any joint variability between EEG and EMG signals. Given that muscular electrical activity as measured by EMG is the direct result of synaptic transmission from the motor neurons innervating the muscle, i.e. that the measured electrical signals could be said to "originate"[10] in the brain's motor cortex, such covariance could be expected in principle to be strong and indeed correlation-based fusion approaches have shown some merit [134]. Scalp EEG does not measure at the level of individual motor neurons and likewise surface EMG is not easily decomposed into constituent Motor Unit Action Potentials; it may be that this coarseness of measurement precludes such relationships from being readily identifiable. Further, the highly somatotopic mapping of the motor cortex may suggest that such a relationship as unlikely to be influenced by the nature of a movement being performed; for two movements using the same muscles, the link between those muscles' fibers & their associated motor neurons is unlikely to vary between them, and hence EMG-EEG covariance may be of low informativity with respect to class in this case. Nevertheless, evaluating this empirically may prove insightful.

[10] See 2.2.1 for a more complete description of this mechanism.

Strategies for Multimodal EMG-EEG Fusion in Same-Hand Gesture Classification

5.1 Aims & Overview

To investigate the impact of leveraging Electromyographic & Electroencephalographic data simultaneously in upper-limb gesture recognition, suitable systems which can use such data to classify gestures need to be established. Chapter 5 therefore seeks to identify viable candidate systems for this problem.

As discussed in Chapter 3 the literature on biosignal classification is vast and while in certain contexts some trends and conventions in modelling choices do arise, a wide range of approaches have been used for this task with limited evidence as to the superiority of any over another and it is thus prudent to explore a variety of options.

This work proposes and compares the following three fundamental architectures (described fully in 5.3.1 below) for fusing EEG & EMG data in gesture classification:

- Feature-level, wherein EEG & EMG data are considered together by a single model to predict gestures.
- Hierarchical, wherein the predictions of a lower-level model based on one data modality are considered alongside data of the other modality by a higher-level model to make the final prediction.
- Decision-level, wherein EEG & EMG data are classified in parallel by separate models, and the predictions of each used to determine a system-level prediction.

The first and last of these are derived from domain-precedented methods of early and late fusion respectively; the Hierarchical approach is believed a novel strategy in this domain.

In the first stage of this investigation, an unbiased determination of suitable modelling choices — those which lead to highly-performing systems — is made for each of the proposed architectures. This is treated as a Combined Algorithm Selection & Hyperparameter Optimisation process [250] over a selection of candidate modelling choices informed by the biosignal literature as detailed in 5.3.3. This enables a fair assessment of the fusion architectures by providing equivalent opportunity for each to use its respective "best-in-class" arrangements, thus offering a more thorough comparison between a broader range of fusion strategies than that which has been done before in the domain (see 3.1).

This process additionally illuminates properties of the various models and techniques being considered, providing insights which can aid future research in the field. Some such findings establish a systematically identified underpinning of evidence to certain trends which, while having precedent in literature, have previously appeared to rely on *a priori* assumptions or to be a result of "herding" [251] as researchers opt to use popular techniques over new ones. As noted by Lotte at al. [30] among others, comparatively few studies on Brain-Computer-Interfaces draw objective comparisons between classifiers tested on the same problem while keeping other factors such as participants and feature extraction methods consistent, and many are insufficiently transparent — or even at risk of bias — in their modelling decisions. Indeed Lotte et al. [30]'s seminal work on BCIs states in conclusion that: "*One difficulty encountered in* studies *concerns the lack of published objective comparisons between classifiers. Ideally, classifiers should be tested within the same context, i.e., with the same users, using the same feature extraction method and the same protocol. Currently, this is a crucial problem for BCI research.*" [30]. Through its application of CASH optimisation to this domain and the unbiased, objective comparisons between systems enabled by doing so, this work addresses precisely that limitation. The use of CASH further provides an evidential basis, which much of the field is lacking, for design decisions taken in the work.

In the second stage of experimentation, the optimiser-identified configurations of the fusion systems are assessed on their ability to generalise to unseen subjects. This emulates the real-world use case of a complete gesture recognition system being provided to novel users, and hence provides an indication of the systems' suitability for hypothetical deployment. The systems are established as capable of being used by new subjects, and the extent to which each system can generalise to new individuals is evaluated & compared across system categories (for example between best-in-class Multimodal Fusion and Unimodal classification systems). This establishes strong candidate gesture classification systems on which to ground subsequent areas of investigation which form the following chapters, namely exploring the extent of the need for user-specific training in Chapter 6, and the evaluation of per-session calibration procedures in Chapter 7.

5.1.1 Aims

The overarching purpose of this chapter's experimental work is to identify a suitable system configuration(s) for classification of same-hand gestures using noninvasive biosignal, which performs accurately over multiple individuals. From this, the following subsidiary Aims are derived:

- **Aim 5.1** *Establish whether a multimodal system can offer better performance than a unimodal one*
    - **Aim 5.1.1** *Establish whether the fusion architecture impacts system quality, & identify performant fusion architectures*
- **Aim 5.2** *Identify modelling choices which can contribute to a multimodal or unimodal system achieving high classification accuracy*
- **Aim 5.3** *Establish a pipeline for the unbiased identifying of a performant multimodal system*

5.1.2 Deployment Approaches

In a real-world situation the user of a deployed gesture-recognition system, such as a prosthesis wearer, would need confidence that the system would perform well for them. A system could be deployed in one of two potential opposing approaches (or some middle ground as explored later in Chapter 6), according to which two categories of system were developed.

The first type of systems are subject-dependent in nature: trained solely on biosignal data of a single given individual and intended for exclusive use by that individual, being tested only on their ability to classify that individual's data. We hence call such subject-specific systems "Bespoke". Such an approach could be intuitively presumed to maximise a system's accuracy by virtue of being specialised to the user, but would would carry practical challenges in the time, cost, convenience, and accessibility implications of said specialisation. It should be noted that while any given Bespoke system trialled in these experiments is trained and tested on data from each individual subject separately, the system-level configuration of such a Bespoke system is not subject-specific. That is, the selection of Bespoke systems' component machine learning models and the tuning of their hyperparameters is not itself tailored uniquely to each Development Set subject — rather a single configuration is found for use with all subjects. This is motivated by a desire to achieve what is described here as "portability" of a Bespoke system — the identification of a model which can be trained on and subsequently used to predict data belonging to unseen individuals, on a subject-specific basis each time. This could potentially mitigate the resource implications of a Bespoke deployment, by avoiding the need for custom designing of a gesture classification system from the ground up for each novel user. Instead it provides a more "universally" applicable classifier configuration, which need only be trained on the new user's biosignal data.

At the opposite end of the spectrum the second category, referred to here as "Generalist" systems, are subject-independent — intended to classify data of a novel subject without any prior access to that individual's data for modelling. As implemented here these systems employ a "Leave-One-Participant-Out" approach, being trained on data provided by all participants *except* for a given individual, and tested on the accuracy with which they make predictions based on that unseen individual's data. They are thus measured on their ability to *generalise* to a new end-user with no adaptation, additional training, or calibration to the user's data. These hence mimic a hypothetical "off-the-shelf" use case, wherein a novel user could make use of the pre-trained system with a minimal barrier to entry, at the expense of sacrificing any potential performance gains arising from specialisation.

As was discussed in Chapter 3, systems of the subject-specific or "Bespoke" nature are vastly more common among literature and particularly dominant among previous studies on multimodal biosignal classification [128, 129, 141, 145, 147, 169]. While some work has seen varying degrees of success in developing user-independent EMG-based gesture classification systems [172–174], and to a lesser extent EEG-based systems [64], achieving strong generalisation performance remains a present challenge for research in this domain [171]. This work's experiments with its "Generalist" systems seeks to further the boundaries of subject-independent biosignal gesture classification and establish its possibility in a multimodal context.

5.2 Methodology

5.2.1 Overview

The overriding aim of this aspect of the research is to develop a gesture identification system that could be used successfully by a naïve subject, be that as it may a pretrained Generalist system the subject could use out-of-box or a Bespoke system to be trained on the subject's data.

Unbiased determination of candidate classification systems is a particular contribution of this work which has been noted as lacking among biosignal literature as discussed in 3.3.2. Lotte et al. [152] for example in reviewing the state of EEG Brain-Computer Interfaces state that: "*many studies did not compare the studied DNN*[1] *to state-of-the-art BCI methods or performed biased comparisons, with either suboptimal parameters for the state-of-the-art competitors or with unjustified choices of parameters for the DNN, which prevents us from ruling out manual tuning of these parameters with knowledge of the test set*". This work makes specific effort to avoid such pitfalls and present fair, unbiased explorations of the approaches to noninvasive biosignal-based gesture classification included within its scope.

To enable modelling choices to be determined in an unbiased way, the system configurations for each fusion architecture to be compared were found through Automated Machine Learning [252]. Rather than select classification models solely on the bases of *a priori* assumptions or their dominance among literature, it is acknowledged that no technique could ever plausibly be universally ideal [252]. Instead a range of possible modelling choices (described in 5.3.3 below) were established, and an automated algorithm used to explore these candidate options & identify those which lead to the highest classification accuracy. Similarly the hyperparameters defining the properties of those candidate models were not tuned manually, but likewise chosen through an automatic assessment of a defined range of options. Some fusion architectures comprise multiple machine learning classifiers. It cannot be guaranteed however that an EMG-based classifier, for example, which has been tuned to perform well in isolation will necessarily be a suitable component of a multimodal system. Optimising the different possible algorithms separately, and assembling an ensemble from those tuned models, could not only be an inefficient allocation of optimisation budget but risk overlooking combinations of modelling choices which are individually suboptimal but jointly effective. The processes of selecting classifiers and setting their hyperparameters, for all models constituent in a given system, were therefore performed simultaneously in a procedure known as Combined Algorithm Selection And Hyperparameter (CASH) optimisation [250].

This CASH optimisation was performed independently for each of the fusion architectures explored in these experiments. This enabled fair comparison between the fusion architectures each on the basis of their respective "best-in-class" systems as identified by the optimisation process, as it was assumed likely that one candidate fusion architecture may be more or less suited to different modelling choices than another. Bespoke and Generalist systems were also considered separately; their distinct natures likewise believed to make it unlikely that the same modelling choices would be optimal in both cases.

[1] While the passage quoted focuses on Deep Neural Networks, the trend Lotte discusses here and in their earlier work [30] to which [152] was a follow-up is not exclusive to studies proposing usage of DNNs

These optimisation routines rely on evaluating the predictive power of various different combinations of modelling choices; the CASH optimisation itself learns from the data used to train and test candidate options and tailors a system's configuration to that data. For this reason it would not be a valid assessment to compare fusion architectures using their "best-in-class" systems solely on the basis of the data used to identify those optimal system designs; this would be a source of data leakage [190, 191]. While such comparisons may be informative they could not be confidently said to generalise beyond the specific data used in the optimisation due to the inherent risk of overfit [31]. Therefore as detailed in 4.2.3 above, a portion of the dataset was reserved for verification of such comparisons as the "Holdout" data, and the remainder (the "Development" data) used for identification of the suitable system configurations. For each fusion approach, the most performant combination of models and their hyperparameters found during CASH optimisation with the Development data was taken as that approach's optimal configuration, and the resulting systems compared on the unseen Holdout data.

5.2.2 Procedure

As noted, the search space describing systems' configurations included both model selection hyperparameters determining the choice of classification algorithms used, and model-level hyperparameters describing the particular nature of the instantiated instances of those selected classifiers. Such model-level hyperparameters are relevant only to classifiers of their associated types; they hence each only affect the overall assembled system when the model selection hyperparameters were of specific categorical values. To enable the optimisation algorithm to allocate resources efficiently, it was thus important to ensure it was made aware of this conditionality. Such capability is inherent to the nature of Tree-Structured Parzen Estimators [253], making them a suitable choice of optimiser in this case.

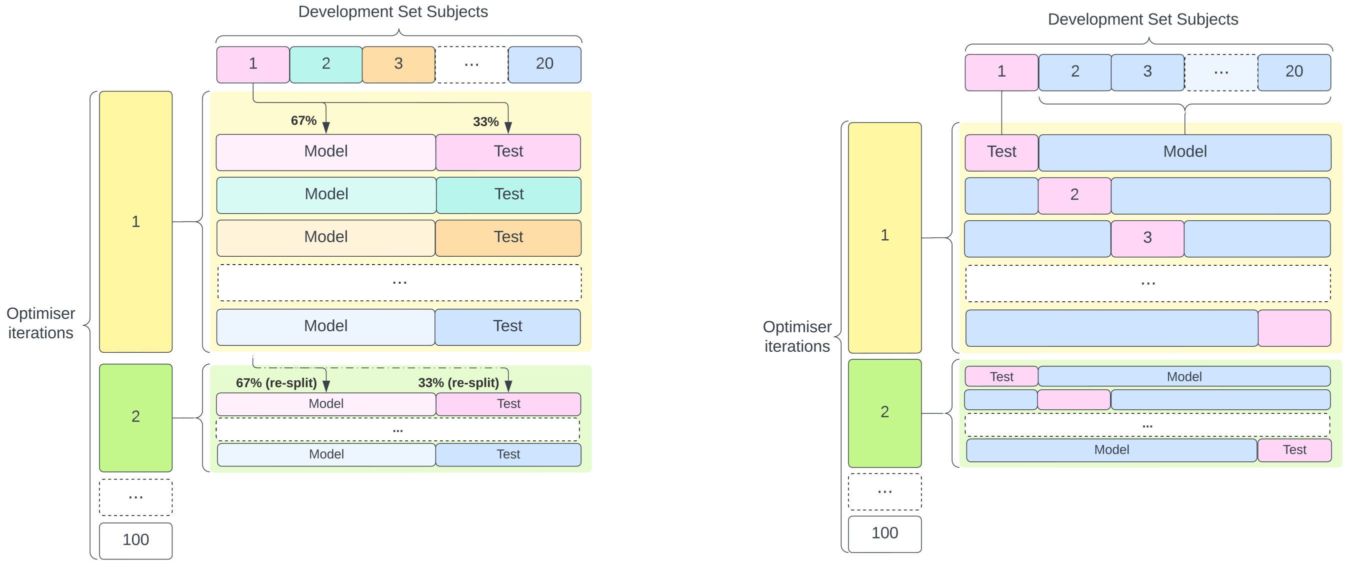
In both Bespoke and Generalist cases, for each Fusion architecture a Tree-Structured Parzen Estimator (implemented in the python package *hyperopt* [254]) was allocated a budget of 100 iterations over which to explore the search space and identify appealing hyperparameter choices. This budget was chosen to balance the benefit of exhaustive exploration of the space and exploitation of individual local minima identified within it, along with the pragmatic time and resource implications of performing both a Bespoke and Generalist optimisation task for each architecture, or subtype thereof, described in 5.3.1.

In each given optimisation routine, for every point assessed in the hyperparameter search space — i.e. each iteration of the optimisation process — a corresponding system with those parameters was assembled. This system was trained for, and tested on, every participant in the Development Set in turn. The arithmetic mean of classification accuracies for those 20 participants was taken as the system's Mean Accuracy, the complement of which was used as the loss function to be minimised by the optimiser.

For both Bespoke & Generalist versions of each Fusion architecture, the set of hyperparameters which provided the greatest mean accuracy across the Development Set in this way was considered the "best-in-class" system. These optimal configurations were then compared to identify the strongest candidate Bespoke and Generalist fusion systems. These were then verified on the Holdout Set to establish the extent to which the identified fusion systems could be used by an unseen subject. Unimodal systems (i.e. those using solely either

EMG or EEG data) were optimised similarly using the Development Set; in both Bespoke and Generalist cases the strongest unimodal candidates were then competed against the aforementioned winning fusion systems on their Holdout performance to evaluate the impact of fusion on classification accuracy.

To address Aim 5.3 by assessing the CASH optimisation's value in identifying suitable system designs, winning fusion systems' Holdout Set accuracies were compared to those of gesture classification systems defined non-algorithmically on inferences drawn from literature (hence referred to as "Literature-Informed").



(a) Bespoke

(b) Generalist

Figure 5.1: Learning process of Bespoke (left) and Generalist (right) systems during CASH optimisation procedure

5.2.3 Data Splitting

5.2.3.1 Optimisation

In the bespoke case, for every assembled system in an optimisation process, the data provided by each individual participant in the Development Set were divided into training and testing splits. This split was performed independently in each of the 100 optimiser iterations, to reduce the likelihood of the optimisation algorithm itself overfitting to a particular subset of participants' data. A random 67% of a participant's data were used to train the system and the remaining 33% reserved for validation. Crucially, this splitting was done on the basis of gesture performances, — all datapoints from any given execution of a gesture were grouped together. This ensured that data collected at consecutive time interval of the same gesture performance were not distributed among the training and testing splits, in efforts to protect against the issues of data leakage discussed in 3.3.2. Had time-adjacent datapoints been allowed to be divided between training and testing splits, any time-series correlations in the data would have risked artificially inflating systems' classification accuracies, due to models learning temporal artefacts rather than genuine motor activity [192]. The split was additionally stratified by class (i.e. the gesture being performed) to ensure an equal and balanced distribution of gestures appeared in both training and testing splits.

Scaling and selection of features was in each case performed on the training split of a participants' data as discussed further in 5.3.2.1 below, and the same transformations applied to their test split. The system was then modelled on the training split, and the accuracy of its predictions on their test split evaluated and reported as the classification accuracy for that participant. The arithmetic mean of these per-participant classification accuracies was as noted above recorded as the architecture's accuracy for that point in the hyperparameter search space, and was thus the optimisation target.

Generalist systems were by contrast evaluated on a Leave-One-Participant-Out basis [172, 173] in each of their optimiser iterations. For each participant N in the Development Set $\mathbb{D}$, all data provided by that participant N were reserved as the testing set, and all data provided by the remainder of the Development subjects $\mathbb{D} - \{N\}$ were used for training. Again care was taken to preserve this separation at all modelling stages: as detailed in 5.3.2.1 training data were standardised and the same transformation subsequently used to scale N's data, and features were selected on the basis of the $\mathbb{D} - \{N\}$ training data with N's data being reduced to those same features. The candidate system, configured according to the hyperparameters at that iteration's point in the search space, was then trained on the $\mathbb{D} - \{N\}$ training data, and the accuracy of its predictions of N's data recorded as the per-participant accuracy for subject N. As in Bespoke systems, the arithmetic mean of all 20 Development subjects' accuracies computed in this manner was the optimisation target.

5.2.3.2 Validation

The ultimate evaluation of gesture classification systems, such as the comparing of the most promising fusion & unimodal systems or between those derived from the optimisation-based pipeline just outlined & those defined solely by literature inferences, was performed by validating their generalisation ability to wholly unseen data. In these tests, systems' classification accuracies on the Holdout Set were hence assessed to avoid the common pitfall in Brain-Computer Interface research of insufficient validation of findings risking misleadingly high accuracies being reported, as previously discussed in 4.2.3 above.

It should be noted explicitly here that "systems" being validated does not refer simply to individual trained classification models being tested on Holdout data. The "findings" of the described experiments with Development Set data are in the suitable system architectures, configurations, and hyperparameter choices; it is these which are evaluated as follows.

For Bespoke systems, data from each of the five Held-out subjects were split in the same manner as described for Development subjects above; 67% of their gesture performances were used for learning — including scaling and selection of features as well as actual training of the subject-specific model — and the accuracy of gesture predictions made on the remaining 33% of their data evaluated. To mitigate the random effect of the random nature of the train/test split, scores for each Holdout subject were calculated as the mean of 100 repeats of such evaluations. This was carried out separately for each Holdout subject in turn to establish per-participant Holdout accuracies for each given system configuration being validated in this way.

In the Generalist case, systems of the configurations-under-validation were assembled and trained on the entire Development Set of 20 subjects. These trained systems were then used to predict the entirety of the data provided by each one of the 5 Holdout participants in turn. Here as there was no splitting of data, a system's per-subject scores were exactly repeatable, thus Generalist systems needed be tested only once on each Holdout subject for validation.

5.3 System Design

5.3.1 Fusion Architectures

Three fundamental fusion architectures (some with distinct sub-types) are presented here, drawing on and extending those with precedent among the biosignal literature. As discussed in Chapter 3, precedent for multimodal biosignal fusion in this domain is relatively limited; many studies which have drawn on both EMG & EEG data have done so either in "decoupled" ways, with each data source utilised for distinct purposes, or at separate "sequential" stages of a process. Of those which truly use multiple types of biosignal data simultaneously for gesture classification, the more popular strategies could be broadly separated by the stage of the conventional machine learning pipeline at which the modalities are fused — in particular, whether this is pre- or post - classification (i.e. "early" or "late" fusion respectively). The architectures considered here, as outlined below, thus span Early Fusion in the "Feature-Level" approach, Late Fusion in the "Decision-Level", and additionally propose the "Hierarchical" strategy — which can be seen as a synthesis of the two.

5.3.1.1 Feature-Level

In the Feature-Level Fusion architecture, illustrated in Figure 5.2, features derived from EEG & EMG data are merged prior to their classification by a single model. Such an approach has been explored in various works including [130, 137, 142], though is at times variously also described as "signal-level" or "data-level". Here the term "feature-level" is preferred as a more accurate reflection of the stage at which the modalities are merged, in contrast to approaches such as that of [129] wherein epochs of processed EMG & EEG data were provided to a CNN whose convolution layer extracted features from the merged data, or similarly [128] in which Common Spatial Patterns were extracted from a set of joint EEG & EMG data.

Joining the modalities at this stage allows for a "hands-off" approach; the learning process of a system's classifier determines the way in which information carried by each datatype is combined, and the extent to which each is drawn upon. Given that EMG-based gesture classification is typically considered a less complex problem than EEG, and its class-discriminative patterns are often more clearly identifiable from fewer features, Feature-Level Fusion systems may be encouraged to prioritise EMG data. This could allow a system to be less vulnerable to noisy EEG data. However, it may also lead greedier algorithms to pay insufficient attention to the complex patterns in EEG data during training, and fail to fully exploit the data available to them.

Considering this, the Feature-Level Fusion architecture as explored here has two sub-types: Joint Selection and Separate Selection. In both, data of the two modalities are first processed, and statistical features extracted, independently. In the Separate Selection variant, informative features are then selected from EMG and EEG featuresets independently, and joined into a single featureset for model training. This approach ensures the selected feature ensemble exploits both the available data modalities, though risks a reduction in the overall information captured in the featureset due to the potential for inclusion of EMG features which are highly correlated with included EEG features, or vice versa. In the Joint Selection case the features of both data types are firstly joined, and feature selection is performed on this merged set to determine an informative collection of features on which to train the subsequent classification model. This approach is more able to reduce duplication of information through avoiding inclusion of highly correlated features, but may be susceptible to unduly exploiting one data modality less than the other, as it makes no particular effort to select an equivalent number of features belonging to each.

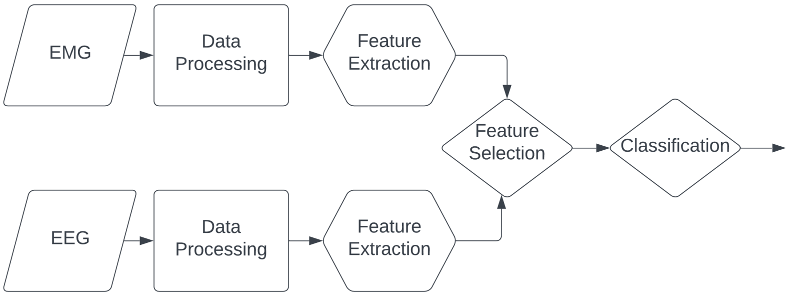


Figure 5.2: Feature-Level Fusion Architecture

5.3.1.2 Hierarchical

The architecture described as "Hierarchical" is believed novel, at least in this domain, though it takes inspiration in part from the principles of those approaches in literature wherein data modalities serve distinct *consecutive* roles in a system (described in Chapter 3 as "gated" or "cascaded") such as that of Du et al. [111] wherein EEG identified the presence and direction of a movement and EMG its intensity, Hooda et al. [112] wherein EEG was used to identify the presence of a foot movement and EMG to classify its type, and Ozdenizci et al. [113] wherein EEG identified the presence of movement of the right or left hands and distinguished between them, EEG & EMG contributed together to identifying the broad category of movement, and EMG alone informed the prediction of the specific hand gesture being performed.

The Hierarchical architecture also incorporates aspects of stacked generalisation techniques. Contrary to a conventional stacking method, wherein a meta-model is trained on the outputs of multiple base classifiers, in this approach the outputs of one data modality's base classifier are joined with the featureset of the other data modality, to be provided together to the “higher-ranking” model. This is performed on a probabilistic basis; for each given sample the lower-level model is used to predict its probability distribution over the four classes, and this distribution is then used to supplement to the sample's entry in the featureset (post-feature-selection) of the higher-level model.

This inherent assignment of rank suggests some interesting properties of a Hierarchical architecture. The information of the lower-ranked data modality is in effect collapsed into one feature — this classwise probability distribution. Should the lower-ranking model be extremely reliable, this will be strongly correlated with the target class; some algorithms may weight such a feature very strongly, and thus if selected for the high-ranking classifier may actually prioritise the lower-ranking model's decisions. More commonly however this consolidation is likely to make the lower-ranked modality easier to "ignore" at the system level. A hierarchical design could thus be robust to increases in task complexity, where such changes degrade the performance of one modality more than another — such as in this biosignal context, wherein EMG systems are typically capable of distinguishing between more gestures than EEG. In such cases, this architecture could minimise any detriment from a weak EEG component by downgrading that model's influence, the system collapsing to a near-unimodal paradigm. Unlike a typical Late Fusion approach however, wherein both modalities are so condensed before a system-level decision is made, the greater retained depth of the higher-ranking EMG data could even allow a Hierarchical system to learn more carefully the parts of its modelling space where EEG can be a valuable component of decision-making, and the parts where it need be weighted lower. This capability would be of particular benefit if there is low overlap between the data reliably classifiable with each datatype — allowing one modality to compensate at the residuals of the other — but where the high-ranking model is "confidently wrong" in its predictions & thus would be given undue preference in a typical Late Fusion weighting-based approach.

Both possible arrangements of data modalities in this Hierarchical architecture were trialled. The case illustrated in Figure 5.3 wherein class probabilities predicted by an EEG model were used to supplement EMG data is referred to hereafter as the "Hierarchical". Domain precedent indicates EMG-based gesture classification to be a comparatively easier problem than that which is EEG-based, suggesting a high likelihood of an EMG model outperforming an EEG one. This case, wherein the model which primarily considers EMG data outranks its EEG counterpart, is thus considered the architecture's "default" configuration. The opposite orientation, wherein probability distributions obtained from an EMG model are joined with EEG data as in Figure 5.4, is hence referred to as the "Inverse Hierarchical".

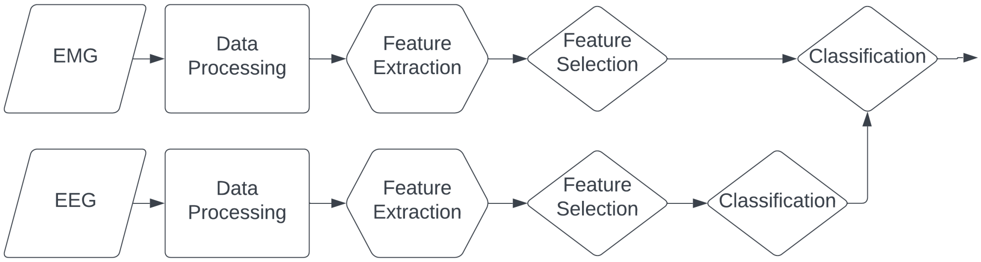


Figure 5.3: Novel Hierarchical Fusion Architecture

To train the higher-ranking classifier, training data were were first split randomly into three folds. For

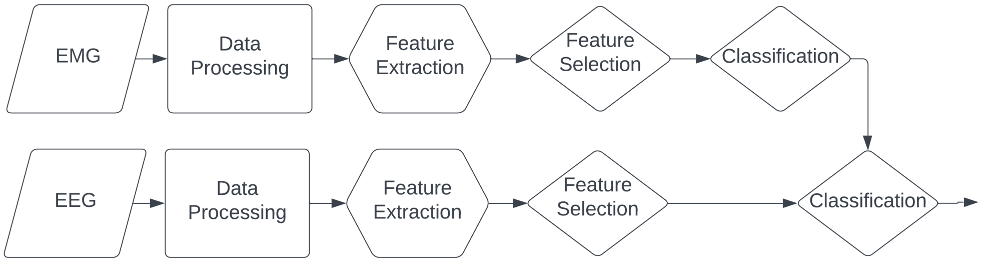


Figure 5.4: Novel “Inverse” Hierarchical Fusion Architecture

each fold K, the lower-ranking model was trained on data of its corresponding signal type in the non-K folds, and used to predict probability distributions for fold K. These probability distributions were then joined with their corresponding instances of the data modality belonging to the higher ranking model. After all folds had been used in this way, the higher-ranking modality's dataset was fully supplemented and could be used to train the higher-ranking model, with the lower-ranking model being then retrained on all folds of its full dataset.

5.3.1.3 Decision-level

The decision-level fusion architecture encompasses a range of methods which utilise predictions made by parallel independent EMG and EEG classifiers to arrive at a final decision, as illustrated in Figure 5.5.

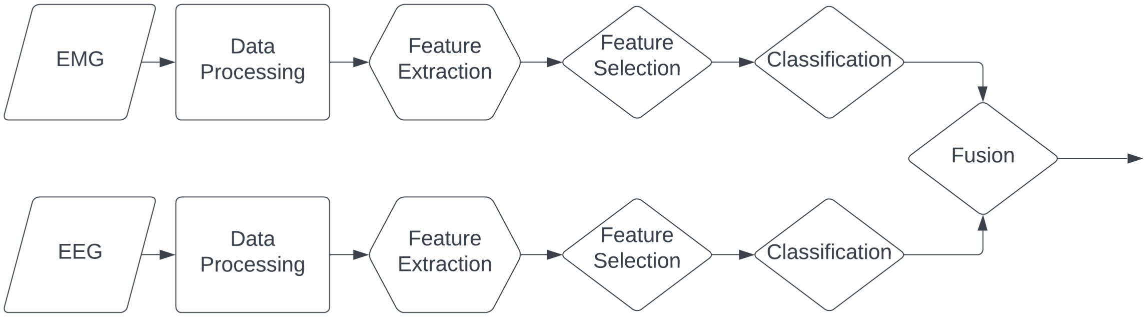


Figure 5.5: Decision-Level Fusion Architecture

By contrast to the “hands-off” approach to combination of Feature-Level Fusion, in this approach a system is forced to make class predictions using both data modalities, even if those predictions are not always considered equally in arriving at a final class decision. This may mitigate the risk of a more complex datatype being unduly ignored during training, depending on the strategy used to combine the two component models' predictions. In cases where their errors are likely to be statistically independent of one another, consensus of the components can imply confidence in their predictions. Should their errors overlap, however, this assumption breaks down as they would be likely to misclassify the same datapoints [255]; instead a more effective fusion technique would draw on one modality at points in the modelling space where the other is known to be less reliable. In the gesture classification context such dependence is difficult to predict — while the unique properties and differing complexities of EMG and EEG likely cause some independent errors, others may share a common cause, such as unexpected variations in the movements performed by a subject. It is hence worthwhile to explore a range of potential strategies for the component classifiers' combination.

The particular methods used in this work are outlined in Table 5.1 and fall under two schools. The first use rule-based techniques to merge the lower-level models' generated distributions into a final system-level decision.A number of studies in the biosignal literature use simple boolean logic to combine classifier outputs – logical AND and OR rules [128, 141, 256] – but such methods are naturally more suited to binary classification tasks [144] such as the detection of movement onset i.e. "move-vs-rest". Applying these strategies to a multiclass problem as in this work would require additional computation steps, such as a layer of one-vs-one ensemble voting; an added complexity thought unlikely to provide sufficient benefit to merit exploration. A more suitable strategy is the averaging of models' predicted probability distributions to form a fused distribution, as in Equation 5.1. The argmax of the fused distribution can then be used to identify the most likely class label.

$$P(y_{fused}) = \frac{P(y_{EMG}) * w_{EMG} + P(y_{EEG}) * w_{EEG}}{2} \tag{5.1}$$

Studies as early as Leeb et al.'s 2010 paper calculated fused classwise probabilities as the arithmetic mean of their constituent EMG & EEG models [146]; multiple subsequent studies have trialled both similarly equal weightings and various alternative distributions [142, 145, 150, 169, 257]. Here, after [141] among others, the Mean was trialled alongside a fixed weighting in favour of each data type: EMG predictions being weighted at 0.75 and EEG at 0.25, and vice versa. Leveraging the opportunity of the CASH optimisation routine, an additional Weighted Average variant was included wherein the distribution of weights over data modalities was itself a hyperparameter that could be tuned. The final rule-based decision fusion method included here was simply the Maximum Rule seen in works such as [145]. This rewards confident models: for each instance, the prediction of the data modality whose estimated probability distribution assigned the greater probability to its winner class was used.

The second family of decision fusion methods use an approach akin to stacked generalisation; one of a range of meta-classifiers is provided the probability distributions with respect to class produced by the lower-level models, and uses these distributions to predict the subject's gesture. This approach is underexplored among biosignal fusion literature, though was attempted by Cui et al. [145]'s work on classifying lower-limb movements as discussed in 3.1.3.2. Here, two candidate linear meta-models (a Support Vector Machine with a linear kernel, and a Linear Discriminant Analysis classifier), and one nonlinear model (a Random Forest) were trialled. To train these meta-models, a system's training data were split into three random folds. For each fold K in turn, data of the non-K folds were used to train base EMG & EEG classifiers, which were then used to make predictions of fold K. After all folds had been predicted in this way the resultant probability distributions were used to train the meta-model, and the base component classifiers were subsequently retrained on all folds of their respective full training datasets. In testing such a system EMG & EEG datapoints were first provided to the low-level models, which generated probability distributions

Algorithm	Description
Mean	Arithmetic mean of EMG and EEG probability distributions
Fixed Weighting favouring EMG	Weighted average of EMG & EEG with weights (0.75,0.25)
Fixed Weighting favouring EEG	Weighted average of EMG & EEG with weights (0.25,0.75)
Fixed Weighting Tunable	Weighted average of EMG & EEG distributions wherein the weight assigned to EEG is a tunable hyperparameter
Maximum	Distribution with the highest probability in its respective highest-scoring class is selected
SVM Stacking	Support Vector Machine used to produce a decision based on probability distributions produced by EMG and EEG classifiers
LDA Stacking	Linear Discriminant Analysis model used to produce a decision based on probability distributions produced by EMG and EEG classifiers
RF Stacking	Random Forest model used to produce a decision based on probability distributions produced by EMG and EEG classifiers

Table 5.1: Summary of candidate Decision-Level Fusion algorithms

accordingly to be supplied to the meta-model.

In all cases, the decision-fusion algorithm outputs a probability distribution estimating the likelihood of a given instance belonging to each of the various gesture classes; the aystem's final classification decision is determined by the argmax of this distribution.

5.3.1.4 Single-Modality Baselines

As points of reference for later comparison, Bespoke and Generalist systems of two additional architectures were developed, consisting simply of a single classifier using either EMG or EEG data on a unimodal basis. This work is exploratory and seeks to identify suitable approaches for fusing these data, not to prove definitively whether such fusion is universally guaranteed to outperform single-mode systems in gesture classification. Such baselines were included however to aid in reviewing systems' performance and judging the merit of the multimodal fusion strategies proposed. Their respective CASH optimisation procedures may also reveal insights regarding suitable configuration choices for future unimodal systems.

5.3.2 Feature Engineering

5.3.2.1 Feature Scaling

Following removal of any identifying attributes and the target class label, featuresets were standardised using scikit-learn's *StandardScaler* such that all features had a mean of zero and a unitary standard deviation. Initial exploratory work had trialled normalisation of features to the range (0,1), standardisation as described, and the absence of any scaling at all, finding minimal apparent difference in the informativity of the resultant data; standardisation was chosen as the scaling method here to ensure compatibility with those machine learning algorithms which assume such a distribution in their training data.

Crucially, for any given modelling process, standardisation was itself performed on only the training data.

The same transformation — that is, a transformation wherein the mean and standard deviation had been computed from the training data — was then applied to the test data, to protect against data leakage. It is acknowledged that, should unseen test data be particularly anomalous from the training data, this approach may risk test data being scaled to values outside the range encountered by a trained model. This is however necessary to ensure strict separation of train and test data to protect against data leakage [191], and to remain analogous to a deployed system of either the Bespoke or Generalist nature. In both cases a deployed system would be trained, on the users' data or that of a wider population respectively, prior to use; it would not be possible to standardise on the join of training and test data. Neither would it be appropriate or viable to standardise test data within itself in real-time. New data samples would be generated continuously, the presence of which could affect the consistency of attempted real-time scaling from sample to sample as the distribution of the total set of test data changed. Such a method would also require a significant cache of historical test data to be accrued over time. More viable would be to apply a single pre-established transformation to incoming subject data, as is emulated here. While bioelectric signal characteristics can vary between individuals, their broad properties discussed in Chapter 2 are consistent at least in orders of magnitude; it is hence likely that statistical features derived thereof would also be likely consistent in scale. Thus it is not anticipated that the risk of novel test data being scaled to extreme outlier values is high enough to be likely to present significant issues, other than perhaps in situations such as sensor breakdown wherein degradation of system performance would already be expected.

5.3.2.2 Feature Selection

The feature extraction process described in 4.3.2 above results in a total number of features which scales more than linearly with the number of raw signals in the dataset; while a fixed number of attributes are extracted from each individual raw signal, the size of the covariance matrix scales quadratically with the number of signals that form its input vector. To enable efficient modelling and avoid overfitting, features needed to be reduced to those most informative with respect to the gesture class. While the particular strengths and limitations of various feature selection methods are not a specific research focus of this work, the decision was made to take a univariate selection of features on the basis of ranking their individual predictive power over the class to enable later assessment (in 5.5.7) of the "popularity" of such features — the frequency at which they were found informative. While transformational methods of feature selection are somewhat more common choices in biosignal research [258], univariate approaches are far from unprecedented; Tryon & Trejos [142], who also sought to characterise the frequency with which features were selected across multiple subjects, used a ranking-based approach (in their case the ReliefF [259]) and indeed found it to be superior to both the Maximum-Relevance-Minimum-Redundancy (MRMR) and Principal Component Analysis techniques.

In this work six channels of EMG data were used, resulting in a featureset 588 attributes wide. Univariate feature selection was used to reduce this to a size more appropriate for modelling; the 15% of features ranked most highly by a one-way ANOVA between feature and class were selected (as implemented with *scikit-learn*'s *SelectPercentile* and *f_ classif* functions) and the remainder discarded, forming a set of 88 features. The choice of 15% as the threshold here was largely arbitrary, with no particular motivation other than being a convenient figure for arithmetic and resulting in a vector of fewer than a hundred features.

The EEG data carried information from twenty individual sensors, more than double the EMG, and hence generated an ensemble of many more features. Selecting from these on the same basis, the highest-ranked 15%, would result in a still unviable number and in EEG features outnumbering EMG by a factor of at least 4. It was additionally anticipated that due to inherent mathematical relationships between some of the features, and the close proximity of EEG electrodes to one another, there may be a high degree of correlation between some EEG features. This could cause a simple feature reduction method to select features with high levels of mutual information, and hence risk reducing the breadth of total information captured.Indeed, exploratory work with univariate reduction indicated a tendency of unimodal EEG models to overfit dramatically, which was surmised to be in part due to a narrower than anticipated breadth of information carried in the featureset. EEG features were hence instead reduced by an L1-norm based selection using *scikit-learn*'s *SelectFromModel* and *LinearSVC*. A linear Support Vector Machine was trained on the data, with a regularisation parameter C of 0.005 and the L1-norm used for penalisation, so as to result in sparsely assigned coefficients (i.e. to encourage many features to be zero-weighted). Thereafter, a fixed number of those features with nonzero coefficients were retained, prioritising those with the greatest coefficients. The training split of each Bespoke system was 1608 samples long, 67% of the total number of samples belonging to a given participant (2400). In Bespoke systems EEG were hence reduced to 40 attributes, the approximate square root of the number of training samples. In a Generalist system, for each subject N under test, the training set comprised all data from the 19 non-N subjects and was hence a factor of at least 19 larger than that of a bespoke system, at 45,600 samples. The square root of this sample length, 214, would be an infeasibly large number of attributes to retain for effective modelling. In Generalist systems EEG were thus reduced to a feature vector of consistent width with EMG, at 88 features.

To ensure parity of available information between fusion architectures, the merged featureset of Feature-level Fusion systems was of equivalent size to the sum of the distinct EMG & EEG featuresets seen in the Decision-level and Hierarchical fusion approaches. In a Bespoke Feature-level Fusion system the joint set thus totalled $(88 + 40 =)128$ features, and in a Generalist $(88 + 88 =)176$. For the Separate Selection subtype 88 EMG features and 40 or 88 EEG features (for a Bespoke or Generalist respectively) were selected from the modalities independently, and these two selections subsequently joined as described in 5.3.1.1. In the Joint Selection variant, these 128 or 176 attributes were selected from the combined set of all EMG and EEG features, using the L1-norm based method outlined above.

As outlined in 5.3.2.1 in relation to feature scaling, to preserve separation between training and testing data the selection of features was in every case performed using only the training set for that particular system, and the test set reduced to the identified array of features. Bespoke systems hence selected features from the 67% of subject data being used for training in a given modelling procedure. This both for each of the the 20 Development Set subjects at each iteration of a CASH optimisation routine, and for each of the 5 Holdout subjects when later verifying generalisability of architectures' optimal systems. In the Generalist case where all data from subjects other than the one under test are used for modelling, throughout the optimisation process the features selected for a subject N were done so based on the data from the 19 non-

N subjects. For validation of Generalist systems, where an architecture's optimiser-identified system was retrained on all 20 Development Set subjects to be tested on each of the 5 Holdout Subjects in turn, features were correspondingly selected using the data from all 20 Development Subjects.

5.3.3 Classification

Each Fusion architecture proposed in 5.3.1 comprises at least one classifier for which a machine learning algorithm must be selected and appropriate hyperparameters identified. A range of machine learning models outlined below, all implemented in *scikit-learn* [260], were selected as candidates based on those with precedent in the literature. As noted in 5.2.1, model-level hyperparameters such as the k in a $k-$ Nearest Neighbours classifier are relevant to system performance only where their associated model is selected; their existence is conditional on the choice of algorithm. The hyperparameter space can hence be best described by a tree structure, as illustrated in Figure 5.6[2].

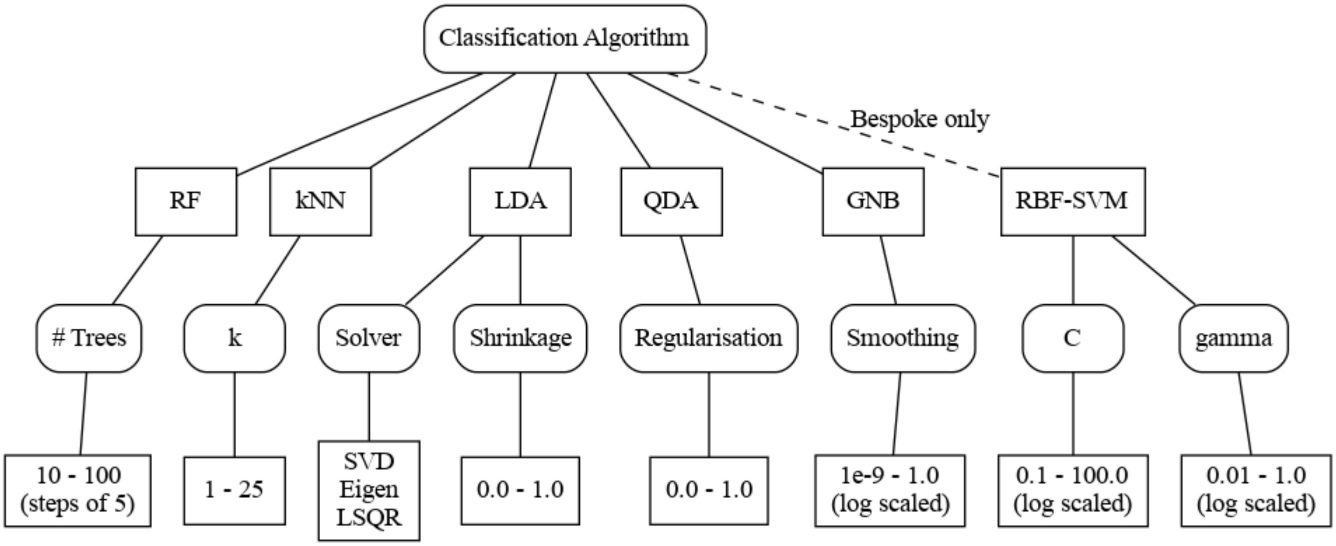


Figure 5.6: Subsection of the joint algorithm-hyperparameter search space describing the nature of a single classifier. EMG and EEG models (whether single unimodal classifiers or components of a multimodal system), and the single models of Feature-Level Fusion systems, each held their own unique but equivalent instances of this hyperparameter tree within a system's overall configuration space. Note that many hyperparameters are conditional & hence only created where the "Classifier" choice is of a given value.

For systems implementing Feature-Level Fusion, and the single-mode EMG and EEG baselines, a single classifier is selected, tuned, and used to classify EMG, EEG, or merged data and thus the tree in Figure 5.6 depicts the full extent of the hyperparameter search space. Hierarchical, Inverse Hierarchical, or Decision-Level fusion systems however comprise two distinct classifiers, one each for EMG and EEG data[3]. The constraints of the various hyperparameters were identical for EMG, EEG, and indeed Feature-Fusion classifiers; the same candidate models and range of value choices were explored in each. It should be stressed however that the

[2] Styled after that of [261].

[3] Strictly, in the Hierarchical and Inverse Hierarchical architectures one classifier is used for EMG or EEG data, and the other for the remaining data type supplemented by predictions from the first, as described in 5.3.1 above.

search space itself was not shared between modalities, nor was any learning in the space shared between fusion architectures, or between Bespoke and Generalist systems of the same architecture. Hyperparameters well-suited for EMG-based classification may not result in models which exhibit good performance with EEG data and vice versa, thus for each distinct classifier in a given fusion system a separate instance of the depicted hyperparameter tree was included within the overall configuration space. In Decision-Level fusion systems the search space additionally contained hyperparameters used to configure the Decision Fusion algorithm, as discussed further in 5.3.3.7 below.

It should be noted that the candidate algorithms included, outlined below, are not necessarily an exhaustive representation of all schools of machine learning classifier and are not intended to be so. Notably, only classical models were considered as candidates for selection; while deep learning has been used successfully in biosignal research [216], classical models show sufficient promise and precedent to merit continued exploration and the literature does not indicate a firm trend for deep learning models consistently providing performance sufficiently superior as to justify their costs. While reviews such as [262] highlight the emerging topic of “big data” among biosignal research and the suitability of deep learning to such experiments, they also highlight the necessity of large datasets for training deep networks without overfit — even the work of Aly et al. [129], who applied deep learning successfully in the fusion of biosignal data, found deep models with more than just two layers to overfit. Systems in this work, particularly the Bespoke type, are by contrast of limited dataset size — and experiments in subsequent chapters seek to reduce training data further in the interests of minimising the burden to users. Lotte et al. [152] corroborate this apparent non-necessity of Deep Learning models, finding that while popular they do not appear to present notable benefit over alternatives for classifying EEG data, and that their high complexity can lead to excessively long training times. Dolopikos [223] further found their Deep Neural Network was beaten not only by a voting ensemble of classical models in classifying EMG data, but also regularly outperformed by individual Random Forests and Support Vector Machines. Shallower Artificial Neural Networks such as Multilayer Perceptrons have also seen some use in BCIs, but similarly do not appear to be routinely worth their additional cost in terms of training, tuning, and prediction time. Kuzborskij et al. [38] found MLPs not to offer routinely better performance than Support Vector Machines (using the Radial Basis Function kernel) on the domain-standard NinaPro EMG dataset [226], and to both be slower and require more intricate hyperparameter tuning.Garrett et al. [263] found ANNs to marginally outperform LDAs in the classification of abstract mental tasks from EEG data, but both to be weaker than SVMs, and to be slow and more computationally intensive than the LDA [264]. Hargrove et al. [265] found MLPs to perform no better than LDAs for EMG classification, & Englehart et al. [242] found LDAs to outperform MLPs with EMG datasets comprising features in both the time and frequency domains; as described in 4.3.2 above, in this work both time and frequency domain features were indeed computed from raw biosignal data. Scheme & Englehart [14] observed ANNs to be of neglibible benefit over kNNs and SVMs and equivalent to or weaker than LDAs and QDAs in classifying 7 gestures from EMG data with both able-bodied and amputee subjects. Gandolla et al. 2017 [266] used two cascaded ANNs to successfully identify a pinch, cylindrical grasp, and closed fist from EMG data at an average of 76% accuracy, though they noted a significant fall in accuracy, by an average of 13%, after electrode removal and repositioning – suggesting perhaps a sensitivity to drift in the data.

Similarly, some techniques more dissimilar to those included for investigation, such as recently emerging approaches using Convolutional Neural Networks to classify image representations of biosignal features as explored by Qi et al. [267], Ashford et al. [268], and Tryon & Trejos [143] among others, are considered outside the scope of this work. Indeed, Qi et al. note that "*inter-subject differences [exceed] the generalization ability of CNN[s]*" [267], suggesting an unsuitability for application to the subject-independent systems of interest to this work, and Tryon and Trejos' work while demonstrating the potential of CNNs in classifying multimodal biosignal data did not find the approach to offer significantly more accurate predictions than EMG classifiers alone.

5.3.3.1 Random Forest

Random Forests [269] are a popular nonlinear classification algorithm which aggregate the predictions of an ensemble of randomised Decision Trees. They see prominent use in a wide range of applications [270] and have much precedent in classification of both EMG and EEG data [139, 145, 147, 195, 211, 271], including in Feature-Level fusion [139], and have in some cases been found to outperform both classical and deep learning models in the classification of EMG data [145], to be capable of outperforming the widely-used Linear Discriminant Analysis model in classifying Kinaesthetic Motor Imagery from EEG data [211], and to be robust to the injection of mis-labelled instances in the classification of Chinese Sign Language gestures from EMG and accelerometer data [271].

The number of decision trees in a Random Forest frequently has an impact on its ability to generalise to unseen data, with larger forests typically providing better generalisation performance [269] by virtue of the greater diversity in the constituent trees (though greater randomness in a forest may not alone guarantee a defence against overfit [272]). It has however been shown that, in addition to increasing computational complexity, increasing the number of trees in a forest is not guaranteed to offer improved classification performance and in fact many datasets will reach a saturation point beyond which the inclusion of further trees has minimal impact on a forest's predictive power, even where the numbers of instances or attributes of the dataset are high [273].
While various biosignal studies which make use of Random Forests provide no details on the choice of forest size [139, 145, 195], Pritchard et al. [147] used the default value in the Waikato Environment for Knowledge Analysis (WEKA) toolbox [274, 275] of 100 trees with some success for EMG & EEG classification. Steyrl et al. [211] investigated this hyperparameter in depth in the context of EEG-BCIs classifying a binary KMI problem, trialling forest sizes of 10 to 5000 trees while also varying the number of features considered as splitting candidates (referred to as "data dimensions at each node" in Steyrl et al.'s work). Their findings suggest diminishing returns from the inclusion of additional trees over 100, where more than 10 features are considered per node — with the benefit of increasing forest size becoming less significant as the number of candidate features increased, consistent with the claims of [273].

Here the number of decision trees in a given Random Forest was a conditional hyperparameter, whose existence was conditional on a Random Forest being used (i.e. it was present only when a Random Forest was chosen for classification), with possible values uniformly distributed between 10 & 100 trees, quantised in discrete steps of 5.

The maximum depth of trees in each forest was 5 nodes, following from early preliminary experimentation. In accordance with the default configuration of *scikit-learn*'s *RandomForestClassifier*, each tree was fit on a number of samples (drawn with replacement) equal to the length of the training set, splits were considered on the basis of minimising Gini impurity, and at each juncture a random number of features equal to the square root of the total number of features were considered as splitting candidates — which is not only the the *scikit-learn* default method of calculating the number of features to consider but has precedent in [211] among others. In EMG and Generalist EEG this computes as $\sqrt{88} \approx 9$, in Bespoke EEG as $\sqrt{40} \approx 6$, in Bespoke feature-fusion $\sqrt{128} \approx 11$, and in Generalist feature-fusion $\sqrt{176} \approx 13$.[4] Nodes were split only where they contained at least 2 samples, and where at least 1 sample would be left on each branch, again in accordance with the default *scikit-learn* implementation. Predicted classwise probabilities of the Forest were computed as the mean of the classwise probability distributions of each of its trees.

5.3.3.2 k - Nearest Neighbours

The k-Nearest Neighbours classifier [276] is a transductive, non-parametric model which classifies data by identifying the k training datapoints closest in the feature space to the datum under test, and classifying the instance according to the labels of those k neighbours. While a kNN conventionally classifies by vote of the selected neighbours rather than probabilistically, to facilitate fusion, probability distributions with respect to class were obtained from the kNN models; these were reflective of the proportion of the neighbourhood voting for each class (all neighbours considered were weighted equally).

While a less popular choice in biosignal literature than competing models, kNNs have seen occasional use in classification of motor activity & KMI from EEG data [145, 277, 278], extensive use in classification of EMG data [38, 173, 189, 238, 240, 279], and have even been explored in studies which merge EMG & EEG at the feature level [137, 139].

The hyperparameter k is naturally of great signficance to the predictive power of a kNN model. Prior studies applying kNNs to biosignal classification do not indicate a consistently superior value of k – indeed intuitively the optimal k for a given problem is likely to be dependent on the properties of the data being modelled – and thus k was made a tunable hyperparameter for the CASH optimisation process to explore. Despite the variation in choices of k among literature, some values appear popular. Kim et al. [238] used a 5-NN classifier in tandem with a Gaussian Naive Bayes model to identify wrist movements from a single EMG electrode, though offered no rationale as to this choice of k (such unexplained parameter choices are not uncommon among the biosignal literature, something which this work strives to avoid). Benalcazar et al. [173] similarly used a 5-knn to classify EMG, justifying it as the closest larger odd integer to $log_2(30)$, 30 being the size of their training and testing datasets). Chen et al., in classifying Chinese number gestures opt for a 10 - neighbours classifier on the basis of their dataset comprising 50 samples of 10 gestures from 6 subjects [233]; though not explicitly stated, this k of 10 may be speculated to be derived with a similar juestification to

[4] See 5.3.2.2 for description of the featureset sizes in various fusion architectures.

that of Benalcazar et al., in that $log_2(50 \times 10 \times 6) = log_2(3000) = 11.56$. Schlögl [278] meanwhile found KNNs to be weaker than other models such as SVMs and LDAs in classifying Motor Imagery from EEG data, but that much higher k values in the order of 50 - 100 offered the best performance. Große Sundrup and Mombaur [279] remarkably found some success in classifying EMG data with $k = 1$, applying a unique approach wherein clusters of neighbours in the featurespace were replaced by a single representative of the neighbourhood, & calculating the distance to said neighbours with Dynamic Time Warping rather than more commonplace distance metrics, though noted that where neighbourhoods are not replaced by cluster representatives, the unitary k value meant distance may be measured to a singular point on the *edge* of a given group of neighbours rather than a representation of its central tendency (thus motivating their neighbourhood representative approach). Others such as Kuzborskij et al. [38], Pritchard et al. [147], and Kim et al.[5] [189] performed optimisation to select their k-values: Kim trialling ks of 1 through 10 neighbours, and settling again on 5, for classifying wrist movements; Pritchard trialling even-number ks between 2 and 20, finding $k = 2$ to be optimal for EMG gesture identification and $k = 12$ optimal for EEG mental state classification; and Kuzborskij trialling ks equal to 1 through 7.

Here, the search space for k was the integers 1 through 25, inclusive. In accordance with the default *scikit-learn* KNN implementation ("*KNeighborsClassifier*"), nearest neighbours were computed with a k-dimensional tree of leaf size 30, and the distance was measured by the Minkowski metric of the 2nd order, i.e. the Euclidean distance [280].

5.3.3.3 Linear Discriminant Analysis

Discriminant Analysis classifiers operate by determining decision boundaries which separate data in a multi-dimensional featurespace (i.e. one where each dimension is one feature of the dataset) in a way that seeks to maximise distance between classes [281]. The Linear Discriminant Analysis model in particular works under the assumptions of each class's feature likelihoods being normally distributed, and of the covariance matrices of all classes being identical [240, 281].

LDAs are well-established as a prevailing algorithm for EEG classification [30] and are likewise well precedented in the classification of EMG data [38, 237, 282, 283] including that of amputees [14, 228], and indeed in classifying motor activity from EEG & EMG data merged at the feature level [130].

Sickit-learn implements three methods, or "solvers", for computing classwise log-posteriors within LDAs: the Least Squares Solution (LSQR), Eigenvalue Decomposition (which optimises the ratio of distance between classes to variation within classes), and Singular Value Decomposition. While some such as Mohd Khairuddin et al. [237] have used the *scikit-learn* default hyperparameters to successfully classify EMG data, notably finding LDAs to both train on and predict much faster than competing algorithms, many other works provide no details on the particular implementation of the LDAs they utilise [228, 283]. Hence all three of these solvers were trialled in this work, as the permissible values of a categorical hyperparameter within the CASH optimisation search space.

While the SVD solver bypasses calculation of the covariance matrix, this is a necessary step in both the

[5] NB: Not the same Kim as that of the aforementioned [238]

Eigenvalue Decomposition and Least Squares Solution algorithms. Shrinkage can be applied to improve this estimation of covariance matrices, by combining multiple estimators [284]. In the *scikit-learn* implementation, applying zero shrinkage will result in the empirical covariance matrix being used, and with a shrinkage value of 1 it will be estimated wholly from the diagonal matrix of variances, with values between these corresponding to a proportional combination of these extremes determining the shrunk matrix [285]. Shrinkage is typically of particular importance in cases where the number of features in a dataset particularly outweighs the number of samples [198, 284]. While the feature selection processes outlined in 5.3.2.2 result in this not being the case in this work, the use of shrinkage by some biosignal literature [75], including that of Jeong et al. [198] by whom the dataset used in this work was originally collected, motivate exploring its influence here for completeness. Therefore where Eigenvalue or LSQR solvers were used, the shrinkage was also a hyperparameter, with a uniformly distributed search space between 0.0 and 1.0.

5.3.3.4 Quadratic Discriminant Analysis

The Quadratic Discriminant Analysis model is a version of a Discriminant Analysis classifier wherein decision boundaries are quadratic, rather than linear as in the LDA. While a somewhat less common model among biosignal literature, QDAs have been found competitive with KNNs and LDAs in classifying the direction of wrist movements from EMG [189], and to be more robust than LDAs in the classification of Chinese number gestures [240] again with EMG data; in the latter study they were described as preferable over the marginally better-performing SVMs on the basis of lesser model complexity. QDAs have also seen precedent in the classification of arm movements from feature-level fused EMG-EEG data by Aly et al. [137] and as component parts of decision-level fusion systems in classifying lower limb movements [145].

Unlike the LDA, in the QDA the covariance matrices of the classes are not assumed to be equal [281]. In *scikit-learn*'s implementation, the Singular Value Decomposition solver is used in all QDAs to enable estimation of per-class covariances while circumventing the need to explicitly compute the covariance matrix. In efforts to overcome the same challenges of features outweighing samples (the "singularity problem") as were addressable in Linear Discriminant Analysis by applying Shrinkage, in the QDA these per-class covariance estimates can be regularised by scaling them in a manner biased towards their diagonal elements [281]. As noted in the discussion of shrinkage in 5.3.3.3, the singularity problem is unlikely to arise in this work, however for completeness & to minimise *a priori* assumptions regarding modelling characteristics, the strength of this regularisation was an optimisable hyperparameter, with a uniformly distributed search space between 0 & 1.

5.3.3.5 Gaussian Naïve Bayes

The Gaussian Naïve Bayes (GNB) algorithm estimates the conditional probability of an instance belonging to a class, given its feature values, through an application of Bayes' theorem wherein features are naïvely assumed independent and their likelihoods with respect to the class assumed to have normal (Gaussian) distributions.

While GNB models see less frequent use in biosignal literature than some other included candidates, they have been used successfully in classifying both EMG & EEG data [1, 172, 233, 238]. Of particular interest, Bird & Pritchard et al. [1] found a GNB to offer the greatest predictive power for wholly unseen EMG data out of a range of popular algorithms, providing some motivation for its inclusion here in light of the interest in Generalist systems.

The assumption of feature independence is unlikely to hold true for the data in this work. Not only is the likelihood of crosstalk between raw recorded biosignals high, particularly in the case of EEG wherein electrodes are both in close proximity to one another and located at some distance from the neural point sources generating the electrical activity they measure, but from each raw bioelectric signal a range of statistical features were extracted (4.3.2); while the Feature Selection methods for EEG and joint EMG-EEG data make some effort to avoid selecting highly correlated features (5.3.2.2), some dependence remains likely. Nevertheless, GNBs are known to often perform at competitive levels even when the assumption of independence between features does not hold true; it is suggested that independence can be violated if dependences are distributed over classes, rather than class-specific [286].

While the GNB has no fundamental tunable hyperparameters, the implementation in *scikit-learn* allows for a "smoothing" of the gaussian distributions by which features' likelihoods are modelled, by adding some proportion X of the largest of all feature variances to the variance of every feature, in effect widening the distributions in the interests of stabilising internal calculations. This variance smoothing appears largely undiscussed in biosignal literature but was here made a hyperparameter within the optimisation search space. Given scikit-learn's default X value of 1×10^{-9}, and that the feature standardisation (5.3.2.1) ought to have resulted in normally distributed features, the search space for this variance smoothing factor X was defined as a logarithmic distribution between 1×10^{-9} & 1, such that potential values were more likely to be in similar orders of magnitude to 1×10^{-9}.

5.3.3.6 Support Vector Machine

Support Vector Machines [287] (SVMs) are natively binary classifiers which operate by seeking a linear decision boundary, called a hyperplane, with the maximal margin between itself and the closest datapoints to it of each class (those points being referred to as the eponymous "Support Vectors"). Their extension from binary classifiers to models capable of multiclass problems is reasonably trivial — *scikit-learn*'s implementation, itself based on that of *libsvm* [288], uses a one-vs-one scheme, i.e. a deconstruction of the problem into a number of binary problems, one for each possible pair of classes. They are also capable of forming decision boundaries which are nonlinear in the native feature space by projecting the data into a higher-dimensional space where it may be more linearly separable [38] and thus a hyperplane able to be found. Such transformation is often done by application of a kernel function to avoid the need for explicit mapping of the data — known as the "kernel trick" [289].

While other kernels have been variously used, including linear kernels for classifying the NinaPro EMG dataset's [195, 226] many distinct hand gestures by Kuzborskij et al. [38], and linear kernels with L1-penalisation for classifying ECoG data by Fujiwara et al. [290] , the Radial Basis Function (RBF) has been found less computationally expensive than other kernels [291] and dominates various uses of kernel functions among biosignal literature, to the point of being described the "*de facto* standard" [106]. Indeed RBF-SVMs have been routinely used in the classification of gestures from EMG [38,106,141,145,147,174,195,223,292,293] data, where they consistently prove among the most popular choices of classifier and have been noted as particularly highly-performing with data from amputees by comparison to competing models [195], and EEG data [141, 145, 159, 160, 291, 294], where they have also been applied with some degree of success in attempts to discriminate individual finger movements of the same hand [159, 160].

An SVM's regularisation hyperparameter C controls, in essence, the trade-off between the width of the margin formed between classes, and the level of accepted misclassification risk. This is sometimes referred to as the "hard-" or "soft-"ness of the margin; where a "hard" margin SVM (corresponding to a higher C-value) determines its discriminative hyperplane on the basis of only those datapoints of each class closest to that boundary, while a "soft"er margin (a lower C) would consider datapoints farther from the boundary (and potentially closer to the centre of a class's cluster in the featurespace) even where this may place some training datapoints on the "wrong side" of the hyperplane. The C hyperparameter determines the extent to which this distance between a such a datapoint and its correct margin boundary penalises the SVM. Where data is not linearly separable, even after projection to a high-dimensional space, tuning this penalty and hence the level of "slack" which is permitted can be of particular importance to balancing underfit and overfit in an SVM. The other key hyperparameter of an RBF-kernel SVM, γ, determines the radius of influence of any given training datapoint chosen as a support vector, with that radius being inversely proportional to γ — a lower γ encouraging a simpler hyperplane as the influences of many support vectors average out, and a higher γ encouraging a hyperplane very tightly fit to the support vectors & hence susceptible to overfit.

While many biosignal studies do not detail their selection of C and γ values [143] or simply note them as having been found through some unspecified optimisation [145, 160], among those that do give further detail there appears little consensus in either the scope of the space to be searched over or the optimal values themselves — suggesting perhaps a high specificity to the problem being modelled. Kuzborskij et al. [38] trialled C values of 2^0 to 2^{16} for EMG classification, optimising these on a per-subject basis. Yong et al., in the identificatation of KMI from EEG, trialled similarly large C values but significantly smaller values in addition, ranging from 2^{-15} to 2^{15} [294], again on a subject-specific basis. Others found success with less extreme values of C: Ameri et al. [292] found an optimal C of 0.2, & that higher values of C were not beneficial and led to longer prediction times. Pritchard et al. [147] found optimal Cs of 2 for EMG classification and 6 for EEG, though with a Linear kernel function rather than the RBF, and Tavakolan et al. [291] tested C values of $0 - 100$. Castellini et al. [174], notable for taking a subject-independent approach to EMG classification in their work, found optimal C values in the region of $10^{1.5}$.

With regard to γ, Kuzborksij et al. [38] searched over $\gamma = 2^{-16}$ to $\gamma = 2^{-2}$, and Yong et al. [294] used values of somewhat similar orders of magnitude, from $\gamma = 2^{-15}$ to $\gamma = 2^3$; Dolopikos et al. similarly found lower γ values superior, with an optimal γ of 10^{-5} [223]. Castellini et al.'s Generalist approach to EMG SVMs found the optimal value of γ to be typically around $10^{-0.5}$ [174], and Tavakolan et al. [291] trialled γs of a similar magnitude, ranging from 0 to 3. Garrett et al. [263] trialled SVMs with γ[6] values of 0.5, 1, and 2, finding γ=0.5 to provide the best classification results. Interestingly Ameri et al. [292], rather than

[6] Referred to in [263] as σ.

simply optimising over a range of fixed values, determined γ as $\frac{1}{N_{features}}$. In the systems investigated here, this formula would equate to $\frac{1}{88} = 0.0114$ for a Bespoke EMG-SVM and $\frac{1}{40} = 0.025$ for a Bespoke EEG-SVM.

In this work Support Vector Machine classifiers using the Radial Basis Function kernel, implemented with *scikit-learn*'s *SVC*, were trialled with the hyperparameter C having a search space logarithmically scaled from 0.1 to 100, & the Kernel coefficient Gamma as a hyperparameter with values logarithmically scaled from 0.01 to 1, clustering values at the lower end of the scale and thus closer to $\frac{1}{N_{features}}$. For computational feasibility, ties were broken by arbitrary selection of the first listed class of tied classes, in accordance with the default configuration in *scikit-learn*.

It should be noted that SVMs were included as a candidate model only in Bespoke systems. The training time for the SVMs used here increases dramatically with greater numbers of training samples & the documentation of scikit-learn itself states that the *SVC*'s computation time scales quadratically with the size of a dataset, further suggesting the model may be unviable where training data exceeds >10,000 samples. In Bespoke systems the traning dataset is 1600 samples long which is viable, however in Generalist system it is in the order of 45,000 (and would be higher still where models were retrained on all 20 development set subjects for use predicting the holdout data); early exploratory work made it evident this high dataset length did indeed make the SVM unviable for Generalist systems.

Coercion of probabilities from SVMs

As described above, SVMs operate on the basis of constructing decision boundaries are not inherently probabilistic; they do not natively lend themselves to producing probability distributions with respect to class. Many of the fusion strategies explored in this work however, particularly the Decision-level and Hierarchical algorithms (5.3.1) assume that class probabilities will be provided. To coerce probability distributions from SVMs, SVM models were wrapped in *scikit-learn*'s *CalibratedClassifierCV*. Using a 5-fold cross-validation process, five copies of the base SVM were constructed and their outputs calibrated by fitting a logistic regression model, a method known as Platt's Scaling [149] used for this purpose in [145, 295] among others. When used to predict new data, the "SVM classifier"'s resultant overall probability distribution was calculated as the average of the probabilities estimated by these five calibrated base models.

5.3.3.7 Tunable Decision-Level Fusion Algorithms

In Decision-level Fusion systems, an additional dimension of the hyperparameter configuration space was the choice of fusion algorithm between the candidate algorithms in Table 5.1. Those classification-based algorithms which incorporate stacking techniques to the system each have their own conditional hyperparameters, which determine the nature of the meta-model. One rule-based algorithm also made use of a conditional hyperparameter, to allow the weighting of a weighted average to be determined by the optimisation procedure. The search space for decision-fusion hyperparameters is presented in Figure 5.7 & outlined below.

Parameter-Weighted Average

Under this rule-based decision fusion method, the weighting w_{EEG} given to probability distributions predicted

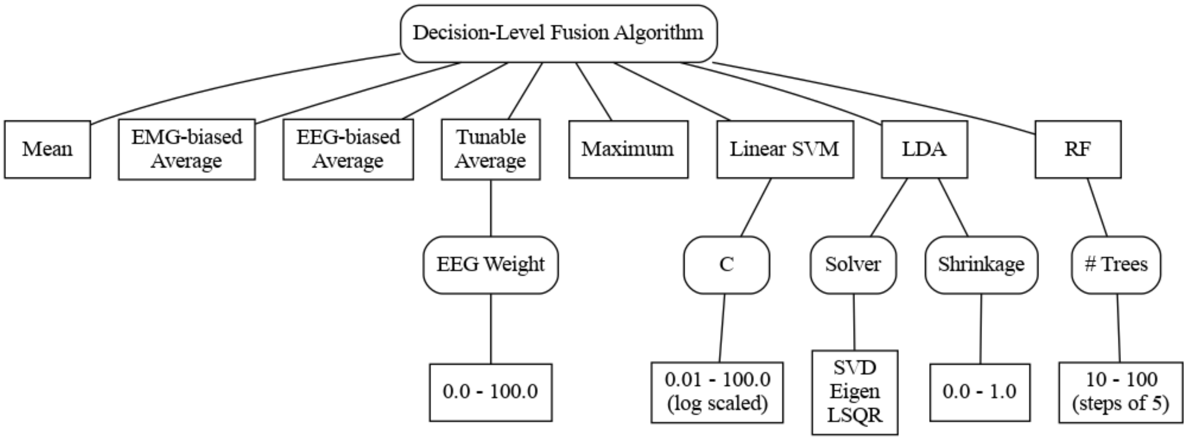


Figure 5.7: Subsection of the hyperparameter search space describing the algorithm used for Decision-Level Fusion. Note that this branch of the hyperparameter tree was only extant in optimisation of Decision-Level Fusion systems.

by the EEG component classifier was a conditional hyperparameter with a uniform distribution between 0.0 and 100.0, expressed as a percentage. EMG predictions were in each case weighted by w_{EMG}, calculated as the complement of the EEG weighting.

Linear Support Vector Machine

The Linear Support Vector Machine metamodel was implemented with *scikit-learn*'s *LinearSVC*; this uses the *liblinear* [296] library internally, which is described as better able to scale to large datasets than the *libsvm*-based *SVC* [260]. This was a motivating factor in the use of a Linear SVM as a candidate metamodel, as it was viable for use in both Bespoke and Generalist systems. The regularisation parameter C had a logarithmically distributed search space between 0.01 and 100, capable of handling lower values than the RBF-SVMs of component classifiers due to the simpler kernel. Default *scikit-learn* values were used for other parameters — L2 norm penalisation and a loss function defined as the square of the hinge loss — except that, as recommended by *scikit-learn* where the number of samples outweighs the number of features, the algorithm was tasked to solve the primal problem rather than the dual.

Linear Discriminant Analysis

Linear Discriminant Analysis metamodels were given an equivalent search space over which to optimise as that of the LDAs used for EMG or EEG classification; the Singular Value Decompositon (SVD), Eigenvalue Decomposition, and Least Squares Solution (LSQR) algorithms were all candidate solvers, and where Eigenvalue or LSQR solvers were used the shrinkage was also a hyperparameter, with a uniformly distributed search space between 0.0 and 1.0.

Random Forest

As in the case of LDAs, the search space for Random Forest metamodels was the same as that of Random Forest component classifiers described above: the number of trees was a hyperparameter with values distributed between 10 & 100 in quantised steps of 5, the maximum tree depth was 5 nodes, each tree was fit on all samples, splits were considered on the basis of minimising Gini impurity, and at each juncture the square root of the total number of features were considered as splitting candidates — which here, where the only two features were the probability distributions supplied by each lower-level model, corresponds to both features being considered.

5.3.4 Defining a "Literature-Informed" system

Aim 5.3 seeks to assess whether the Combined Algorithm Selection & Hyperparameter optimisation pipeline presented here was a helpful tool with which to design a suitable multimodal system. To do so it is necessary as described in 5.2.1 above to compare the "winner" optimiser-identified Bespoke & Generalist systems to ones derived non-algorithmically; that is, the best-informed decisions which could be gleaned solely from literature without performing such an optimisation.

The literature on which one could draw in the identifying of such a system is sparse, and of little unanimity. As discussed in Chapter 3, among research conducted into multimodal biosignal classification it is relatively rare that a range of fusion approaches are evaluated on the same problem with the same data (indeed this is a limitation of much of the wider biogsignal literature [30]), thus making the art of establishing a "best" candidate more difficult. This is as noted central to the motivation for applying CASH optimisation in this work, under the principle that even within a given domain such as biosignal gesture classification, no single model could be expected to prove consistently the most suitable over the totality of all possible unique problems & datasets [252]. Additionally, 3.2 noted that particularly among those works which do investigate multiple fusion strategies [128,142,145] few studies seek to classify between multiple gestures performed by the same limb, and among those that do fewer still distinguish subtly different movements of the same appendage, often opting instead to define broader movement categories such as flexion & extension of the elbow. They are thus quite dissimilar to the nature of the multi-grasp classification task being studied here. Some inferences can be made from works which use EMG or EEG data alone to approach problems more similar to that of this research. Such studies have been influential in determining the range of modelling options for optimisation as described in 5.3.3. It is not necessarily a safe assumption however that approaches which can offer best-in-class performance for unimodal systems will be best suited to being a constituent part of a multimodal system; while useful in establishing candidates, any inferences of absolute superiority obtainable from such works will be limited.

A number of studies fusing data at the feature level [129,130,137] make use of the dataset collected by Li et al. [130] corresponding to opening and closing of the hand and pronation and supination of the wrist. Such gestures, though much coarser than the three distinct grasps used in this work, are self-evidently relevant to the context of prosthesis control and so may suggest these works as good sources of design decisions.Works which both attempt feature-level fusion and compare it to other approaches, however, tend to find that feature-level fusion while successful was inferior to other strategies, such as boolean combinations of binary classifier outputs (albeit detecting only the presence or absence of motor activity) as in Gordleeva [128], or computing weighted averages of classifier probabilities as in Tryon et al. [142]. Thus while Feature-Level fusion is worthy of its inclusion in this work it does appear the strongest candidate as can be inferred from the literature alone. In others which investigate multiple fusion strategies, such as Cui et al. [145], the intensity of a movement rather than its nature defines the classes. Despite this dissimilarity in problem, Cui's work is a rare case wherein not only the rule- or weighting- based decision fusion algorithms as in Tryon [142] and Gordleeva [128] are applied, but also stacking techniques with meta-classifiers processing the EMG- & EEG-derived predictions; indeed these were found superior.

Ozdenici et al.'s study [113] was recognised in Chapter 3 as one of few which does indeed aim to distinguish between grasp types of the same hand, but falls under that set of approaches wherein EEG & EMG are used in dissimilar ways; they propose a cascaded system with EEG is utilised in movement onset detection and in identifying which of a subject's hands is moving, but the subsequent classification between specific gestures relying on EMG alone. Thus despite the similarity of classification problem, their approach is sufficiently distinct to those explored in this study as to be of limited benefit in informing system design here.

Cui et al.'s aforementioned finding of an SVM-meta-model as the superior strategy for subject-specific EMG-EEG fusion [145] motivates its selection as the "Literature-Informed" Bespoke fusion algorithm here. While Cui states that hyperparameters of some models they tested were determined through cross-validation, no further details are provided beyond the use of the RBF kernel and of Platt's scaling for eliciting probability distributions from the SVM (see 5.3.3.6). Thus the *scikit-learn* default values for the hyperparameters C and γ are used, of 1.0 and 1/(n_features) respectively. Both Cui et al.'s work & Tryon et al.'s 2021 study [142] indicate a preference for RBF-SVMs in the component EMG classifier of their subject-specific Decision-Fusion models, though again with no further information provided on their hyperparameter choices, the library defaults must be fallen back upon.

Tryon et al.'s earlier study [141] however, while only distinguishing motion from rest, is particularly notable as one of few which both attempt fusion in a cross-subject manner and in doing so compare a wide range of possible fusion algorithms. Their work provides otherwise rare insight as to suitable design choices for a subject-independent system; they found the Mean rule to offer the greatest accuracy & it is thus used for the "Literature-Informed" Generalist system here. Similar to their aforementioned subject-specific work, the component EMG classifier of Tryon et al.'s Generalist was likewise an SVM. While no hyperparameter details are provided they are described as being implemented with *MATLAB*'s *Statistics and Machine Learning Toolbox*; consulting the documentation of this toolbox [297]suggests the Linear kernel is chosen by default and thus the EMG model within the "Literature-Informed" Generalist system is here a Linear SVM. With no information on the value of C it is again left to the default in *scikit-learn* of 1.0.

For systems' constituent EEG models, notwithstanding the reservations noted above over the extent to which high unimodal classification accuracy may or may not be necessarily an indicator of suitability as a fusion component, to dismiss LDAs' overwhelming popularity in EEG classification [30] would be to wilfully disregard much of the literature rather than to draw on it. EEG models in both Bespoke and Generalist cases were hence defined as LDAs using Singular Value Decomposition, the *scikit-learn* default solver.

The Literature-Informed Default systems are therefore defined in full as in Table 5.2.

Design Element	Implementation	Source
Bespoke		
Fusion Algorithm	SVM-metamodel (Platt's scaling)	[145]
	RBF Kernel	[145]
	C = 1.0	*scikit-learn* 0.24.2 default
	$\gamma = \frac{1}{n\_features}$	*scikit-learn* 0.24.2 default
Component EMG classifier	SVM (Platt's scaling)	[142,145]
	RBF Kernel	[142,145]
	C = 1.0	*scikit-learn* 0.24.2 default
	$\gamma = \frac{1}{n\_features}$	*scikit-learn* 0.24.2 default
Component EEG classifier	LDA	[30]
	SVD Solver	*scikit-learn* 0.24.2 default
Generalist		
Fusion Algorithm	Mean	[141]
Component EMG classifier	SVM (Platt's scaling)	[141]
	Linear Kernel	*MATLAB Statistics and Machine Learning Toolbox* default [297]
	C = 1.0	*scikit-learn* 0.24.2 default
Component EEG classifier	LDA	[30]
	SVD Solver	*scikit-learn* 0.24.2 default

Table 5.2: Literature-Derived Baselines

5.4 Test Procedure Overview

The overall testing procedure by which gesture classification systems were developed, assessed, and compared, described in further detail throughout Section 5.5, is summarised as follows:

1. CASH optimisation routines, each afforded a budget of 100 iterations, were performed for Bespoke and Generalist settings of every Fusion architecture and both Unimodal EMG & EEG systems using Development Set data, with mean per-participant accuracy as the objective function.

2. For each CASH optimisation routine, the configuration of algorithm choice(s) & hyperparameter values which maximised mean per-participant accuracy was determined as the "canonical" set of modelling choices for that respective Fusion or Unimodal architecture in that setting.

3. The Fusion and Unimodal architectures whose canonical configurations had achieved the highest predictive power on the Development Set were chosen to be the "candidate" Fusion and Unimodal systems for each setting.

4. The candidate Fusion and Unimodal systems of each setting were compared on their ability to predict the unseen Holdout Set data, using a paired difference test over the 5 Holdout Set subjects.

   - In the Bespoke setting these were modelled on a subject-specific basis (with a 67/33 train-test split) for each Holdout subject.
   - In the Generalist setting these were modelled on all 20 Development subjects' data, and used to predict the data of each Holdout subject.

5. Baselines derived from literature inferences for each setting were used to predict the Holdout Set data in the same manner above. These were compared against the CASH-derived candidate Fusion systems using a paired difference test over the 5 Holdout Set subjects.

6. Those Fusion architectures which did not attain the highest Development Set accuracy (i.e. all those other than the previously selected candidates) in each setting were subsequently assessed on their ability to predict the Holdout Set data. All Fusion architectures were then compared against each other with pairwise tests in both Bespoke and Generalist settings.

5.5 Results

This section addresses the various Aims outlined in 5.1.1 in turn. Firstly, Sections 5.5.1 & 5.5.2 address Aims 5.1 & 5.1.1 respectively, investigating multimodal and unimodal systems' classification abilities on Development Set data and verifying findings with the reserved Holdout Set. Sections 5.5.3, 5.5.4, 5.5.5, & 5.5.6 explore Aim 5.2, each looking at the modelling decisions contributing to fusion systems' configurations in different contexts; Section 5.5.7 then considers aspects of Aim 5.2 specific to the use of Unimodal EEG systems. Finally, Section 5.5.8 seeks to investigate Aim 5.3 by evaluating the efficacy and benefit of the outlined approach involving Combined Algorithm Selection & Hyperparameter Optimisation as a strategy for devising accurate biosignal fusion systems for this gesture classification task.

5.5.1 The merit of Fusion

From the defined aim of this chapter to "*Establish whether a multimodal system can offer better performance than a unimodal one*" (Aim 5.1) we can derive the null hypothesis that: "*A multimodal system's performance will be no greater than a unimodal one*", or more formally:

$$H_0 : \mu_{fusion} - \mu_{unimodal} \leq 0. \tag{5.2}$$

To test this, candidate multimodal and unimodal systems, each afforded the same optimisation budget, must be put forwards.

5.5.1.1 Selecting candidate systems

Tables 5.3 presents for each fusion architecture the modelling choices of the best-performing Bespoke system found in CASH optimisation, and the mean of the classification accuracies reached by that system for each of the 20 Development Set subjects. Table 5.4 presents the same for Generalist systems.

The highest-scoring architecture over the Development Set is selected from each of these cases for comparison against a unimodal approach, to address Aim 5.1. In the Bespoke case this is the Hierarchical Fusion system, and in the Generalist the Feature-Level Fusion, of the subtype which performed feature selection after joining EMG & EEG data.

Architecture	Accuracy (mean across subjects)	Modelling choices
Decision-level	87.83	Fusion Algorithm: Max EMG Model: SVM C: 98.9189 Gamma: 0.0131 EEG Model: Random Forest Number of trees: 85
Feature-level (Separate selection)	86.13	Linear Discriminant Analysis Solver: Singular Value Decomposition
Feature-level (Joint selection)	86.48	Linear Discriminant Analysis Solver: Singular Value Decomposition
Hierarchical	**88.98**	EEG Model: Quadratic Discriminant Analysis Regularisation: 0.4559 Supplemented EMG Model: Support Vector Machine C: 19.4037 Gamma: 0.0138
Inverse Hierarchical	83.68	EMG Model: Quadratic Discriminant Analysis Regularisation: 0.3325 Supplemented EEG Model: Random Forest Number of trees: 75

Table 5.3: Peak mean classification accuracy over Development Subjects achieved in CASH optimisation of Bespoke fusion architectures & corresponding system configurations.

Architecture	Accuracy (mean across subjects)	Modelling choices
Decision-level	71.67	Fusion Algorithm: Linear Support Vector Machine C: 0.0538 EMG Model: Linear Discriminant Analysis Solver: Least Squares Solution Shrinkage: 0.2349 EEG Model: Linear Discriminant Analysis Solver: Eigenvalue Decomposition Shrinkage: 0.3693
Feature-level (Separate selection)	72.03	Linear Discriminant Analysis Solver: Eigenvalue Decomposition Shrinkage: 0.0235
Feature-level (Joint selection)	**72.30**	Linear Discriminant Analysis Solver: Least Squares Solution Shrinkage: 0.1871
Hierarchical	71.68	EEG Model: Linear Discriminant Analysis Solver: Singular Value Decomposition Supplemented EMG Model: Linear Discriminant Analysis Solver: Singular Value Decomposition
Inverse Hierarchical	71.68	EMG Model: Linear Discriminant Analysis Solver: Least Squares Solution Shrinkage: 0.0229 Supplemented EEG Model: Linear Discriminant Analysis Solver: Singular Value Decomposition

Table 5.4: Peak classification accuracy achieved in CASH optimisation of Generalist fusion architectures & corresponding system configurations. NB that both the Hierarchical & Inverse Hierarchical systems noting a mean accuracy of 71.68% is not a typographical error.

Tables 5.5 and 5.6 present the optimiser-identified Bespoke and Generalist Unimodal systems respectively. In both cases, the EMG-based systems are the clear winner and hence are chosen as the unimodal candidates to compete with the above mentioned multimodal systems.

Data modality	Accuracy (mean across subjects)	Modelling choices
EMG	**87.78**	Support Vector Machine C: 4.1725 Gamma: 0.0126
EEG	54.80	Linear Discriminant Analysis Solver: Least Squares Solution Shrinkage: 0.038

Table 5.5: Peak Development Set accuracy in Bespoke Unimodal CASH optimisation & corresponding configurations

Data modality	Accuracy (mean across subjects)	Modelling choices
EMG	**68.90**	Linear Discriminant Analysis Solver: Eigenvalue Decomposition Shrinkage: 0.0744
EEG	49.11	Linear Discriminant Analysis Solver: Least Squares Solution Shrinkage: 0.435

Table 5.6: Peak Development Set accuracy in Generalist Unimodal CASH optimisation & corresponding configurations

5.5.1.2 Competing multimodal and unimodal gesture classification systems

For a valid comparison, both the selected fusion and unimodal systems are tested on the same held-out dataset described in 5.2.3.2. To statistically analyse performances, the five holdout subjects are considered the test's sample; as the observations (i.e. the classification accuracies achieved by a system for each subject) in the systems' samples are taken from the same set of subjects, they are compared by a paired t-test. The null hypothesis (5.2) is of a form which specifies direction and thus a one-tailed test performed.

In the Generalist case results as described above in 5.2.3.2 results are exactly repeatable and so here both systems are tested on each Holdout Subject in the holdout set only once, the results of which can be seen in Table 5.7.

A paired Student's t-test assumes a normal distribution of the differences between pairs and a homogeneity of variances across the conditions; before applying the test these assumptions are verified. The Shapiro-Wilk test has been found more powerful than alternatives including at small sample sizes [298] for testing normality of paired differences. This was used here as implemented in *R Version 4.2.0* [299]; the resulting W-statistic is 0.90103 at a p-value of 0.4156. This is well above both the $\alpha = 0.05$ confidence level, thus the hypothesis that differences were normally distributed is not rejected & this assumption is satisfied. A simple F-test using *R*'s *var.test* allows comparison of the population variances, with a resulting F-statistic of 0.65033 and

Subject	System	
	Unimodal EMG	Feature-level Fusion (Joint selection)
1	0.59333	0.66333
6	0.70875	0.74750
11	0.79333	0.82333
16	0.69708	0.72458
21	0.66375	0.71208

Table 5.7: Candidate Generalist Fusion & Unimodal classification accuracies on Holdout Set

p-value of 0.6869. The null hypothesis, that the ratio of variances is equal to 1, is hence not rejected and the assumption of equal variances is also satisfied.

The paired one-tailed t-test between candidate Fusion and Unimodal Generalist systems reports a t-statistic of 5.5761, with a p-value of 0.002535. The estimated mean difference in accuracy was 0.04291667, with the lower bound of the 95% confidence interval equal to 0.02650891. This suggests we can reject the null hypothesis (5.2) in the Generalist case, and conclude that the optimised Generalist Feature-level Fusion system can be expected to offer a performance boost over the similarly-optimised Generalist Unimodal EMG system in classifying the four same-hand gestures.

In the Bespoke case, the train/test split for each subject was as described in 5.2.3.2 random in nature, meaning that a given system configuration selected in 5.5.1.1 is liable to achieve slightly different accuracies for a subject if run multiple times. Therefore as previously outlined each Bespoke architecture was tested on each Holdout subject 100 times, and the means of these repeat measures presented here in Table 5.8.

Subject	System	
	Unimodal EMG	Hierarchical Fusion
1	0.8032	0.8293
6	0.8249	0.8345
11	0.9455	0.9467
16	0.8376	0.8324
21	0.8650	0.8697

Table 5.8: Candidate Bespoke Fusion & Unimodal classification accuracies on Holdout Set (means across 100 trials)

Here we must again verify the assumptions of the one-tailed paired t-test. The Shapiro-Wilk test computes a W-statistic of 0.93253 at a p-value of 0.6137, failing to reject the null hypothesis thus indicating normality of the paired differences. Comparing the two systems' variances results in an F-statistic of 0.81544 at a p-value of 0.848, again failing to reject the null hypothesis and so suggesting equality of variances.

Here in contrast to the Generalist case, the optimised Bespoke multimodal system is not found to offer improved performance over the equivalently-optimised Bespoke unimodal system. The paired t-test results in a t-statistic of 1.3771 with a p-value of 0.1203, above the $\alpha = 0.05$ threshold indicating the trend observable among Table 5.8's results in favour of fusion not to be statistically significant.

It is worth recalling here that as noted in 4.1.1.1 the participants in Jeong et al.'s dataset used in this work were all able-bodied individuals [198]. Electromyographic data collected from amputees has routinely proven more challenging to classify than that of able-bodied individuals: Menon et al. for example found classification errors for transradial amputees to be 31.5% greater than those of able-bodied participants [228], and Scheme & Englehart similarly found lower classification accuracies of transradial amputees' data compared to able-bodied individuals, across a range of classifiers [14]. This is unsurprising; various factors including the potential presence of scar tissue [14], the size of residual limb [43] & degree of voluntary control over residual forearm muscles, and the site of amputation [228] — many of which are highly variant between amputees — result in an observable reduction in the level of information carried by amputees' EMG signals [13]. It may hence be that while subject-specific fusion has provided no significant benefit in classification accuracy over an equivalent unimodal strategy here, in those amputees for whom the performance of EMG classifiers may be diminished, a multimodal fusion approach could provide an accuracy boost significant at the 0.05% confidence level, & indeed greater than the estimated mean difference between fusion and unimodal performances of 0.728% observable here. Indeed one of the seminal works on fusing biosignals for gesture classification, that of Leeb et al. [146], emulated precisely this effect by artificial attenuation of EMG amplitude, and posited the inclusion of EEG data to be of value in such situations.

Of course, that Hierarchical Fusion reached accuracies of equivalent levels to the Unimodal system here is not necessarily a guaranteed indication of its usefulness — it cannot be categorically assured that the top-level model of the Hierarchical system did not simply learn to disregard the classification probabilities estimated by its constituent EEG-LDA. Nevertheless these results clearly demonstrate the potential of the novel Hierarchical architecture for multimodal biosignal fusion, and indicate the merit of further research exploring the impact of such a fusion strategy in cases where the informativity of EMG data is diminished.

5.5.2 The impact of Fusion Architecture

Having identified fusion systems as capable of surpassing unimodal ones in the Generalist case, and performing at a similar par in the Bespoke, Aim 5.1.1, to "*Establish whether the fusion architecture impacts system quality, & identify performant fusion architectures*" is subsequently investigated.

5.5.2.1 Holdout Set performance of Fusion Architectures

Tables 5.9 and 5.10 present the classification accuracies of the fusion architectures proposed in 5.3.1 for Bespoke and Generalist cases respectively, each using their optimal configurations as identified in Tables 5.3 and 5.4, on the Holdout dataset (with Bespoke scores again being the means of 100 trials as outlined above). It is immediately evident that in every architecture the modelling choices determined as suitable by the CASH optimisation process were able to generalise beyond the Development Set to some degree; scores are consistently above the chance level in all cases. The theoretical chance level for a balanced 4-class system is 25%. As noted by Müller-Putz et al. [300] however, in reality the threshold for determining better-than-chance performance is modified both by the confidence level (α) and the number of trials of each class being tested. Here, all participants performed each gesture 50 times in each of three recording sessions; a total of

150 performances of each class. Treating each performance as a single "trial"[7] , this means that in Generalist systems wherein all of a given subject's data were used for testing, the number of trials per class were 150. The upper confidence interval of the chance level is hence approximately 29% at the α=0.05 confidence level, and approximately 30% at α=0.01 [300]. In the Bespoke case, wherein two-thirds of each participants' data were used for training, 50 trials per class remained for testing; the upper limits of chance-level results at the 0.05 and 0.01 confidence levels would be approximately 32% and 35% respectively [300]. It is clear from Tables 5.9 and 5.10 that all both Bespoke and Generalist systems of all fusion architectures achieved performance well above these thresholds for all Holdout Subjects.

Architecture	Classification Accuracy					
	1	6	11	16	21	Mean
Decision-level	0.8125	0.8144	0.9467	0.8292	0.8661	0.8538
Feature-level (Separate selection)	0.7812	0.8170	0.8980	0.7891	0.8287	0.8228
Feature-level (Joint selection)	0.7747	0.8101	0.8960	0.7816	0.8114	0.8148
Hierarchical	0.8293	0.8345	0.9467	0.8324	0.8697	0.8625
Inverse Hierarchical	0.7518	0.7942	0.8900	0.7619	0.7907	0.7977

Table 5.9: Bespoke classification performance of the proposed fusion architectures on Holdout Set data, each using the respective configurations identified by CASH optimisation in Table 5.3 (scores averaged over 100 trials).

Architecture	Classification Accuracy					
	1	6	11	16	21	Mean
Decision-level	0.6500	0.7408	0.8138	0.715	0.7146	0.7268
Feature-level (Separate selection)	0.6729	0.7446	0.8238	0.7158	0.7108	0.7336
Feature-level (Joint selection)	0.6633	0.7475	0.8233	0.7246	0.7121	0.7342
Hierarchical	0.6721	0.7450	0.8125	0.7021	0.7054	0.7274
Inverse Hierarchical	0.6479	0.7371	0.8213	0.7146	0.7117	0.7265

Table 5.10: Generalist classification performance of the proposed fusion architectures on Holdout Set data, each using the respective configurations identified by CASH optimisation in Table 5.4.

While classification accuracies for these unseen Holdout Subjects are of broadly similar levels to the mean accuracies over the 20 Development Subjects reached in optimisation, comparing tables 5.3 and 5.9 shows Bespoke Holdout performance to be weaker than optimisation performance. This is not surprising and highlights the risk of overfit inherent to the CASH optimisation process. While in every iteration of any optimisation process data were split such that the specific model configuration being evaluated had no sight of the test data, the optimiser itself as a higher-level learning process ultimately had knowledge of all

[7] In actuality, the time-window segmentation procedure (see 4.3) means that each gesture performance contributes multiple instances to the dataset. For the purposes of determining chance level thresholds, the worst-case assumption is taken here: that the *effective* number of trials contributed by each gesture performance is 1. The upper confidence limits of a chance-level result for larger numbers of trials per class would be closer to the theoretical chance level of 25%.

20 subjects (per Figure 5.1), and hence its modelling choices may be tailored to those individuals. This is precisely the over-optimisation phenomenon described by Hosseini, Powell, et al. [31] which — as has been discussed previously in 3.3.2, 4.2.3, & elsewhere — was a key motivation for the decision to validate findings using wholly unseen Holdout data as is done here rather thank risk reporting inflated accuracies on the basis of non-generalisable results.

Given that, it may be surprising that in the Generalist case, through comparison of tables 5.4 and 5.10 it can be seen that mean accuracy was marginally higher in validation across Holdout participants than in optimisation over Development. It should however be noted that as outlined in 5.2.3.2 above, for verification each optimiser-identified Generalist system configuration was retrained on all 20 Development Subjects's data before being tested on each Holdout Subject in turn. Whereas, during optimisation a Leave-One-Subject-Out strategy was used: they had in each case been trained on 19 of the Development Subjects and tested on the remaining one, repeated for each Development Subject in turn. This inclusion of a 20th subject's data for training represents a 5.26% increase in the total amount of data available to a system when tested on a Holdout Subject than on a Development, and also by inclusion of an additional individual among the training set a widening of the dataset's diversity. Either or both of these factors may account for this apparent performance boost.

5.5.2.2 Trends among fusion architecture performance

Considering the first clause of Aim 5.1.1, to "*Establish whether the fusion architecture impacts system quality*", the obvious null hypothesis to be tested is that "*fusion architecture does not impact system quality*". However to fulfil also the Aim's latter part, to "*identify performant fusion architectures*", requires not only testing for simply the presence of significantly different performances among architectures, but to assess whether some architectures are stronger or weaker than others. Fusion architectures are hence compared on a pairwise or "all-vs-all" basis, done here with Tukey's method in *R*, controlling for the effect of between-subject performance variation (discussed further in 5.5.2.3 below) by using the participant number as a blocking factor.

Hypothesis	Estimate	Std. Error	t value	p value
decision − hierarch	-0.00874	0.00656	-1.333	0.6758
feat_join − hierarch	-0.04776	0.00656	-7.284	**<0.0001**
feat_sep − hierarch	-0.03972	0.00656	-6.058	**0.0001**
inv_hierarch − hierarch	-0.06480	0.00656	-9.883	**<0.0001**
feat_join − decision	-0.03902	0.00656	-5.951	**0.0002**
feat_sep − decision	-0.03098	0.00656	-4.725	**0.0019**
inv_hierarch − decision	-0.05606	0.00656	-8.550	**<0.0001**
feat_sep − feat_join	0.00804	0.00656	1.226	0.7369
inv_hierarch − feat_join	-0.01704	0.00656	-2.599	0.1176
inv_hierarch − feat_sep	-0.02508	0.00656	-3.825	**0.0111**

Table 5.11: Pairwise comparisons of Bespoke fusion architectures (implementing optimally-identified configurations), using mean accuracies per Holdout subject per architecture over 100 trials

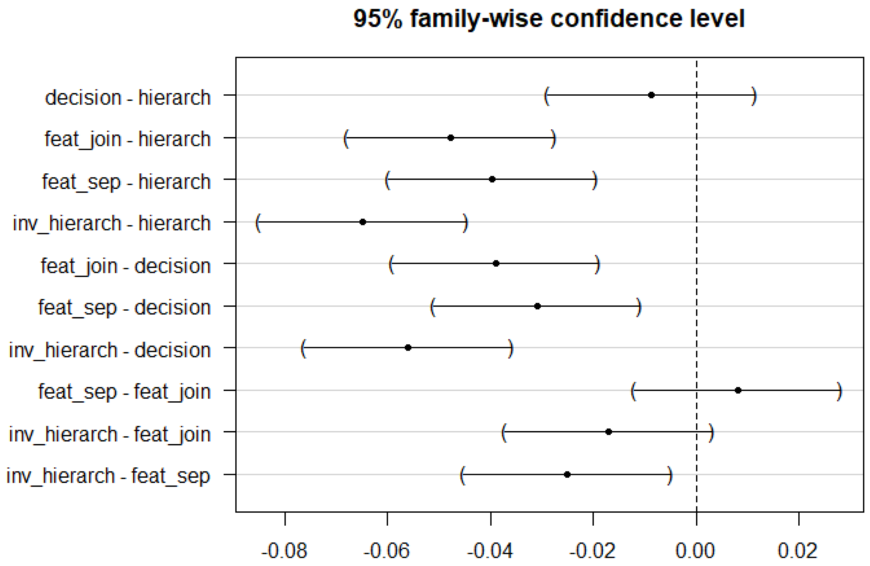


Figure 5.8: 95% Confidence Intervals of differences in means between fusion architectures in Bespoke systems (mean accuracies over 100 trials), each using CASH-identified configurations, estimated by Tukey pairwise contrast accounting for participants as blocks. (Differences are significant if the corresponding interval does not contain 0).

It can be seen from the comparisons presented in Table 5.11 and the visualisations in Figure 5.8 that in Bespoke fusion, the Hierarchical and Decision-Level architectures both outperformed all except each other. Of those outlined in 5.3.1, these two architectures are likely among the most able to "de-prioritise" the information carried by EEG. In the Hierarchical architecture EEG data is in effect condensed into a single feature (the probability distribution estimates by its component EEG classifier), and the optimal Bespoke Decision-level system's use of the Max rule rewarded the more confident constituent classifier, which considering the lower separability of EEG data may be assumed to more often be the EMG model.

Given the superiority of unimodal EMG systems' performance to that of unimodal EEG systems (Table 5.5), it would follow that a system more heavily influenced by EMG would plausibly be superior to one which under-utilises EMG. This is of course not to suggest such systems would be better served by disregarding EEG data *entirely* (else an optimised Decision-Level fusion system could never be expected to outperform its component EMG model, which was not borne out in results). It may however indicate that the subset of data which was classifiable with information carried by the EMG and that classifiable with the information captured by EEG were heavily overlapped. That is, datapoints misclassified by an EMG model may be not likely to be classifiable by an EEG model; the two were not able to compensate for each others' "blind spots".

This may suggest that, all other things being equal, the limiting factor in system performance was less related to the particulars of the machine learning algorithms applied, but could have derived from the informativity of the features (described in 4.3.2) extracted from the raw biosignal data. Hargrove et al. [265] found that in EMG classification the choice of appropriate feature vector was often of much more significant influence than the selection of model itself. Instability in the underlying ground truth from which EMG and EEG measurements were taken, such as irregular movements by participants, could even be a factor[8].

[8] see further discussion in 5.5.7

While not done here as it was not a central focus of investigation, future work could seek to track individual datapoints' likelihood of correct classification across data modalities & system architectures. This would help verify whether certain gesture performances proved inherently more difficult to classify by both EMG & EEG models, and could motivate exploration of modelling approaches which devote particular attention to residuals, such as gradient boosting methods, to overcome this. Conversely, if different data proved to be at risk of misclassification by EMG & EEG, such techniques could be adapted into an extension of the Hierarchical Fusion approach wherein one datatype was used explicitly to focus on the residuals of the other — a principle not dissimilar to the "error-correcting" approach to multimodality discussed in 3.1.2.1.

Whilst 5.5.1.2 found the candidate Bespoke Fusion system to not significantly outperform a Unimodal EMG model in this study, as was discussed there is reason to expect such a finding may not necessarily generalise to amputees. The promise shown by multimodal approaches in performing on-par with the unimodal EMG model on unseen data makes it evident that future work on the use of subject-specific biosignal fusion systems with amputees, ought to explore the Hierarchical & Decision-Level architectures as key lines of investigation.

Pairwise comparisons between fusion architectures for Generalist systems, again performed with Tukey's method, are presented in Table 5.12 and visualised in Figure 5.9. Among subject-independent systems, the Feature-level Fusion (with joint feature selection), found in 5.5.1.2 to outperform a Unimodal EMG-based system, did not in fact perform significantly differently from other architectures at the 95% confidence level. It is though noted that in the observed classification accuracies across Holdout subjects (Table 5.10) no other architecture outperformed this approach, consistent with its superior mean accuracy on the Development Set during optimisation (Table 5.4). The magnitude of this dominance however is incredibly small; it may be that the relatively low sample size in this work precludes such a small effect from being detected with confidence at the $\alpha = 0.05$ level.

Hypothesis	Estimate	Std. Error	t value	p value
decision − feat_join	-0.00733	0.00434	-1.688	0.468
feat_sep − feat_join	-0.00058	0.00434	-0.134	1.000
hierarch − feat_join	-0.00675	0.00434	-1.554	0.545
inv_hierarch − feat_join	-0.00767	0.00434	-1.765	0.425
feat_sep − decision	0.00675	0.00434	1.554	0.545
hierarch − decision	0.00058	0.00434	0.134	1.000
inv_hierarch − decision	-0.00033	0.00434	-0.077	1.000
hierarch − feat_sep	-0.00617	0.00434	-1.420	0.625
inv_hierarch − feat_sep	-0.00708	0.00434	-1.631	0.500
inv_hierarch − hierarch	-0.00092	0.00434	-0.211	1.000

Table 5.12: Pairwise comparisons of fusion architectures (using optimally-identified configurations) in Generalist case.

Notwithstanding that possibility, it appears that for a subject-independent system, across and regardless of fusion mechanism, the incorporation of EEG data can provide greater predictive power than use of EMG

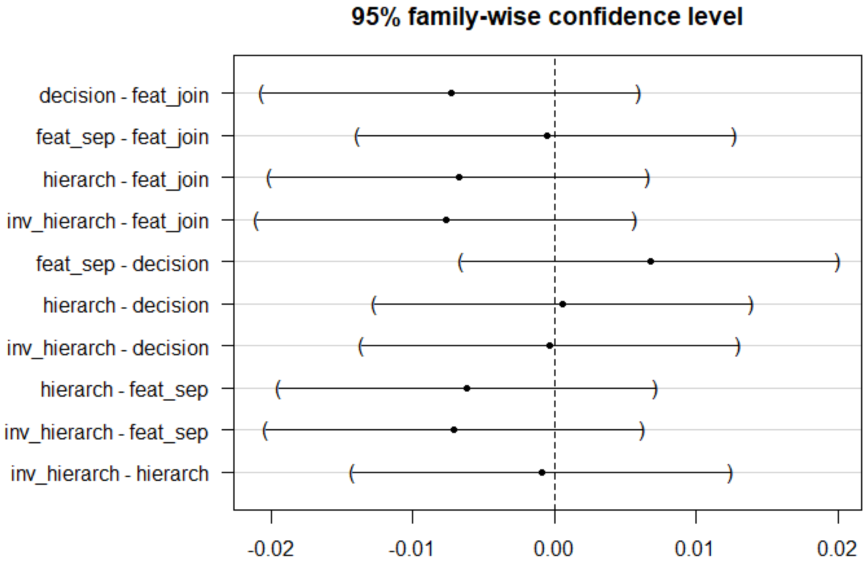


Figure 5.9: 95% Confidence Intervals of differences in means between fusion architectures in Generalist systems, each using CASH-identified configurations, estimated by Tukey pairwise contrast accounting for participants as blocks. (Differences are significant if the corresponding interval does not contain 0).

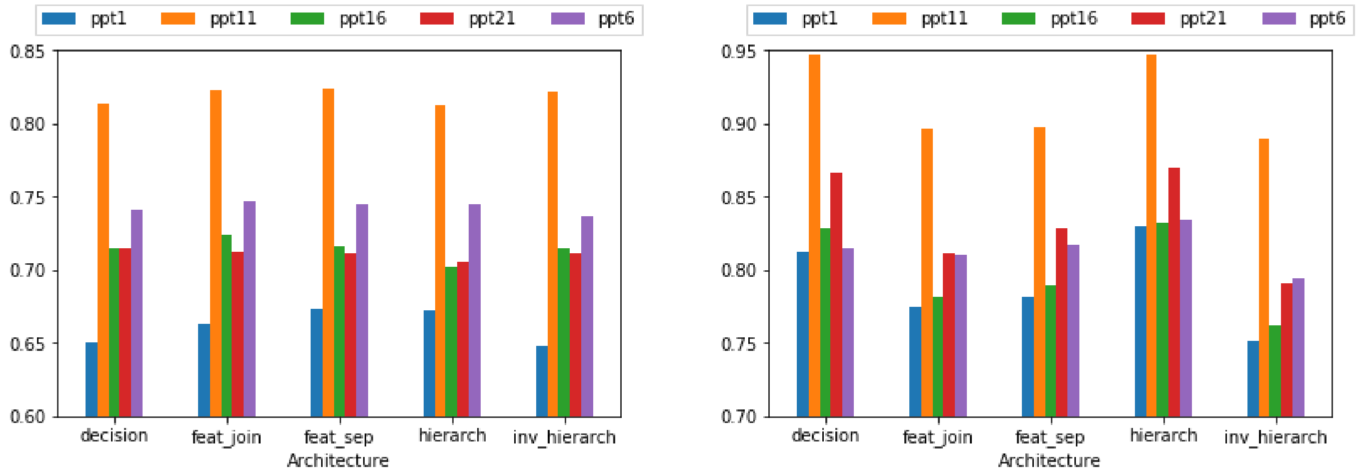
data alone in classifying same-hand gestures from unseen subjects. While under the given optimisation budget no fusion architecture was here able to achieve performance better than any other, it is implausible that no such superior fusion approach exists even if it is not found within the scope of the algorithms trialled in this work. Given the clear impact of EMG-EEG fusion in improving subject-independent classification ability demonstrated here, and the various benefits to prosthesis users in terms of to cost, time, and convenience which achieving subject-independence in gesture classification could enable, these results strongly motivate future research to assess a wider range of fusion strategies.

Such work should also investigate in depth the impact of the inherent increase in model complexity caused by incorporation of EEG to a system on prediction speed and required computational load. The latter is particularly in the context of a potential deployed prosthesis control system wherein component cost is an important consideration. The potential barriers presented by the cost and convenience implications of adding EEG to a system should be weighed not only against the degree of meaningful performance boost achievable by fusion as it relates to day-to-day usage however. If incorporation of EEG can enables better generalisability of classification systems, this should be weighed also against the barriers to access which may *already* be presented intrinsically by subject-specific systems, due to the degree of tailored support a prosthesis user may need in the setting up of such a system and the compounding effect of healthcare inequalities, particularly for disabled people, on the timeliness & likelihood of receiving such support. While well outside the scope of this work and the academic expertise of the author — the insight of researchers in disability theory and related social sciences, and of amputees with lived experience of prosthesis use, will doubtless be needed to do such investigations justice — implications such as these are vitally important and it would be remiss not to give them due acknowledgement.

5.5.2.3 Variation among participants

As indicated by the per-subject results in Tables 5.9 & 5.10, the various systems' respective abilities to generalise to the unseen Holdout Subjects varied by subject. To help highlight such differences Figures 5.10b & 5.10a present this variation for Bespoke and Generalist cases visually. As a reference, Figures 5.11a and 5.11b similarly present per-subject accuracies of Unimodal systems.

This illuminates some interesting trends. It is immediately apparent for example that Participant 11 routinely outperforms the other Holdout Subjects in both Bespoke and Generalist fusion. Participant 1 meanwhile routinely underperforms in Generalist systems, yet is not dramatically worse than some other subjects in many Bespoke cases. This gives confidence in the generalisation ability of the Bespoke system configurations identified by the CASH optimisation. It can be inferred from Figure 5.10a that Participant 1's data may differ notably from that of the 20 Development Set subjects used to train the Generalist systems, hence the diminished ability for their data to be informative in classifying Participant 1. Despite this apparent difference however, Bespoke systems were indeed able to be trained and utilised with Participant 1. While they saw a distinctly lesser degree of success they nevertheless achieved accuracies nearing those of their contemporaries in the Decision-Level and Hierarchical Fusion systems (themselves the two most accurate architectures for Bespoke fusion on average at the group level, as previously discussed). Thus the optimiser's learning of suitable modelling choices from Development Subjects was indeed somewhat transferable even to Participant 1, even where learning at the level of direct model training (as in Generalist systems) was less so.

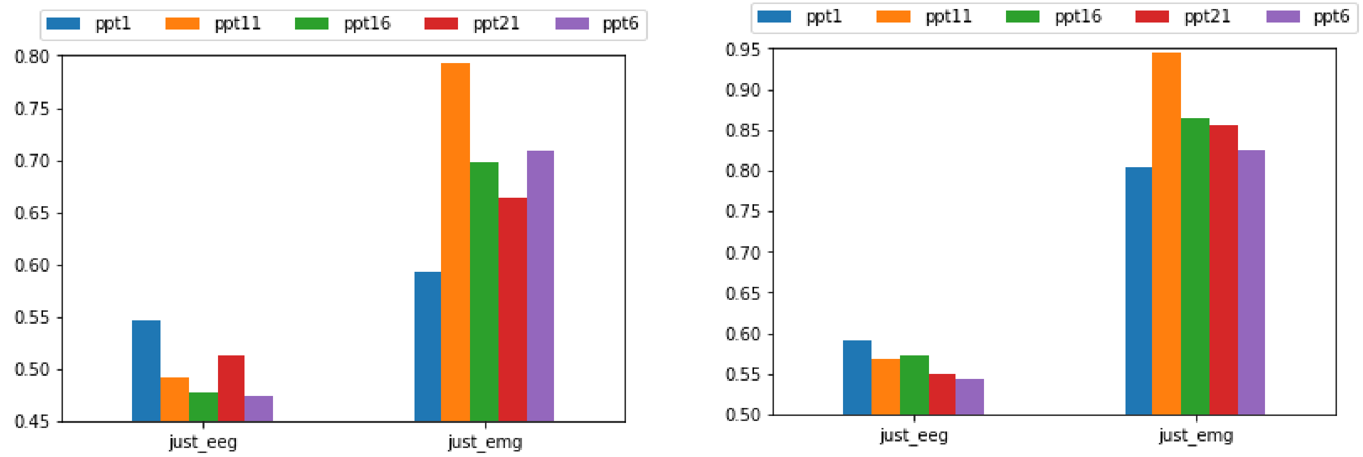


(a) Generalist

(b) Bespoke

Figure 5.10: Fusion architectures' accuracies across Holdout Subjects using optimal configurations; note that trends among fusion algorithms are present but scores are highly modified by subject. (NB datapoints in the Bespoke (right) case are the mean of 100 trials of a given system with a given subject)

Considering Figure 5.10b together with Figure 5.11b it can be further observed that for Participant 1, both the Bespoke Decision-level & Hierarchical fusion systems were stronger than the Unimodal EMG baseline. This is interesting in the context of Participant 1's Generalist Unimodal EMG baseline being notably weaker than other holdout subjects' and, per Figure 5.11b, their Bespoke Unimodal EEG baseline being stronger



(a) Generalist Unimodal Holdout generalisation performance.

(b) Bespoke Unimodal Holdout generalisation performance.

Figure 5.11: Unimodal systems' accuracies across Holdout Subjects in Generalist (left) and Bespoke (right) cases. (NB datapoints in the Bespoke (right) case are the mean of 100 trials of a given system with a given subject)

than others in the Holdout Set. These results may suggest that where EMG performance was poor, the Decision-level & Hierarchical fusion algorithms were able to successfully use the informative EEG data to supplement EMG predictions, thereby improving overall performance.

Additionally, considering this subject's status as an apparent outlier with exceptionally high EEG performance and low EMG performance, it should be recalled that systems were optimised for mean accuracy across all 20 Development Subjects, and hence likely tend towards configurations tailored to a "typical" user representative of the central tendency of that set. In the optimisation of Bespoke systems, this may lead to a configuration which is due to preference aggregation equivalently suboptimal for all subjects, while any given individual would actually be better served by a different system configuration. This is of course a feature, not a bug, of the CASH optimisation pipeline. The motivation was to identify a system configuration which could be trained and tested on unseen subjects' data reliably, and thus avoid the need for a dedicated CASH optimisation with each new user of a system. It may be however that the optimally-identified systems here were particularly ill-suited to Participant 1 due to their outlier status, and their gestures could be better classified by a system which drew on the EMG & EEG data in a different way. In the context of Aim 5.3, to "*Establish a pipeline for the unbiased identifying of a performant multimodal system*"[9], this is acknowledged as a potential limitation of the proposed optimisation pipeline.

Despite such stark performances differences between Holdout Subjects trends in the relative accuracies of architectures do appear to present in similar ways across them, consistent with the statistical analysis of such inter-architecture differences presented above.

[9] Discussed further in 5.5.8 below

5.5.3 On EEG Feature Informativity

Aim 5.2, to "*Identify modelling choices which can contribute to a multimodal or unimodal system achieving high classification accuracy*", is deliberately broad in scope. One category under the umbrella of "modelling choices" is the design decisions taken prior to any actual selection or training of machine learning models, specifically the ensemble of features extracted from data. Here, considering the established challenges of multi-gesture classification from noninvasive neuroimaging data and the distinctly less "solved" nature of this problem by comparison to EMG-focused research, and the vastly more complex nature of neurological biosignals than muscular, EEG modelling is considered of greater interest and hence given particular focus. While as discussed in Chapter 4 the feature ensemble (from which informative features were selected according to 5.3.2.2) was static throughout the work and not a primary target of investigation, exploring the informativity of some of the statistical features in Table 4.2 with regard to EEG data reveals some interesting insights.

Feature selection in Bespoke systems was performed separately for each subject as described in 5.3.2.2; in each case the 40 EEG features chosen were those determined as most informative for that specific subject. The consistency of various features' informativity, as a proxy for their importance, can be inspected by reviewing how frequently they were selected by these subject-specific systems [142]. A number of EEG features were found consistently informative across subjects in the Development Set during optimisation, being chosen by a sizeable subset of the population despite the individually tailored nature of this selection.

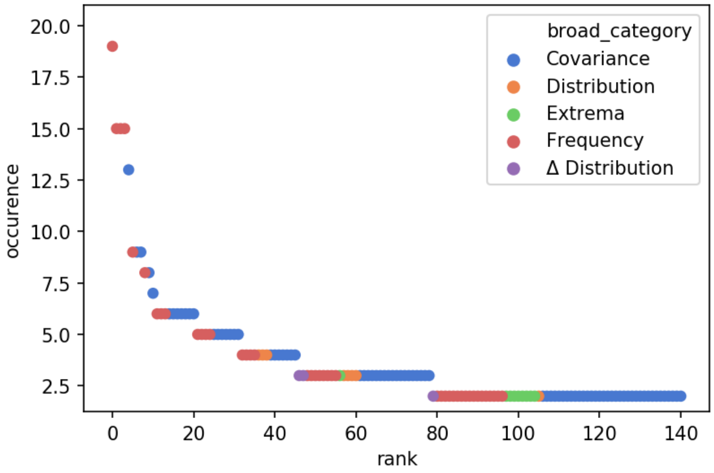


Figure 5.12: Co-occurrence of Bespoke-selected EEG features, grouped by feature category

Figure 5.12 presents[10] the number of Development Set subjects for whom each EEG feature was selected, labelled "occurrence", for all those features selected by more than one subjects' Bespoke Unimodal EEG system[11]. Here the Time- and Frequency- domain features outlines in Table 4.2 are grouped into broad categories: all frequency bandpowers of any signal as "Frequency"; all features relating to the covariance

[10] styled after [301]

[11] Features from the lag window (see 4.3) are not counted as occurring twice as this is ultimately the same feature at a shifted point in time

matrix, its logarithm, or its eigenvalues as “Covariance”; all features relating to signal means or standard deviations as “Distribution”, all features relating to the maxima and minima as “Extrema”, and any paired differences between half- or quarter- windows of a given type of feature as the delta (Δ) of that type, as in e.g. “Δ Distribution”.

Table 5.13 records the occurrence rates of each of these feature categories across all EEG features selected for any Development Subject’s Bespoke Unimodal EEG system. It is evident that features related to Covariance and to Frequency Bandpowers dominate, consistent with established domain knowledge, as outlined in 2.2.2.1, that much of movement-related neural activity is encoded in the frequency domain.

Covariance	177
Frequency	58
Extrema	42
Distribution	20
Δ Distribution	13
Δ Extrema	3

Table 5.13: Selection rates of EEG feature families in Bespoke Unimodal EEG systems

Dividing these feature families into narrower categories as in Figure 5.13 enlightens that among covariance-related features, the matrix logarithm of the covariance matrix is more consistently found informative than the elements of the covariance matrix itself. Additionally the bandpower of the Delta ($0.5 < f \leq 4$Hz) wave proves consistently informative in more than half of the subjects.

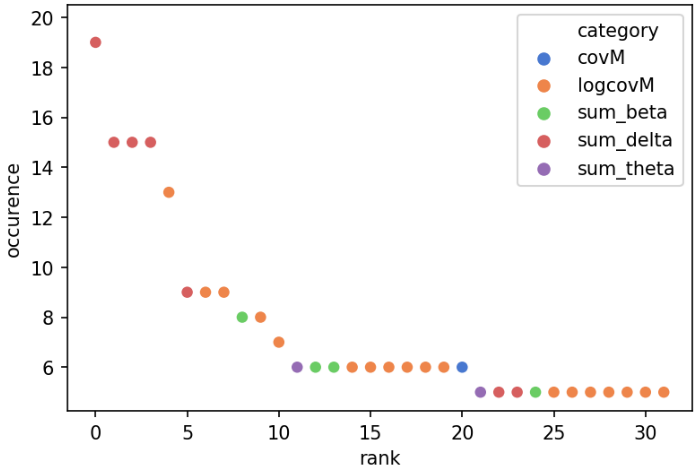


Figure 5.13: Co-occurrence of Bespoke-selected EEG features chosen for at least 5 (one quarter) of subjects, grouped by subdivided category

Figure 5.14 displays the specific individual EEG features most commonly found informative in Bespoke Unimodal EEG systems across the 20 Development Set participants. The Delta bandpower at electrode #13, which corresponds to electrode site FC6 of the extended International 10-20 System (Figure 4.2, above) was among nearly every single subject’s chosen 40 most informative EEG features. The same at electrodes 0, 3, and 16, 10-20 electrode sites FC5, C5, and C6, were chosen by three quarters of subjects.

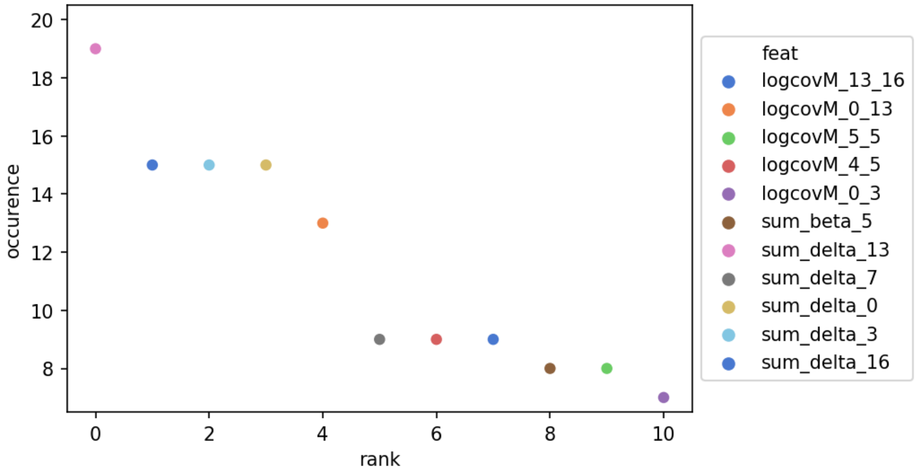


Figure 5.14: Co-occurrence of EEG features selected by ≥ 7 subjects' Bespoke systems

Table 5.14 presents the occurrence of each feature category among all EEG features selected in Generalist systems for any development set subject. It should be recalled that the leave-one-subject-out cross-validation procedure described in 5.2.3 for Generalist systems' optimisation means that for a given subject N, to avoid data leakage features were selected based on their informativity over the training set — which for N was comprised of the data of all 19 subjects *except* for N. Hence only one-nineteenth of the data differed between different subjects' Generalist training sets; they were significantly similar and it is therefore unsurprising that the inter-subject similarity between selected features is higher here than in Bespoke systems, and hence the total number of distinct features selected across all subjects fewer.

It is clear that again Covariance-related and Frequency Domain features dominate, again consistent with prior findings regarding the nature of neural signals. Such preference is more profound here than in the Bespoke case, with very few Time-Domain features being selected at all. Considering Generalist feature selection is on the basis of 19 subjects' data, this suggests that Frequency Domain features are not only more informative by nature than Time Domain features but also that the manner of their informativity is itself more consistent across subjects.

Covariance	76
Frequency	45
Distribution	4
Extrema	3
Δ Distribution	2

Table 5.14: Selection rates of EEG feature categories in any subjects' Generalist Unimodal EEG

Table 5.15 presents those EEG features which were selected in Generalist systems for all 20 Development Set subjects. Of note these are all either Frequency Domain features or those related to Covariances between electrode signals.

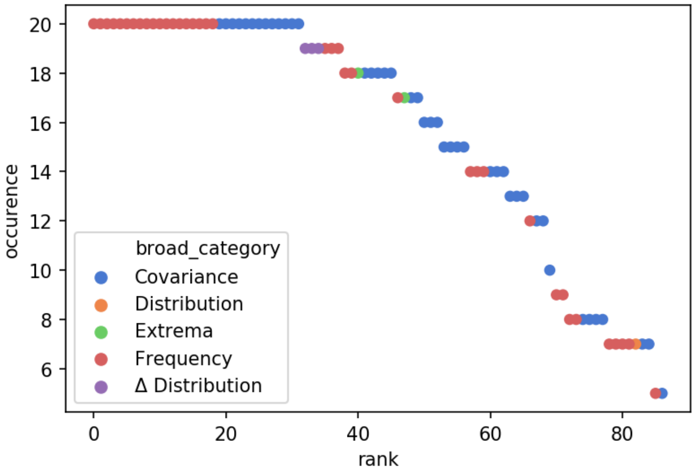


Figure 5.15: Occurrence of EEG features selected in Generalist systems for at least 5 (one quarter) of subjects, grouped by category

Feature			
logcovM 4-5	logcovM 3-16	logcovM 3-8	covM 2-2
logcovM 0-3	logcovM 5-5	logcovM 19-19	logcovM 7-19
covM 4-13	logcovM 13-16	logcovM 2-2	logcovM 12-13
logcovM 0-13	sum delta 0	sum alpha 15	sum alpha 13
sum alpha 10	sum delta 5	sum delta 16	sum beta 6
sum delta 13	sum delta 3	sum theta 3	sum beta 4
sum alpha 4	sum beta 9	sum beta 5	sum theta 16
sum delta 1	sum beta 11	sum beta 8	sum alpha 7

Table 5.15: EEG features selected in all 20 subjects' Generalist systems; presented here in no specific order.

As described above in 5.3.1, there are two feature selection methods in Feature-Level fusion systems. In the "Separate Selection" subtype, EMG & EEG features are selected independently prior to the datatypes being joined; the EEG features chosen are hence identical to those precedingly described in Unimodal EEG systems and so do not warrant further analysis. Under "Joint Selection" however, feature selection was performed over the join of EMG & EEG featuresets.

As can be seen in Figure 5.16 and Table 5.16, in this latter case Covariance and Frequency Domain features continued to dominate the regularly-selected EEG features in Generalist systems. Despite this, the Overlap Coefficient (Equation 5.3) [302] between the set of 69 EEG features selected for at least half of the subjects in Generalist Unimodal EEG systems and the 54 EEG features likewise selected for at least half of the subjects in Generalist Feature-Level Fusion systems is only 0.537 — indicating that when considered alongside EMG, the relative importances of EEG features were different. This perhaps suggests that some of those EEG features found informative when considering EEG alone, but of lesser interest when EMG is included, may have high levels of redundancy with the information carried by EMG features. Detailed analyses of such potential EEG-EMG feature correlations are left for future work.

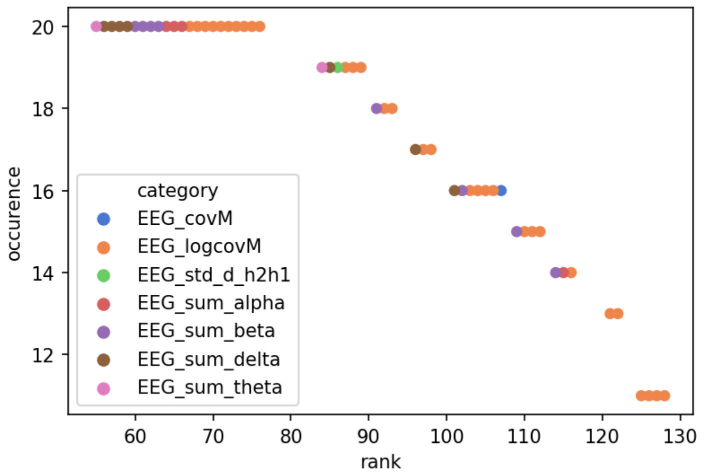


Figure 5.16: Occurence rates of EEG features, grouped by category, in Generalist Joint-Selection Feature-Level Fusion. For legibility EMG features, though present in Feature-Level Fusion, are not plotted here.

$$overlap(X,Y) = \frac{|X \cap Y|}{min(|X|,|Y|)} \tag{5.3}$$

Covariance	61
Frequency	41
Distribution	2
Extrema	1
Δ Distribution	1

Table 5.16: Selection rates of EEG feature categories in Feature-Level Generalist. Across all features selected at least once by any subjects' Generalist Feature-Level fusion systems, the 106 recorded here were EEG features, and a further 99 were EMG.

Implications

5.5.3.1 Delta-Band Oscillations

The persistent informativity of the Delta wave here is somewhat surprising. Motor cortex activity such as the event-related desynchronisation of the *mu* rhythm is typically found at frequencies corresponding broadly to the Alpha and low Beta bands [72, 73, 247], while Delta activity in the brain is more often associated with sleep [67,303], learning [68], auditory sensory processing (including when found in the Motor Cortex [304]), and even decision-making [69], as discussed in Chapter 2. Nevertheless some studies have found low-frequency EEG to be informative in discriminating upper limb gestures. Ofner et al. [75] were able to use EEG filtered between 0.3 and 3 Hz to discriminate between six distinct movements (*bicep* flexion & extension, wrist pronation & supination, and hand opening & closing) at up to 42% accuracy, and on an aggregated move-vs-rest basis at up to 81%, with LDAs. Iturrate et al. [162] were able to distinguish between two grasp types (a whole-hand cylindrical "power grasp" and a thumb-and-forefinger "precision pinch" grip) with EEG from just eight electrodes near the contralateral motor cortex filtered in the 1 - 6 Hz range, reporting a 79% average accuracy and distinct observable differences in EEG activity at electrode C3 between the two classes. Indeed Iturrate's work followed prior evidence from Pistohl et al. [167] wherein low frequency neural activity measured by ECoG had been found informative in distinguishing the same two grasps, though such low-frequency neurosignal components have more often been found, in ECoG as well as more rarely in EEG or MEG, to carry information regarding the trajectory of movements, i.e. their direction and speed [59, 305, 306], rather than the nature or form of a gesture itself. Hamel-Thibault et al. [307] found contralateral motor cortex Delta oscillations to predict individuals' choice of hand to move when asked to perform an arm-reaching movement at high speed. In this work however all movements were performed with the right hand only [198], suggesting it unlikely that such a discriminatory effect could explain the entirely of the Delta oscillation's informativity.Schalk et al. [77] observed apparent correlations between low frequency neural activity measured with ECoG and the direction of movement of a joystick controller, but posited that these were in fact highly spatially-localised time-domain amplitude correlations which they named “local motor potential”s, rather than truly frequency-domain phenomena.

The mechanisms of motor control are complex and it should of course be noted that the apparent informativity of Delta-band oscillations in this study does not necessarily predicate such oscillations playing a primary role in the control of voluntary upper-limb movements. Considering the encoding of directional reaching related information in low-frequency motor cortex oscillations, a limitation of this work is the consistency of object placement during data collection; the spatial arrangement objects corresponding to the three different grasp types was unchanged throughout the study and hence the classes correlated with object position. While they were in close proximity to one another, and not placed in opposing directions relative to the participant, it cannot be discounted that the small differences in reach direction when grasping each object could contribute to the informativity of Delta oscillations here. Taken in the context however of such grasp-discriminative ability of the Delta wave having been identified in a small number of preceding studies, this certainly appears an area worthy of further exploration by subsequent research.

This apparent informativity of low-frequency EEG does potentially present an interesting challenge with regard to real-time deployment of a gesture identification system. As described in 4.3.2 above, in this work features were extracted on a time-windowed basis, with windows of one second in length. Even prior to considerations such as sampling frequency, this presents an absolute theoretical minimum on the frequencies for which signal components can be accurately identified by the Fast Fourier Transform; only a signal of 1Hz or greater can complete a full cycle in the one-second window. Delta oscillations however are often characterised as being within the boundaries of 0.5 - 4 Hz, and indeed neural oscillations below 0.5Hz may prove of interest to future research in this area — potentially motivating use of a wider time-window for segmentation. Conversely however a *shorter*-duration window would allow for more frequent decision-making by the system; this may prove beneficial in reducing the delay between a users' movement intention and the system's response, which has been suggested to optimally lie around 100 milliseconds for powered prostheses [222], but at cost of restricting the systems' ability to identify low-frequency signals.

These two conflicting priorities could perhaps be managed by adjusting the overlap between sequential time-windows. A window for instance of a 2 second duration, but 95% (190 millisecond) overlap, would allow for signal components as low as 0.5Hz to be identified while enabling predictions to be made every 100ms, albeit ones based on data significantly similar to that of their immediately preceding and succeeding neighbours. This approach also would not be without limitations; the widening of windows would decrease their temporal specificity which may particularly coarsen those features related to signals' distributions over that window. Given the short duration of the gesture performances in the dataset used in this work, each being only 3 - 4 seconds, alternate durations and overlaps of time-windows were not trialled; empirical assessment of the impact of the considerations discussed here is left for future work in the field.

5.5.3.2 Ipsilateral Activity

The second surprising observation to be noted here is the informativity of signals recorded from electrodes sites at the ipsilateral hemisphere. It is well-established that, in broad terms, much of the brain is organised contralaterally — its left hemisphere controls the right side of the body, and vice-versa. It follows therefore that for movements of the right hand, as in this study, the brain region of most interest would be the motor cortex's opposite, left, side[12]; indeed the relevance of contralateral motor cortex EEG to both real and imagined movements has been long demonstrated [247].

Certainly, features derived from signals found in the contralateral hemisphere were indeed found informative; as shown in Figure 5.14 the Delta oscillations at electrodes 0, 3, & 7, corresponding to electrodes FC5, C5, and CP5 (Figure 4.2) and the Beta oscillation at electrode 5, corresponding to electrode C1 situated over the contralateral primary motor cortex, were selected by many participants' Bespoke systems,

[12] As all gestures in this study were performed with the right hand, anatomy hereafter is described relative to the right hand; that is, the right side of the brain referred to as "ipsilateral" and the left as "contralateral". Such use is convention but readers are nevertheless reminded these terms are inherently relative; where describing, for example, an EEG electrode as measuring the "contralateral hemisphere" (such as C3, as in Figure 4.2), this should be taken to mean "contralateral to the right hand".

and as per Figure 5.15, Delta power at electrodes 0 & 3 and Beta power at electrode 5, along with a great many other contralateral features including Beta power at electrode 4 (C3, directly over the hand-relevant area of the motor cortex), were likewise selected in Generalist systems for all subjects. Occuring at similar rates however, across both Bespoke and Generalist systems, were features derived from ipsilateral electrodes. Delta bandpower of channels 13 & 16, corresponding to electrodes FC6 & C6 respectively, were among the most consistently selected features in the Bespoke case. Theta bandpower at electrode 13 (FC6) was selected for around a third of subjects in Generalist cases; interestingly, while not a popular choice among Bespoke systems, Theta at electrode 16 (C6) was selected in Generalist systems for all subjects, perhaps suggesting that its individual informativity was rarely among the highest, but that it was informative in similar ways across participants.

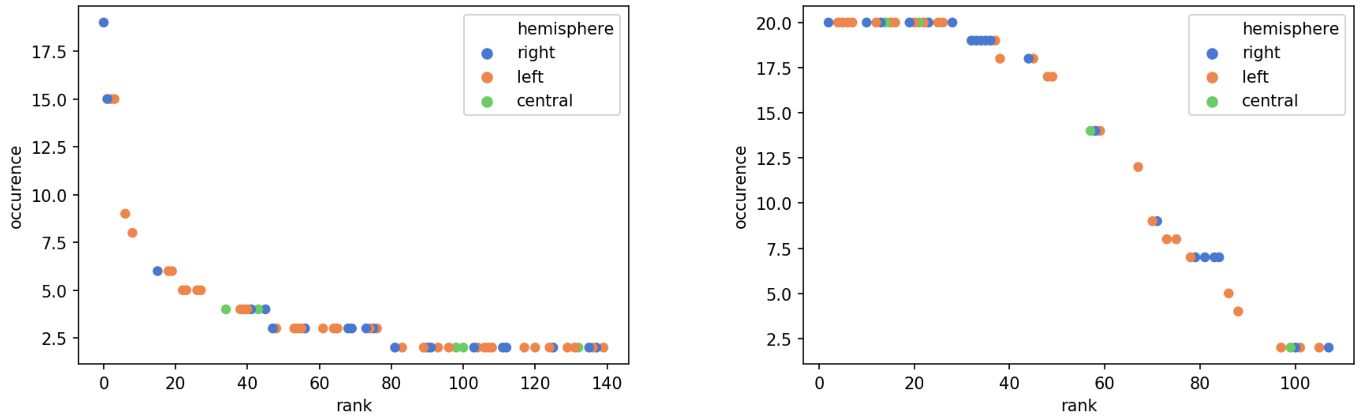
While as mentioned the bulk of motor control is known to take place contralaterally to a movement, there is evidence of the ipsilateral hemisphere playing a role [308]. Bundy & Leudhardt, reviewing the nature of ipsilateral motor activity, noted that "*Although the majority of primary motor cortex neurons alter their firing rate solely with movements of the contralateral hand, a separate small subset of neurons change their firing rate solely during movements of the same-sided hand*" [309].

Wisneski et al. [310] found movement-related activity ipsilaterally in the premotor cortex, which could be successfully used in conjunction with contralateral activity in a cursor-control BCI. Fujiwara et al. [290] found high-frequency (>64Hz) ipsilateral activity, measured with ECoG, to be as informative for decoding movement as being of the wrist, shoulder, or ankle as contralateral activity, and further that ipsilateral-trained models were able to generalise to contralateral data, suggesting a similarly structured neural representation of the body. Ames and Churchland [165] found in rhesus monkeys that while at the neuron level many neurons were active during movements of both contralateral and ipsilateral arms, and indeed some were more active in ipsilateral movements than contralateral, when Principal Component Analysis was applied, distinct and orthogonal subspaces were found across neurons of both hemispheres relating to activity of the left and right arms; Heming et al. [311] similarly found separability between the neural representations of ipsilateral and contralateral limbs.

Further, patterns of ipsilateral activity specific to different fingers have been found by Diedrichsen et al. [312] which were highly similar to those found contralaterally but interestingly did not present during bimanual movements (simultaneous movements of both hands). Subsequent work by Berlot et al. [313] evidenced these ipsilateral representations as being related to active movement rather than the accompanying somatic sensory input[13], and that the spatial distribution of such ipsilateral and contralateral activity differed.

It is believed that the movement-related changes in ipsilateral motor activity may be partially explained by them playing an inhibitive function [314]. Some f-MRI and TMS studies [315–317] have observed decreased levels of ipsilateral activity in parallel with increased contralateral activity during execution of a movement, and this effect being more pronounced during movements of an individual's dominant hand [316, 317]. Likewise, event-related desynchronisation of the contralateral *mu* rhythm is known to be often accompanied by

[13] Recall as noted in 2.1.2 that the somatosensory cortex lies immediately adjacent to the primary motor cortex; the two are sufficiently related that they are often together referred to as the sensorimotor area (or even the "sensorimotor cortex" — though the particulars of debates over neurological terminology are distinctly beyond the scope of this work).



(a) Bespoke; Contralateral (left side): 61, Ipsilateral (right side): 59, Central: 16

(b) Generalist; Contralateral (left side): 28, Ipsilateral (right side): 22, Central: 4

Figure 5.17: Co-occurences of EEG features selected for at least two subjects in Bespoke (left) and Generalist (right) systems, grouped by hemisphere. NB that due to greater cross-subject consistency in Generalist feature selection, the total number of Generalist features shown here is lower than the Bespoke, but this is not reflective of the number of features selected in a given system of the respective natures.

a *mu* synchronisation in the ipsilateral motor cortex [73, 247]. Such effects are thought to be a part of "inter-hemispheric inhibition" — in essence, the mechanism by which inadvertent movements of the "wrong" limb are prevented via communication between the brain's hemispheres [318]. This inhibitive function and the related interhemispherical communication can not however be said to account for the entirety of the ipsilateral hemisphere's role; movements of ipsilateral limbs have been induced in monkeys by electrical stimulation of the ipsilateral motor cortex even after surgical removal of parts of the contralateral motor cortex [319]. Interestingly, it has been found that among amputees, attempted movements of the residual muscles relevant to hand movements correlated with ipsilateral motor cortex activity only in those who did not experience phantom limb pain [320]; the causality of this is not precisely known but given that motor imagery exercises have been observed to reduce phantom limb pain in some patients it seems plausible that some underlying mechanism contributes both to the severity of the pain and the irregularity of motor cortex activity.

While it is perhaps not altogether surprising in and of itself that ipsilateral activity could be found informative in discriminating hand gestures as in this task, it is quite remarkable that ipsilateral activity was of an informativity seemingly on a par with contralateral activity. As shown in Figure 5.17a, of those EEG features selected for more than one subject[14] in the Bespoke case, 43.4% corresponded to ipsilateral electrodes, compared to 44.9% deriving from contralateral electrodes (the remainder being associated with centrally-located electrodes Cz or CPz). In the Generalist case ipsilateral features made up a smaller proportion of the commonly-selected features, at 40.7% to 51.9% contralateral as in Figure 5.17b, than in Bespoke systems, but nevertheless a greater contribution than may have otherwise been assumed *a priori*.

It has been suggested [309,310] that power attenuation in the Gamma frequency band may be of particular

[14] Excluding features related to the covariance matrix and its derivatives, which cannot be said to relate exclusively to a single electrode's signal.

relevance to the interhemispheric inhibitive effect; it is noted here that Gamma activity at channel 4 (electrode C4, located most closely to the region of the ipsilateral motor cortex relevant to the hand), while not a frequently selected feature in Bespoke systems, was selected as informative in Generalist systems for 17 of the 20 subjects.

What should be noted is that the vast majority of studies mentioned here (and indeed many not explicity discussed) which have investigated movement-related ipsilateral motor cortex activity have done so only with neural imaging techniques much higher-fidelity than the scalp EEG used in this work. Functional Magnetic Resonance Imaging was used by [312, 313, 315, 316]; others utilised ECoG [290, 310, 311], which as [310] notes offers both a much wider bandwith of measurable frequencies and a much higher spatial resolution between independent electrode channels (0.125cm separation for ECoG to 3.0cm for EEG), and some studies involving rhesus monkeys have measured individual neurons at the single-unit level [165]. To identify an apparent informativity of the ipsilateral motor cortex with regard to hand grasp type from EEG data as in this work is notable — particularly considering that Delta oscillations, not previously identified as a fundamentally central aspect of informative ipsilateral activity appear here to be similarly informative to their contralateral counterparts. This provides a clear motivation for ipsilateral low-frequency EEG to be considered as a potentially informative source of movement-related information in future work, and for the exact nature of its contribution to motor control to be investigated more thoroughly.

5.5.3.3 Neural Geography

It is also of note that many features derived from electrodes not situated directly over the primary motor cortex were seen to be informative here. As discussed, features including particularly the Delta oscillations at electrodes FC5, C5, C6, and FC6 (Figure 4.2), among other sites, were routinely selected for both Bespoke and Generalist systems for many subjects; indeed, the Delta bandpower at FC6 was the only individual feature to be consistently selected in Bespoke systems across all but one participant, as shown in Figure 5.14. While brain size and shape is certainly variant across individuals, typically in EEG recordings electrodes C3 and C4 are expected to capture hand-related motor cortex activity in the contralateral and ipsilateral hemispheres respectively.

For electrodes both more distal (further “left” or “right” from the centre of the body) and more anterior (or *ventral*, further to the “front” of the body) to be so informative here then is perhaps initially surprising. Indeed, there may be mundane explanations — misalignments in the EEG electrode cap placement, or use of an ill-fitting cap which insufficiently stretches over the head, could plausibly cause primary motor cortex activity to be measured at unexpected electrodes which had inadvertently ended up lying over it. Given however that the likelihood of such issues as poor electrode cap fit would intuitively be modified by the particular size & shape of a participants’ head, it would be surprising for such issues to present systematically across participants and they therefore are unlikely to have caused this informativity. Additionally, as noted in Table 5.15, many features at the more conventionally expected electrode sites were found informative, particularly in the Generalist case. Were electrode misalignment to blame here, one could expect the “usual” electrodes to no longer provide data as informative.

Rather, it would appear that genuine movement-related activity, discriminative with regard to the type of hand gesture performed, was indeed being measured outside the primary motor cortex. One of the most likely contributors to this is the premotor cortex. Located anterior to the primary motor cortex, as shown in Figure 5.18, the premotor cortex is conventionally known to be involved in the planning stages of voluntary movements, before motor execution. Premotor cortex activity has not however been found to solely take place at the pre-movement stage. Berlot et al. [313] found that spatial distributions of ipsilateral finger representations differed from those in the contralateral hemisphere — with contralateral representations being more localised to Brodmann areas 3a and 3b[15], i.e. the primary motor cortex, while ipsilateral representations were stronger in the premotor cortex. Similarly Wisneski et al. [310] found that while movement-related contralateral activity was located in the primary motor cortex, ipsilateral activity in the premotor cortex was relevant to voluntary movements. It is hence wholly plausible that those electrodes situated more ventrally, i.e. the FCX "row" including FC5, FC3, FC4, and FC6 (channels 0, 1, 12, & 13 respectively) in Figure 4.2, may well be measuring premotor cortex activity. Indeed, as in Figure 5.19, FC5 and FC6 specifically may be expected to measure the premotor cortex to some degree.

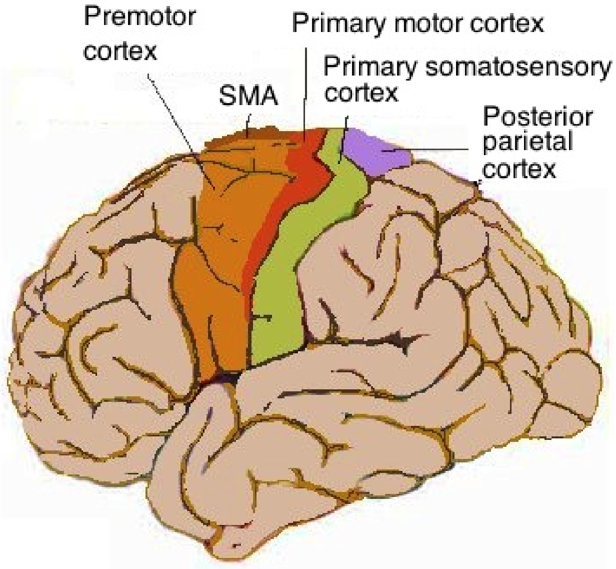


Figure 5.18: Location of premotor cortex, highlighted orange (NB: SMA refers to the Supplementary Motor Area) [321]. Used under CC BY-SA 3.0 [`https://creativecommons.org/licenses/by-sa/3.0/deed.en`].

It is also important to note that the traditional assumption of hand movement neural activity being local to electrode C3 contralaterally, and C4 ipsilaterally, may itself be less safe than believed. The classical "cortical homunculus" model of the motor cortex as outlined in 2.1.2 is one of a heavily somatotopic mapping, as presented in Figure 2.3. According to such a model we would expect hand-related motor activity to be measurable mainly at electrodes C3 and C4, and signals at electrodes C5 and C6 to be more closely associated with movements of the face. This association is sufficiently embedded that some works place very significant focus on C3 activity for hand movement identification [162]. Again it is important to highlight that C3 and

[15] see 2.1.2 above for more thorough discussions of the neural topology

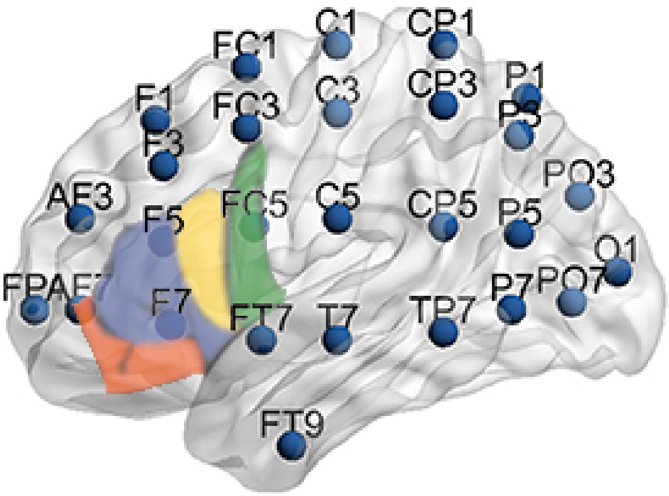


Figure 5.19: Standard EEG electrode labels of the 10–10 system (see Figure 4.2), with Brodmann areas 47 (orange), 45 (blue), 44 (yellow), & 6 (green) highlighted. Note that FC5 here is labelled at Brodmann's area 6 (the premotor cortex) and adjacent to Brodmann's area 44, which though typically associated with speech has been found [322] to be active in the control of hand movements.
Image created through overlaying Fig. 1 of [323], originally published under CC-BY 3.0, and Fig. 6 of [324], originally published under CC-BY 4.0 and created using BrainNetViewer [325].

C4 were indeed found informative here. Both Beta and Alpha bandpowers at C3 (channel 4), and Alpha power at C4 (channel 15), were features selected as informative in Generalist systems for all 20 subjects (Table 5.15). The informativity of the more distal C5 and C6 electrodes however is indeed somewhat surprising. Recent work by Muret et al. [326] may shed some light on this. Assessing relationships between distinct body parts and regions of the somatosensory cortex, they did indeed find a high degree of univariate selectivity of cortical regions in keeping with the conventional homunculus model. However when applying multivariate analysis, they found relevant information, while strongest at the conventional highly selective sites, to actually be distributed throughout the cortex, as represented in Figure 5.20. While a new model, their findings are not entirely without precedent; Schieber et al.'s neuron-level measurements in monkeys suggested the hand area of the motor cortex to be less strictily somatotopic with respect to individual fingers than otherwise assumed [327] — though other work such as [166] has succesfully spatially discriminated finger movements in humans with ECoG. Of particular interest to this work are Muret et al.'s observations that:

- *"Two movements performed by one body part (e.g., the hand) could be dissociated well beyond its primary region"*;
- *"Two actions done with the same body part can be differentiated in non-primary regions of the homunculus"*.

Given the focus of the problem in this work is the differentiation of movements performed by the same body part, it would seem wholly plausible this more distributed somatotopy could account for the informativity of electrodes such as C5. Clearly this topic is a developing one in the literature and no attempt is made here to conclude what will doubtless be, as with any proposed revision to a long-established model, a robust debate

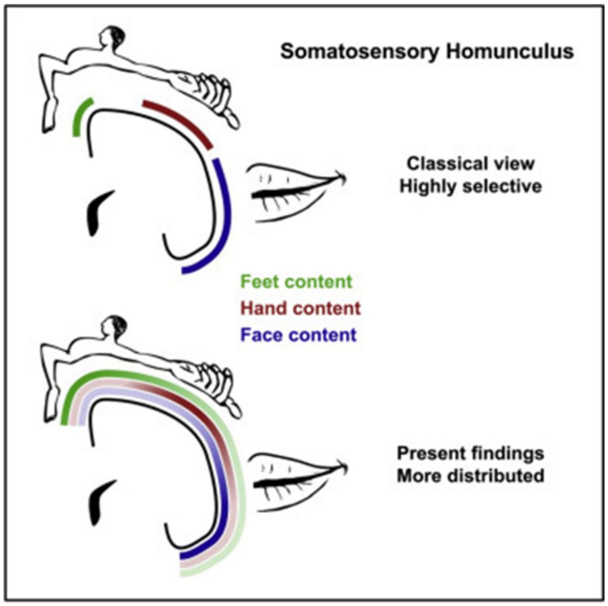


Figure 5.20: Distributed nature of the somatosensory homunculus, according to Muret et al. [326]. Used under CC-BY 4.0 [https://creativecommons.org/licenses/by/4.0/]

investigated much more thoroughly in the near future. Nevertheless, these ideas prove to be insightful lenses through which to consider the findings of this work relating to EEG informativity.

5.5.4 Modelling Choices

The trialling of various modelling choices for same-hand gesture classification performed as part of this works' Combined Algorithm Selection and Hyperparameter optimisation processes allows for an exploration of the impact of those choices, to better identify suitable modelling strategies for these data as an additional means of furthering Aim 5.2. The influence of hyperparameters in certain models deemed promising due to their appearance in optimiser-chosen systems (see 5.5.1.1) is explored here. Appendix A presents, without analysis, trends in hyperparameters of those algorithms not deemed of particular interest; considering these models' lesser relevance this is in many cases provided on the basis of Unimodal systems alone.

5.5.4.1 Linear Discriminant Analysis classifiers

Bespoke

As seen in Table 5.5, the single best-performing Bespoke EEG-only system used a Linear Discriminant Analysis classifier. The LDA is a popular model choice among EEG literature [30] and so its contribution to the winning configuration here is perhaps unsurprising. It is not uncommon however for EEG studies to utilise LDAs with minimal justification for their selection; on the basis simply of precedent or unsubstantiated *a priori* claims as to their superiority. The CASH optimisation here thus provides an opportunity to explore such precedence empirically.

Figure 5.21 presents the mean classification accuracy across Development Subjects for each iteration of the Bespoke Unimodal EEG system's optimisation routine, grouped by model choice (as outlined in 5.3.3, unimodal systems consisted of only a single classifier). Individual models' hyperparameters were conditional on the choice of classifier; hence while they may influence within-group variation, when comparing between groups we can consider the model choice to be the defining variable separating them.

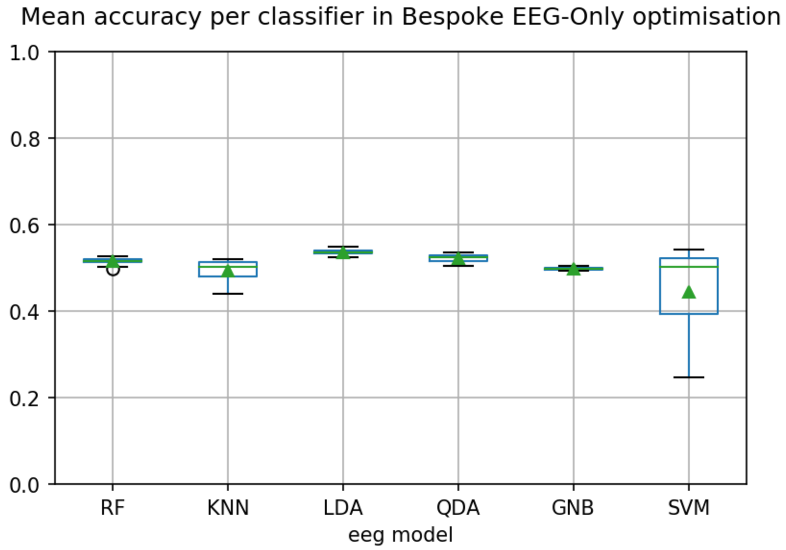


Figure 5.21: Mean Development Set accuracies achieved by different models in optimisation of Bespoke Unimodal EEG system

By observation LDAs appear to offer consistently higher performance than other classifier types. This superiority can be verified statistically through pairwise comparisons between classifier choices. Of the various post-hoc tests for multiple comparisons, options here are limited for multiple reasons. The nature of the CASH optimisation procedure means that more promising hyperparameter choices were explored with more of the optimiser's budget; the six groups here are not of equal sample size. Additionally, it would not be safe to assume each classifier was equally sensitive to the optimisation of its hyperparameters. Some classifiers' performances may be more heavily influenced by tuning than others and hence have exhibited a wider range of performances throughout the optimisation process; the usual assumption of equal variances between groups is therefore also violated. Finally, considering the previously noted unequal occurrence rates of each classifier type, given these groups are drawin from only 100 optimisation iterations the sample sizes of some are likely to be low. The Dunnet T3 test [328] is recommended where variances are unequal and sample sizes both unequal and low [329] and thus is used here, as implemented in the *PMCMRplus* package [330] in *R* [299]. The Dunnett T3 pairwise comparison results presented in Table 5.17 indicate there was indeed a significant difference at the $\alpha = 0.05$ level between the performance of LDAs and that of all other algorithms except Support Vector Machines.

Hypothesis	t value	p value
GNB – LDA	-29.624	$<$**2.22e-16**
KNN – LDA	-5.099	**0.0059**
QDA – LDA	-4.065	**0.0275**
RF – LDA	-8.697	**2.99e-06**
SVM – LDA	-2.873	0.1717
KNN – GNB	-0.545	1.0000
QDA – GNB	7.086	**0.0003**
RF – GNB	6.876	**3.26e-05**
SVM – GNB	-1.667	0.7542
QDA – KNN	3.239	0.0784
RF – KNN	2.530	0.2737
SVM – KNN	-1.472	0.8606
RF – QDA	-1.788	0.6809
SVM – QDA	-2.428	0.3248
SVM – RF	-2.210	0.4316

Table 5.17: Pairwise comparisons of EEG classifier choices in optimisation of Bespoke EEG-Only systems using Dunnett's T3 test for multiple comparisons with unequal variances

As the most promising candidate model for Bespoke Unimodal EEG classification, LDAs were trialled more frequently in optimisation, accounting for 44 of the 100 iterations. It could hence be posited that this greater budget for hyperparameter fine-tuning could account for the dominance of LDAs. The relatively low variance in LDA scores would however indicate this not to be the case; even the weakest LDA was more accurate than many other models and the performance differential between it & the strongest LDA was low.

Indeed, there did not appear to be any significant effects of the values chosen for hyperparameters (detailed in 5.3.3.3) of Bespoke EEG-LDAs. A one-way ANOVA between the choice of LDA solver and accuracy indicated no significant differences among group means & a pairwise Dunnett T3 comparison between them verifies a lack of significant effects between paired solver options; for brevity these are presented in Appendix A as Figure A.1 and Table A.1 respectively. The Pearson Correlation Coefficient (PCC) suggests only a weak and statistically insignificant ($p = 0.134$) linear trend between the Shrinkage hyperparameter and accuracy, & similarly Spearman's rho suggests no significant monotonic relationship, as per Figure 5.22[16]. Both Pearson's and Spearman's tests are performed here, and for all other correlation tests of models' hyperparameters hereafter in 5.5.4 & in Appendix A. To control the risk of a Type I error arising from this simultaneous testing, Bonferroni adjustments are applied; as there are two hypotheses the reported p-values are doubled. The Bonferroni correction process is known to be conservative in cases where hypotheses are themselves related [331, 332], which is likely to apply here — a linear correlation will logically often imply a rank correlation — and thus the adjusted p-values reported can be considered upper bounds for the actual p-values.

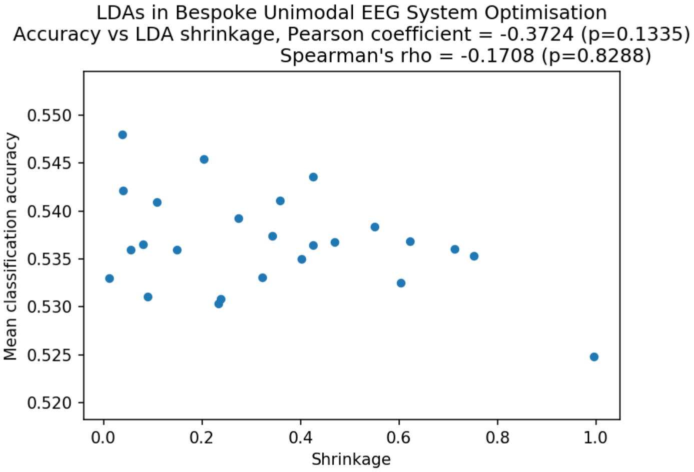
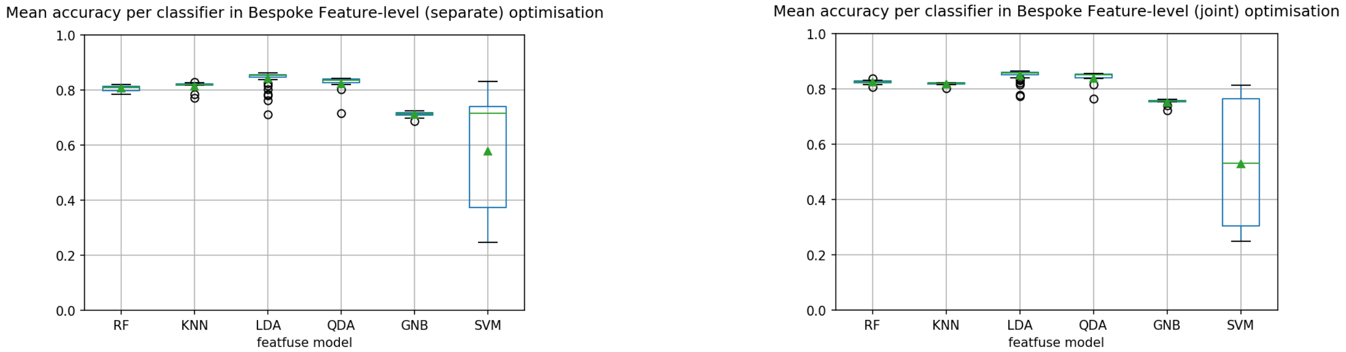


Figure 5.22: Mean Development Set accuracy against LDA Shrinkage in CASH optimisation of Bespoke Unimodal EEG system. Reported p-values adjusted by Bonferroni correction.

This apparent insensitivity to the tuning of its hyperparameters suggests that their greater optimisation budget was not in fact the proximate cause of LDA's superior performance in Bespoke Unimodal EEG-Only systems. Rather, these results corroborate the established preference among EEG-BCI literature for the LDA classifier and provide an evidential basis for their superiority.

[16] It is not assumed that LDAs using the Least Squares Solution solver and those using Eigenvalue Decomposition would have notably different relationships with the degree of shrinkage. This hyperparameter was thus shared among both these LDA subtypes during optimisation as described in 5.3.3.3, and scores from LDAs of these two solvers are pooled when modelling the effect of shrinkage here and henceforth. For completeness, correlation coefficients separated by solver are presented in Appendix A as Table A.6.

Such preference for LDA classifiers is seen also in Bespoke Feature-Level Fusion systems, again both in terms of their "best-in-class" configurations as in Table 5.3 and by visual inspection of per-classifier group means: Figure 5.23 presents accuracies grouped by machine learning algorithm for both Separate-Selection and Joint-Selection subtypes of the Feature-Level Fusion architecture. The Dunnet T3 all-vs-all comparisons presented in Table 5.18 demonstrate this dominance of LDAs over other Feature-Level Fusion classifiers — inclusive of SVMs, in contrast to the aforementioned Unimodal EEG systems — to be significant in all cases except the Quadratic Discriminant Analysis. The QDA it should be recalled is itself a close relation of the LDA [281], as described above in 5.3.3.4.



(a) Separate selection

(b) Joint selection

Figure 5.23: Development Set accuracies achieved by different models in CASH optimisation of Bespoke feature-level fusion system with separate feature selection (left) & joint feature selection (right)

Hypothesis	t value	p value
GNB – LDA	-23.221	<**2.22e-16**
KNN – LDA	-3.896	**0.0093**
QDA – LDA	-1.633	0.7776
RF – LDA	-6.403	**1.14e-06**
SVM – LDA	-3.955	**0.0280**
KNN – GNB	15.685	**8.72e-12**
QDA – GNB	9.473	**2.47e-05**
RF – GNB	19.788	<**2.22e-16**
SVM – GNB	-2.002	0.5467
QDA – KNN	0.663	0.9999
RF – KNN	-1.236	0.9561
SVM – KNN	-3.530	0.0562
RF – QDA	-1.401	0.8946
SVM – QDA	-3.617	**0.0438**
SVM – RF	-3.417	0.0677

(a) Separate selection

Hypothesis	t value	p value
GNB – LDA	-20.536	<**2.22e-16**
KNN – LDA	-9.465	**4.13e-12**
QDA – LDA	-1.267	0.9421
RF – LDA	-6.631	**3.43e-06**
SVM – LDA	-4.702	**0.0084**
KNN – GNB	15.840	**2.72e-10**
QDA – GNB	9.479	**1.08e-06**
RF – GNB	15.327	**6.06e-13**
SVM – GNB	-3.267	0.0865
QDA – KNN	2.412	0.3240
RF – KNN	1.701	0.7387
SVM – KNN	-4.241	**0.0176**
RF – QDA	-1.717	0.7260
SVM – QDA	-4.510	**0.0114**
SVM – RF	-4.320	**0.0155**

(b) Joint selection

Table 5.18: Pairwise comparisons of classifier choices in optimisation of Bespoke Feature-Level Fusion systems with separate feature selection (left) & joint feature selection (right) using Dunnett's T3 test

By contrast, in optimisation of the Bespoke single-mode EMG system LDAs offered mean accuracies not significantly different to those of other models — except the Gaussian Naïve Bayes which, as the pairwise Dunnett T3 contrasts seen in Table 5.19 illuminate, was consistently weaker than others. That the superior accuracy of LDAs in EEG classification is not reflected here in EMG models suggests this effect not to be due to any unforeseen intrinsic advantage offered to them in these experiments, thus strengthening the support these results give to the common belief among literature of their suitability to EEG-based modelling.

Hypothesis	t value	p value
GNB – SVM	-5.392	**6.12e-05**
KNN – SVM	0.134	1.0000
LDA – SVM	0.232	1.0000
QDA – SVM	0.198	1.0000
RF – SVM	-0.411	1.0000
KNN – GNB	25.966	$<$ **2.22e-16**
LDA – GNB	10.974	**1.45e-08**
QDA – GNB	17.387	**1.11e-15**
RF – GNB	27.199	$<$ **2.22e-16**
LDA – KNN	0.236	1.0000
QDA – KNN	0.214	1.0000
RF – KNN	-2.841	0.1234
QDA – LDA	-0.091	1.0000
RF – LDA	-1.315	0.9319
RF – QDA	-1.992	0.5415

Table 5.19: Pairwise Dunnett's T3 comparisons of EMG classifier choices in optimisation of Bespoke EMG-Only systems

As in EEG systems, the choice of solver had no significant effect on Bespoke EMG-LDAs' accuracy (Figure A.2 & Table A.2 of Appendix A). Unlike their EEG counterparts however Figure 5.24 demonstrates the shrinkage of Bespoke EMG-LDAs was perfectly negatively correlated with accuracy (Spearman's ρ = -1).

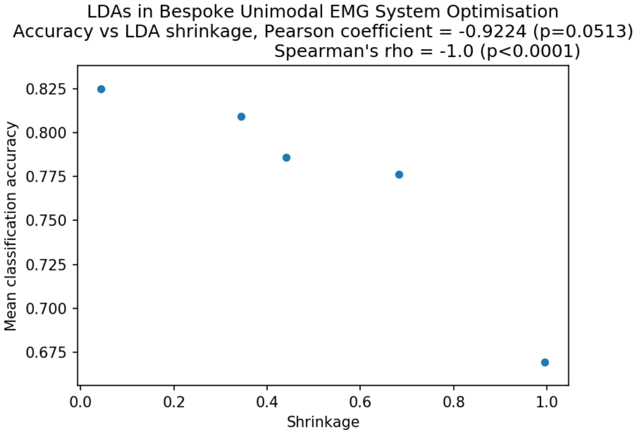
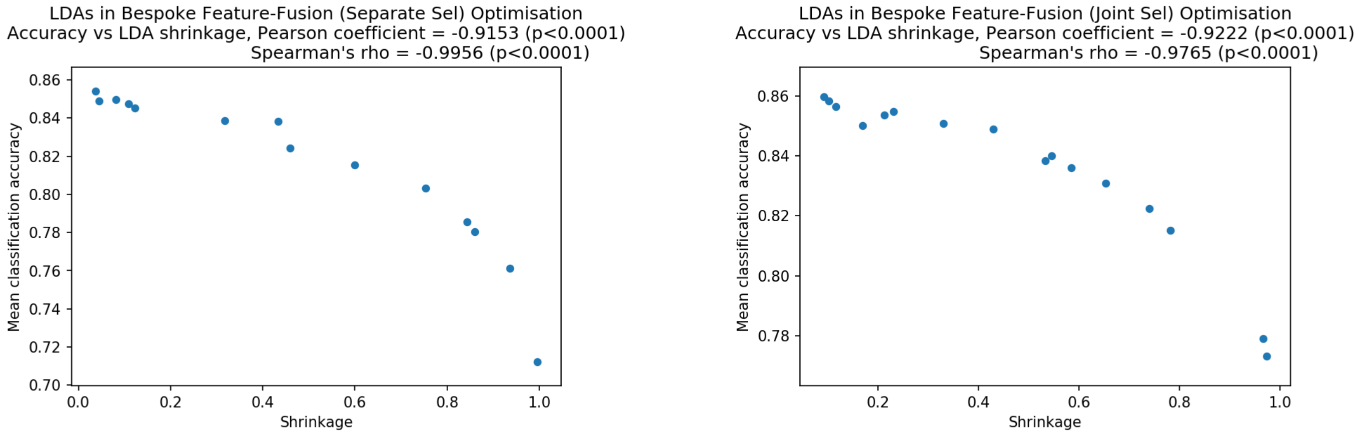


Figure 5.24: Mean accuracy against LDA Shrinkage in CASH optimisation of Bespoke Unimodal EMG system. Reported p-values adjusted by Bonferroni correction.

Shrinkage and performance were similarly correlated for Bespoke Feature-Level Fusion LDAs; with a PCC of -0.9153 (p<0.0001) in the Separate Selection variant & -0.9222 (p<0.0001) in the Joint Selection as seen in Figure 5.25. Considering the noted correlation among Unimodal EMG-LDAs and the lack of such an effect in EEG-LDAs, this could plausibly be due to the presence of EMG among the Feature Fusion dataset.



(a) Separate selection

(b) Joint selection

Figure 5.25: Accuracy against LDA Shrinkage in optimisation of Bespoke Feature-Level Fusion systems with Separate feature selection (left) & Joint feature selection (right)

In the optimisation of both subject-specific Feature-Level Fusion variants, the solver used for LDAs (the dominant algorithm, as noted above) again did not have a significant effect on their classification accuracy, as seen in Figure A.3 & Table A.3 of Appendix A. In contrast to the apparent total ambivalence of both Bespoke EEG and EMG LDAs to the choice of solver however, there does appear to be a slight observable trend in favour of the Singular Value Decomposition method. While not found to be a statistically significant ($0.1 > p > 0.05$) pattern here, it is interesting for both EMG & EEG LDAs to be solver-agnostic yet where LDAs use EMG & EEG simultaneously there to present the possibility of a preference. Indeed, as seen in Table 5.3 the single best-performing Bespoke Feature-Level Fusion systems of both subtypes identified in CASH optimisation utilised the SVD solver. Future work may find interest in exploring the impact of LDA solver choice on Feature-Level EMG-EEG Fusion in greater depth.

Furthermore, as outlined in 5.3.3.3 the shrinkage parameter of an LDA controls the degree of estimation involved in its computing of covariance matrices; in the case of zero shrinkage the entire empirical covariance matrix is used, while with a shrinkage of 1 it is estimated from the diagonal matrix of variance [284, 285]. An LDA using the SVD solver however calculates log-posteriors while bypassing computation of the covariance matrix entirely; the shrinkage technique is not used. This possibility that SVD solvers may be preferred for Feature-Level Fusion, when taken together with the finding that lower shrinkage values were preferred for non-SVD solvers, could imply that in this case the diagonal matrices of variances were poor estimators of the covariance matrices — and thus the more these were relied on by Feature Fusion LDAs, the more their performance was degraded.

Generalist

Table 5.4 shows that remarkably LDAs were the consistent single favourite machine learning model in all Generalist Fusion architectures. It should be recalled here that as discussed in 5.3.3.6, for pragmatic reasons Support Vector Machines (a popular choice among Bespoke systems for classifying EMG data, as per Table 5.3), were excluded from consideration by Generalist optimisation. Interesting is that making a typically high-performing nonlinear model[17] unavailable in this way did not result in Generalist systems preferring other nonlinear models such as the Random Forest, but rather the LDA, a linear classifier. While the impact of models' linearity is not a focus of this work, future research could investigate this further by exploring the use of kernelised nonlinear extensions of the LDA, such as the kernel Fisher Discriminant Analysis [333].

In optimisation of Generalist Unimodal EEG systems, consistent with their Bespoke equivalents, LDAs not only provided the single best-performing configuration but proved significantly more accurate than other classifier choices, as demonstrated by Table 5.20's Dunnett T3 test results.

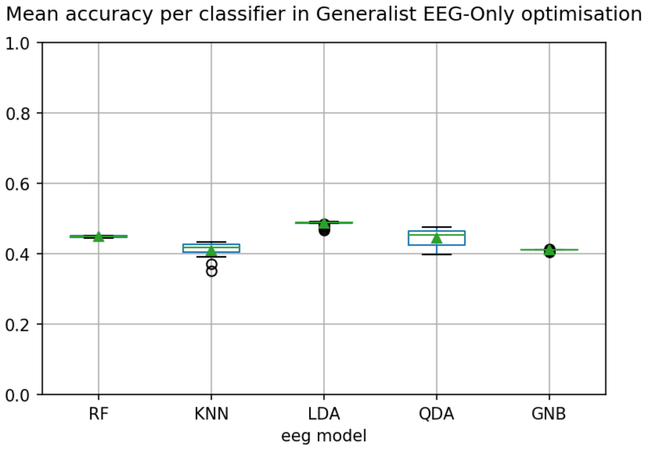


Figure 5.26: Accuracies achieved by different models in optimisation of Generalist Unimodal EEG classifier

Hypothesis	t value	p value
GNB – LDA	-80.311	$<$ **2.22e-16**
KNN – LDA	-11.184	**3.11e-06**
QDA – LDA	-5.405	**0.0020**
RF – LDA	-41.049	$<$ **2.22e-16**
KNN – GNB	-0.210	1.0000
QDA – GNB	4.437	**0.0090**
RF – GNB	42.334	$<$ **2.22e-16**
QDA – KNN	3.457	**0.0213**
RF – KNN	5.689	**0.0009**
RF – QDA	0.473	0.9999

Table 5.20: Pairwise comparisons of classifier choices in CASH optimisation of Generalist Unimodal EEG systems using Dunnett's T3 test

Interestingly here a correlation between shrinkage and performance is found despite none having been observable in Bespoke EEG-LDAs, with a PCC of -0.9189 significant at the $\alpha = 0.05$ level. As seen in Figure 5.27 the relationship does not appear wholly linear, nevertheless Spearman's ρ indicates a rank correlation coefficient of 0.8828. As in the Bespoke case however, the accuracy of Generalist EEG-Only LDAs continues to not be significantly influenced by the choice of solver (Appendix A, Figure A.4 & Table A.4).

[17] While not inherently nonlinear, the SVM is made capable of nonlinear modelling by use of the RBF kernel (see 5.3.3.6).

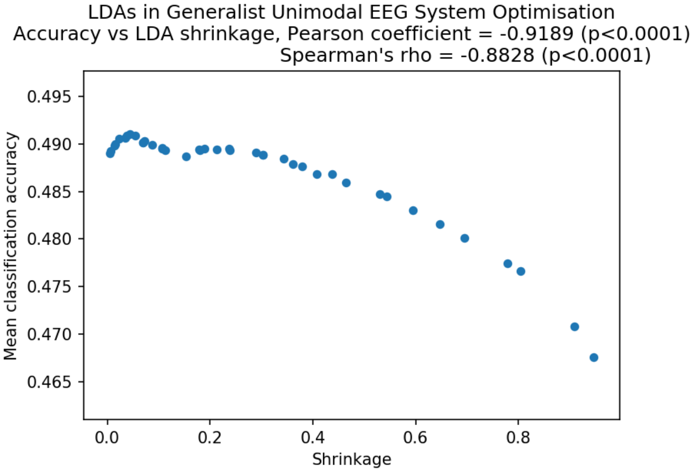


Figure 5.27: Development Set accuracy correlated against LDA Shrinkage in optimisation of Generalist Unimodal EEG system. Reported p-values adjusted by Bonferroni correction.

As seen in Figure 5.28, in EMG LDAs likewise offered more accurate Generalist single-mode classification than all other model types, and similar to the Bespoke EMG case described above the GNB model was consistently outclassed. Table 5.21 demonstrates that these were in fact the only significant trends among Generalist Unimodal EMG systems. Again a negative correlation between shrinkage & performance (PCC of -0.9194, p<0.0001) was found. The only significant effect among solvers of Generalist EMG-LDAs was a preference for SVD over the Least Squares Solution; neither were significantly different from the Eigenvalue Decomposition method. These can be seen in Appendix A, Figure A.6 & Table A.5.

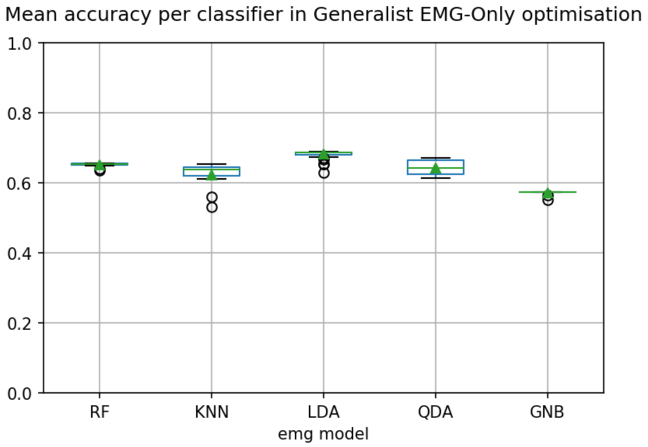


Figure 5.28: Accuracies achieved by different models in optimisation of Generalist Unimodal EMG classifier

Hypothesis	t value	p value
GNB – LDA	-43.600	< **2.22e-16**
KNN – LDA	-5.331	**0.0022**
QDA – LDA	-6.143	**0.0005**
RF – LDA	-12.674	**2.08e-14**
KNN – GNB	4.561	**0.0060**
QDA – GNB	11.200	**1.64e-06**
RF – GNB	29.348	< **2.22e-16**
QDA – KNN	1.649	0.6538
RF – KNN	2.611	0.1756
RF – QDA	1.322	0.8511

Table 5.21: Pairwise comparisons of EMG classifier choices in optimisation of Generalist Unimodal EMG systems using Dunnett's T3 test

The Dunnett T3 pairwise comparisons in Tables 5.22a and 5.22b demonstrate LDAs were the statistically significantly superior choice for Generalist Feature-Level Fusion. Tables 5.23a and 5.23b indicate that

Hypothesis	t value	p value
GNB − LDA	-37.403	<**2.22e-16**
KNN − LDA	-7.137	**0.0001**
QDA − LDA	-5.404	**0.0015**
RF − LDA	-14.335	**3.60e-14**
KNN − GNB	3.856	**0.0204**
QDA − GNB	7.076	**7.47e-05**
RF − GNB	18.210	<**2.22e-16**
QDA − KNN	1.842	0.5168
RF − KNN	2.680	0.1507
RF − QDA	0.364	1.0000

(a) Separate selection

Hypothesis	t value	p value
GNB − LDA	-80.336	<**2.22e-16**
KNN − LDA	-7.702	**2.56e-05**
QDA − LDA	-6.047	**0.0005**
RF − LDA	-30.827	<**2.22e-16**
KNN − GNB	4.778	**0.0028**
QDA − GNB	9.788	**1.48e-06**
RF − GNB	30.486	<**2.22e-16**
QDA − KNN	2.327	0.2317
RF − KNN	1.621	0.6731
RF − QDA	-1.634	0.6650

(b) Joint selection

Table 5.22: Pairwise comparisons using Dunnett's T3 test between classifier choices in optimisation of Generalist Joint Selection (right) and Separate selection (left) Feature-Level Fusion

where EMG & EEG features were selected independently prior to joining the data there were no significant differences in accuracy between LDA solvers, but where selected from joint data the SVD was weaker.

Hypothesis	t value	p value
LSQR − SVD	-1.193	0.5712
Eigen − SVD	-1.716	0.2574
Eigen − LSQR	0.613	0.9007

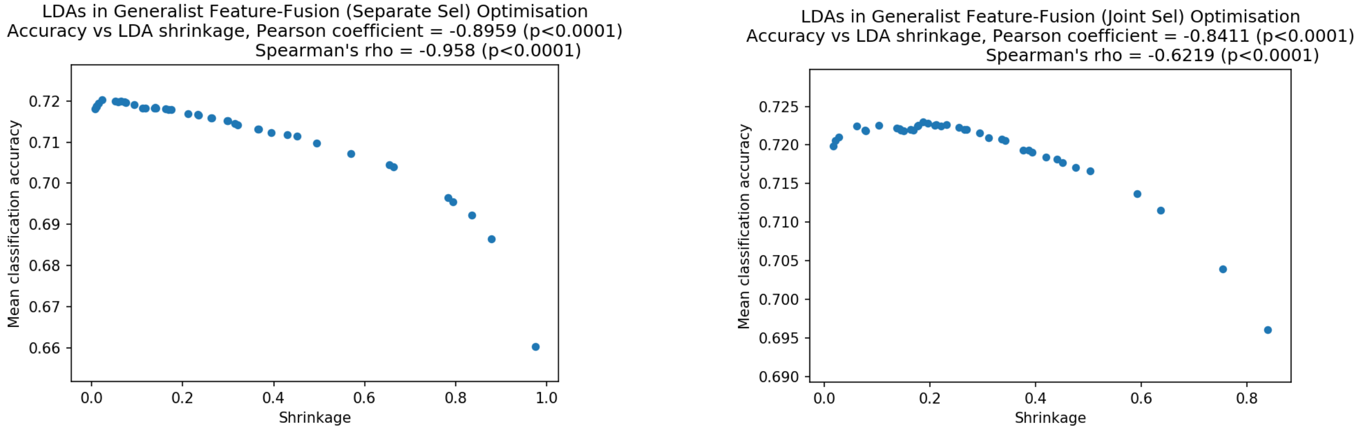
(a) Separate selection

Hypothesis	t value	p value
LSQR − SVD	9.555	**8.66e-12**
Eigen − SVD	8.645	**0.0002**
Eigen − LSQR	-0.156	0.9980

(b) Joint selection

Table 5.23: Pairwise Dunnett T3 comparisons of LDA solvers in optimisation of Generalist Feature-Level Fusion systems with Separate (left) & Joint (right) feature selection

Here there was again a clear negative correlation between LDAs's Shrinkage and predictive performance in both Separate (PCC: 0.8959, p<0.0001) and Joint Selection (PCC: 0.8411, p<0.0001) cases, as per Figure 5.29. Given that the overlap coefficient between EEG features selected separately and those appearing in joint selections was approximately 0.5 (discussed in 5.5.3 above), it is interesting that the impact of both solver choice and shrinkage differs between LDAs of the two Feature-level Fusion subtypes. Such impacts of feature selection strategies on Feature-Level Fusion are left for future work to explore in greater depth.



(a) Separate selection

(b) Joint selection

Figure 5.29: Development Set Accuracy against LDA Shrinkage in optimisation of Generalist Feature-Level Fusion systems with Separate (left) & Joint (right). Reported p-values adjusted by Bonferroni correction.

5.5.4.2 Support Vector Machines in EMG

Figure 5.30 presents the classification accuracies achieved by different machine learning algorithms in optimisation of the Bespoke single-mode EMG-only system. While the single "best-in-class" configuration for this baseline architecture made use of a Support Vector Machine (Table 5.5), it is notable that as seen previously in Table 5.19 there was no significant difference in group means between SVMs' accuracies and those of other models, except the GNB which consistently underperformed.

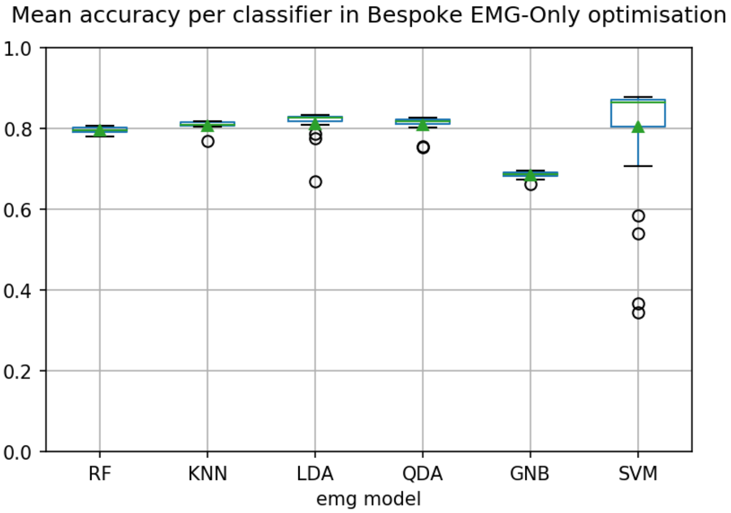
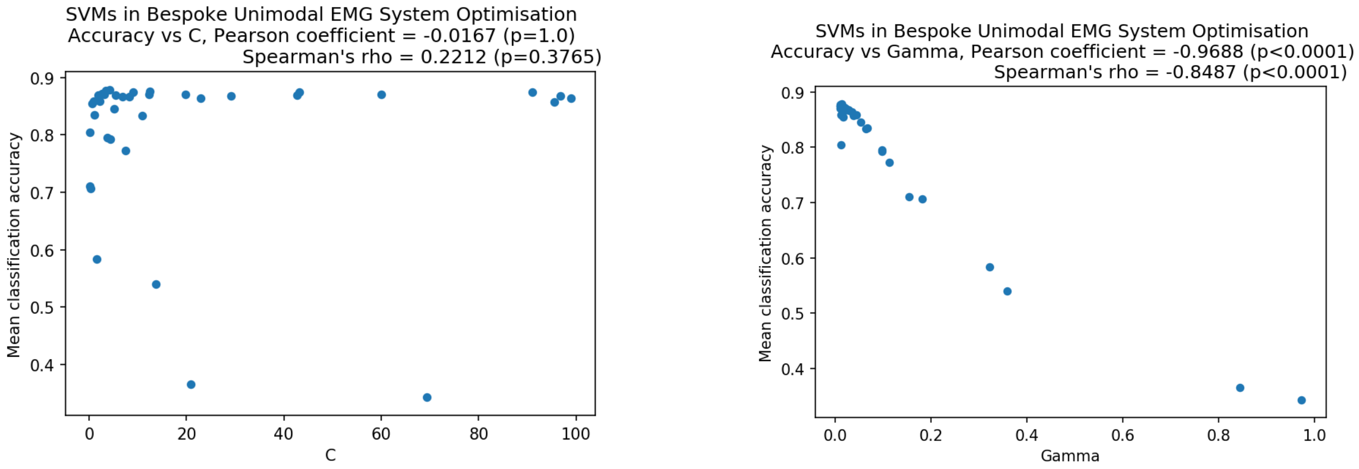


Figure 5.30: Accuracies achieved by different models in optimisation of Bespoke Unimodal EMG system

It should be noted however that EMG-SVMs' accuracies were highly dependent on their hyperparameters — sufficiently so that as seen in Figure 5.30 they provided both the overall highest and lowest Unimodal EMG classification accuracies when initialised with different configurations of hyperparameter values. Figures 5.31a and 5.31b show that while there was no discernible trend in classification accuracy according to the C value, γ had a clear influence on the SVMs' performance and appears to account for much of the variance in their scores. A narrower range for gamma, such as an exclusion of values > 0.2, could perhaps have resulted in an improved EMG-SVM performance which was consistently within or exceeding the interquartile range seen in these results (Figure 5.30), i.e. $> 70\%$, and that the variance in accuracy would accordingly be so reduced as to result in SVMs systematically outperforming other models[18]. It may be pertinent to note that while not wholly consistent, as there is an observable downtick in performance among the very lowest values of γ in Figure 5.31b, in general the better-performing lower γs are closer in value to that which would be obtained by the $\gamma = \frac{1}{N_{features}}$ method of γ determination (which would here $= 0.1136$) used by Ameri et al. [292] as noted in 5.3.3.6.

This high susceptibility to hyperparameter tuning, taken alongside the evident ability of EMG-SVMs to reach high performance levels, perhaps suggests that despite the lack of significant difference among group

[18] Similar trends (a lack of influence of C and a negative correlation between performance and γ) are reflected in Feature-Level Fusion, and even in Unimodal EEG SVMs, as can be seen in A.1.1.4



(a) C

(b) gamma

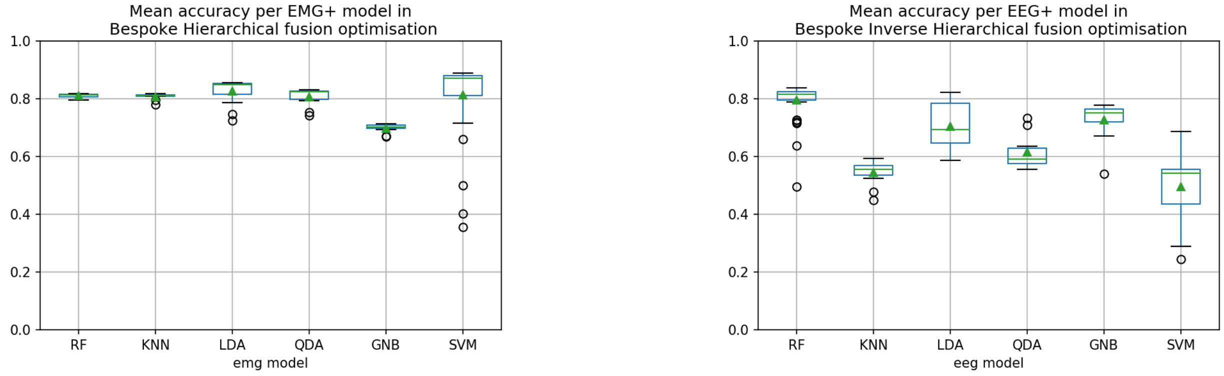
Figure 5.31: Influence of SVM hyperparameters C (left) and gamma (right) on Bespoke Unimodal EMG-SVM accuracies in optimisation.
Reported p-values adjusted by Bonferroni correction.

means (Table 5.19), SVMs may well be a strong candidate for EMG classification — just one which may require more hyperparameter tuning to reach competitive levels.

5.5.5 Preferences in top-level models among Hierarchical systems

As a less established fusion strategy the Hierarchical (& Inverse Hierarchical) architecture, while not found to provide the most accurate classification here, merits some inspection for the benefit of any future work which seeks to develop or extend this approach further.

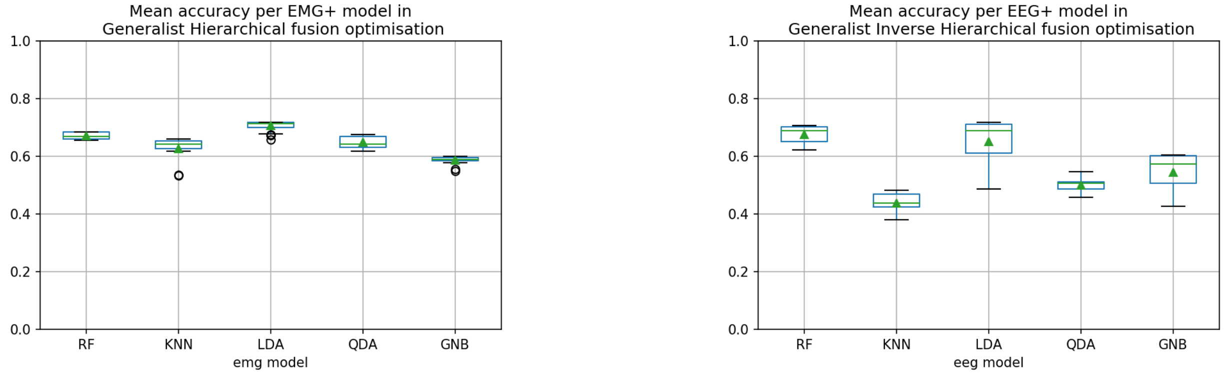
Of particular interest is the “higher-ranking” or “top-level” classifier of such a system — whether that be an “EMG+” model primarily classifying EMG but supplemented with EEG predictions as in the Hierarchical Fusion case (Figure 5.3), or an “EEG+” model where EEG is supplemented with EMG as in the Inverse Hierarchical (Figure 5.4) — and the effect of its model choice on performance. The extent of a Hierarchical fusion system’s ability to fully exploit the information provided by both its data modalities will depend upon the ability of this top-level model to learn neither to ignore the classwise probabilities predicted by the lower-ranking model, nor to be beholden to them, but instead to use them in conjunction with its other features in such a way that they are drawn upon when likely to be reliable and have lesser influence when not.



(a) Hierarchical (where EMG is top rank)

(b) Inverse Hierarchical (where EEG is top rank)

Figure 5.32: Mean accuracies across Development Subjects achieved by different top-level models in CASH optimisation of Bespoke Hierarchical & Inverse Hierarchical systems



(a) Hierarchical (where EMG is top rank)

(b) Inverse Hierarchical (where EEG is top rank)

Figure 5.33: Mean accuracies across Development Subjects achieved by different top-level models in CASH optimisation of Generalist Hierarchical & Inverse Hierarchical systems

Figures 5.32 and 5.33 present visually the accuracies reached throughout optimisation by various different higher-ranking classifiers in the Bespoke and Generalist cases respectively, and Tables 5.24 & 5.25 the results of pairwise Dunnett T3 comparisons between classifiers.

Hypothesis	t value	p value
GNB – SVM	-5.540	**2.98e-05**
KNN – SVM	-0.302	1.0000
LDA – SVM	0.534	1.0000
QDA – SVM	-0.313	1.0000
RF – SVM	-0.185	1.0000
KNN – GNB	20.072	**2.22e-16**
LDA – GNB	11.501	**2.78e-10**
QDA – GNB	10.433	**5.20e-07**
RF – GNB	24.861	<**2.22e-16**
LDA – KNN	1.683	0.7522
QDA – KNN	-0.078	1.0000
RF – KNN	0.586	1.0000
QDA – LDA	-1.368	0.9227
RF – LDA	-1.517	0.8469
RF – QDA	0.335	1.0000

(a) Hierarchical (where EMG is top rank)

Hypothesis	t value	p value
GNB – RF	-3.223	0.0678
KNN – RF	-16.336	<**2.22e-16**
LDA – RF	-3.267	0.0745
QDA – RF	-8.410	**1.44e-05**
SVM – RF	-6.917	**0.0003**
KNN – GNB	-7.877	**2.27e-06**
LDA – GNB	-0.656	0.9999
QDA – GNB	-4.029	**0.0093**
SVM – GNB	-4.938	**0.0030**
LDA – KNN	5.525	**0.0010**
QDA – KNN	3.095	0.0870
SVM – KNN	-1.095	0.9793
QDA – LDA	-2.735	0.1624
SVM – LDA	-4.189	**0.0086**
SVM – QDA	-2.565	0.2399

(b) Inverse Hierarchical (where EEG is top rank)

Table 5.24: Pairwise comparisons of top-level models in optimisation of Bespoke Hierarchical (left) & Inverse Hierarchical (right) systems using Dunnett's T3 test

Hypothesis	t value	p value
GNB – LDA	-25.681	<**2.22e-16**
KNN – LDA	-6.143	**0.0005**
QDA – LDA	-9.242	**4.62e-07**
RF – LDA	-8.802	**1.68e-07**
KNN – GNB	2.949	0.0941
QDA – GNB	8.933	**5.81e-08**
RF – GNB	16.719	<**2.22e-16**
QDA – KNN	1.648	0.6554
RF – KNN	3.501	**0.0380**
RF – QDA	3.538	**0.0207**

(a) Hierarchical (where EMG is top rank)

Hypothesis	t value	p value
GNB – LDA	-5.272	**0.0003**
KNN – LDA	-15.902	<**2.22e-16**
QDA – LDA	-12.973	<**2.22e-16**
RF – LDA	1.716	0.6018
KNN – GNB	-5.341	**0.0005**
QDA – GNB	-2.368	0.2425
RF – GNB	6.473	**3.06e-05**
QDA – KNN	5.811	**9.00e-05**
RF – KNN	17.706	<**2.22e-16**
RF – QDA	15.042	**1.71e-13**

(b) Inverse Hierarchical (where EEG is top rank)

Table 5.25: Pairwise Dunnett T3 comparisons of top-level models in optimisation of Generalist Hierarchical (left) & Inverse Hierarchical (right) systems

Immediately observable is that the pattern of classifier-wise performance in the Bespoke Hierarchical case (Figure 5.32a), wherein the higher-ranking classifier primarily received EMG data, is remarkably similar to that of the Bespoke Unimodal EMG system (Figure 5.30). A similar resemblance can be seen between the Generalist Hierarchical's "supplemented" EMG models (Figure 5.33a) and the Generalist Unimodal EMG (Figure 5.28). Indeed as seen in 5.5.1.2 the top-level models of the optimal Hierarchical systems implemented classifiers of the same types as the optimal Unimodal EMG systems in both Bespoke and Generalist cases; the Hierarchical optimisation has selected for models well-suited to classifying EMG data.

This suggests a degree of indifference in Hierarchical systems to their component EEG classifiers; that in effect the top-rank model is learning to ignore the EEG predictions. In Appendix A, Figures A.13a and A.14a demonstrate that consistent with this inference there were no significant identifiable effects of the lower-ranking EEG classifier choice over Hierarchical systems' accuracies. It may be that the collapsing of information carried by EEG data into classwise probabilities, which considering Unimodal EEG systems' lesser accuracies than those of Unimodal EMG (Table 5.5) are likely to be somewhat low in predictive power, allows it to be more easily "ignorable" than in a strategy such as the Feature-Level Fusion wherein classifiers see EMG features alongside a less condensed representation of EEG.

Not explored here is the possibility that the low-rank classifier choice could have local effects specific to a given high-rank classifier option; that certain top-level classifiers were better served by certain types of lower-ranked model. To enable such an investigation in depth a more exhaustive combinatorial search of the low- and high- rank model choices would be necessary; this seems an obvious place for future research into Hierarchical Fusion strategies to begin investigation.

Performance differences between top-level classifier choice in the Inverse Hierarchical architecture, seen in Figure 5.32b & Table 5.24b for the Bespoke case and Figure 5.33b & Table 5.25b for Generalist systems, were more dramatic. Interestingly these trends do not reflect those observed in Unimodal EEG classification (Tables 5.17 & 5.20) to the same extent that trends among Hierarchical top-level modelling choices reflected those of Unimodal EMG, though some features such as the suitability of LDA classifiers are observable.

It would follow from this that the ways by which an Inverse Hierarchical system's top-level classifier are learning from its available data are distinct from those of a Unimodal EEG system. This can reasonably be assumed to be due to the presence of the low-level EMG classifier's predictions, which are as noted likely to be of reasonable accuracy. Indeed, Figures A.13b and A.14b illuminate that the component EMG models did have an influence over Inverse Hierarchical systems' accuracies in a way that component EEG models did not over the Hierarchical.

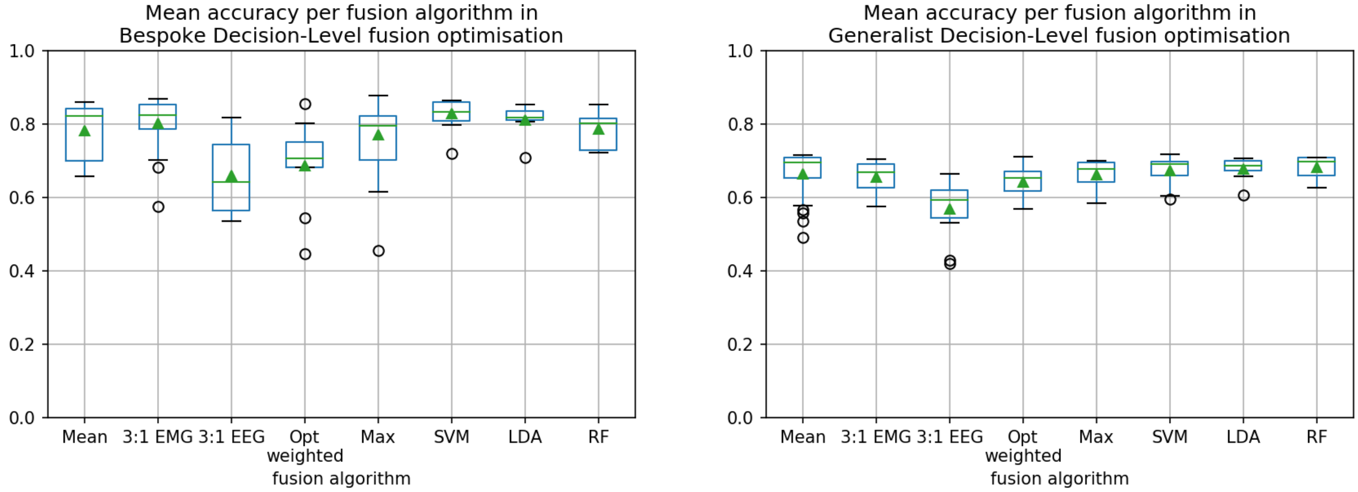
This is not to suggest however that Inverse Hierarchical systems' performance was driven solely by the accuracy of their lower-ranking EMG classifier. The overally mean Development Set accuracy of both Bespoke & Generalist optimiser-identified systems was greater than that of their respective EMG component models. In the Bespoke Inverse Hierarchical system, whose mean system-level accuracy was 83.68 per Table 5.3, the constituent EMG-QDA had a marginally lower classification accuracy of 82.11% (congruous with the performance of Bespoke Unimodal EMG-QDAs seen in Figure 5.30). In the Generalist case, where the Inverse Hierarchical system's mean fusion accuracy was 71.68 as per Table 5.4, the accuracy of the constituent EMG-LDA was 68.83% — a difference nearly double that of the Bespoke case.

Perhaps noteworthy is that across both Hierarchical and Inverse Hierarchical architectures, non-linear top-level models were preferred for Bespoke systems (a Random Forest and an RBF-kernelised SVM), whereas linear models (an LDA in both approaches) were optimal in the Generalist case. It is interesting that this distinction is along the Bespoke – Generalist line, rather than Hierarchical – Inverse. Considering the relatively high accuracy of Unimodal EMG systems, and thus the expected high informativity of their predicted class probabilities, it could be speculated that the Inverse Hierarchical optimisation would be expected to recognise this & hence prefer nonlinear top-level models capable of capturing the subtleties of the EMG-derived probability distributions. Of course, nonlinearity of a selected top-level model would not necessarily guarantee it not to risk insufficiently prioritising the EMG-based predictions — nor indeed to avoid overfitting by modelling primarily around the EMG predictions and using EEG features only to add complexity to the model which would not generalise to new data.

5.5.6 Decision-Level Fusion

5.5.6.1 Decision Fusion Algorithms

Figures 5.34a and 5.34b present the scores achieved by the various decision fusion algorithms in optimisation of Bespoke and Generalist Decision-Level Fusion systems.



(a) Bespoke

(b) Generalist

Figure 5.34: Mean Development Set accuracies achieved in CASH optimisation of Decision-Level Fusion systems grouped by Decision Fusion algorithm

Here, all trials using each given fusion algorithm are grouped together. However, in reality it is not necessarily a safe assumption that the various algorithms would interact with different EMG & EEG classifier options in the same way. Hence to assess the presence of performance differences between fusion algorithms, the EMG & EEG classifier choices need be treated as blocking factors. They are thus modelled as random effects in a linear mixed-effects model, and Tukey's Honestly Significant Difference test used for post-hoc pairwise comparison between fusion algorithms using the *glht* function of *R*'s *multcomp* package. As presented in Tables 5.26a and 5.26b, Tukey's HSD demonstrated significant differences in accuracies to be found only between certain algorithms (for brevity, those not found significant are omitted here but can be seen in Appendix A).

Unsurprisingly given the respective performances of single-mode EMG & EEG classifiers (5.5.1.1), a Weighting Average fixed in favour of EEG is routinely a weak choice of decision fusion algorithm here; a result consistent with the findings of Tryon et al. [141] wherein this strategy performed poorly in identifying the onset of elbow movement. Beyond this, other trends are somewhat less clear. Those stacking-based algorithms which use linear meta-models — and **not** the nonlinear Random Forest — appear more suitable than various rule-based algorithms for Generalist systems, but this strength seemingly does not translate to the Bespoke. While further investigation would be needed to draw conclusive inferences, it is speculated that this strength of stacking algorithms could indicate a polarisation of the informativity of EMG in the Generalist case. Suppose that EMG-based probability distributions were liable to either be highly accurate or

Hypothesis	Estimate	Std. Err	p value
3:1 EEG – Max	-0.10617	0.03053	0.0114
3:1 EMG – 3:1 EEG	0.14185	0.03213	<0.001
LDA – 3:1 EEG	0.16790	0.03842	<0.001
Mean – 3:1 EEG	0.13214	0.03608	0.0055
RF – 3:1 EEG	0.14428	0.03610	0.0016
SVM – 3:1 EEG	0.16705	0.03240	<0.001
Tuned WA – 3:1 EMG	-0.10004	0.03119	0.0281
Tuned WA – LDA	-0.12609	0.03651	0.0124
SVM – Tuned WA	0.12525	0.03140	0.0016

(a) Bespoke

Hypothesis	Estimate	Std. Err	p value
3:1 EEG – Max	-0.07725	0.01769	<0.001
LDA – Max	0.07235	0.01987	0.0062
3:1 EMG – 3:1 EEG	0.09392	0.01764	<0.001
LDA – 3:1 EEG	0.14998	0.01766	<0.001
Mean – 3:1 EEG	0.06401	0.01393	<0.001
Tuned WA – 3:1 EEG	0.05626	0.01700	0.0201
RF – 3:1 EEG	0.09805	0.01610	<0.001
SVM – 3:1 EEG	0.10440	0.01379	<0.001
Mean – LDA	-0.08597	0.01650	<0.001
Tuned WA – LDA	-0.09372	0.01937	<0.001
SVM – Mean	0.04038	0.01221	0.0199
SVM – Tuned WA	0.04813	0.01512	0.0300

(b) Generalist

Table 5.26: Pairwise comparisons using Tukey's HSD test of Decision-fusion algorithms in optimisation of Generalist (right) and Bespoke (left) Decision-level Fusion systems. Pairs in which significant effects ($p < 0.05$) were identified are included here; all pairs not presented saw no significant differences. Full pairwise comparisons, including corresponding z-values (omitted here for brevity), can be found in Appendix A, Tables A.7 & A.8.

highly inaccurate, rather than a moderately accurate "reasonable guess". If in some of those cases where EMG data were uninformative, the EEG data offered greater predictive power, an algorithm which could learn to identify when EMG probabilities were likely to be unreliable and could "choose between" the EMG & EEG models' decisions may be more suitable than a rule-based method which combined them mathematically.

For the purposes of this work the CASH optimisation procedure was used to allow modelling choices to be made in a fair, unbiased, & algorithmic way rather than to determine conclusively a superior approach. That these tests do not reveal an unambiguous groupwise "winner" decision algorithm is thus not of particular concern. Ultimately as in Tables 5.3 & 5.4 the Max rule has been selected for Bespoke Decision-Level Fusion systems and the Linear SVM for Generalists; these selections are not predicated on the chosen algorithms being consistently significantly more performant than all others. It does however preclude more nuanced conclusions from being drawn from these comparisons of decision-fusion algorithms. Nevertheless, in the context of prior work such as that of Cui et al. [145], which successfully fused EMG & EEG with both rule-based and metamodel-based methods to classify the intensity of lower limb movements, these results indicate the need for further research into Decision-Level Fusion methods for gesture classification. That these Decision-Level Fusion systems proved able to achieve accuracies commensurate with those of their competitor architectures demonstrates the merit of the approach. Future work may consider evaluating the methods used here along with other established late fusion strategies such as Bayesian fusion [334,335] further. An exhaustive gridsearch for optimisation may enable assessment of algorithms' compatibilities with different combinations of EMG & EEG classifier choices.

5.5.6.2 Component Models

By the same mechanism described above for the Decision Fusion algorithms, Tukey's method can further be used to assess the effects of EMG & EEG classifier choices on Decision-Level Fusion systems' accuracies, as presented in Tables 5.27 and 5.28. Again only significant effects are shown here; the complete pairwise comparisons can be seen as Tables A.9 and A.10 of Appendix A.

Hypothesis	Estimate	Std. Err	z value	p value
GNB – SVM	-0.08720	0.02393	-3.643	0.0036
KNN – GNB	0.11150	0.03092	3.606	0.0042
LDA – GNB	0.09967	0.03181	3.134	0.0206
QDA – GNB	0.10469	0.03010	3.478	0.0065

(a) Bespoke

Hypothesis	Estimate	Std. Error	z value	p value
GNB – LDA	-0.10299	0.01244	-8.276	<0.001
KNN – LDA	-0.03978	0.01228	-3.240	0.0103
RF – LDA	-0.06069	0.01211	-5.010	<0.001
KNN – GNB	0.06321	0.01534	4.121	<0.001
QDA – GNB	0.08160	0.01507	5.414	<0.001
RF – GNB	0.04229	0.01521	2.781	0.0422
RF – QDA	-0.03930	0.01410	-2.788	0.0413

(b) Generalist

Table 5.27: Significant differences ($p < 0.05$) identified by pairwise Tukey comparisons between EMG classifier choices on Bespoke (left) and Generalist (right) Decision-Level Fusion systems' mean Development Set accuracy in optimisation. Pairs not presented saw no significant differences; for full pairwise results see Appendix A Tables A.9a & A.9b.

Hypothesis	Estimate	Std. Err	z value	p value
GNB – RF	-0.10988	0.02720	-4.039	<0.001
SVM – GNB	0.08575	0.02911	2.945	0.0361

(a) Bespoke

Hypothesis	Estimate	Std. Err	z value	p value
GNB – LDA	-0.07840	0.01428	-5.488	<0.001
KNN – GNB	0.06796	0.01757	3.869	<0.001
QDA – GNB	0.04344	0.01585	2.741	0.0464
RF – GNB	0.09413	0.01469	6.410	<0.001
RF – QDA	0.05069	0.01322	3.834	0.0011

(b) Generalist

Table 5.28: Significant differences ($p < 0.05$) identified by pairwise Tukey comparisons between EEG classifier choices on Bespoke (left) and Generalist (right) Decision-Level Fusion systems' mean Development Set accuracy in optimisation. Pairs not presented saw no significant differences; for full pairwise results see in Appendix A Tables A.10a & A.10b.

In the Bespoke case, the use of a GNB for the component EMG model routinely resulted systems significantly less accurate than those using alternative EMG classifiers. The effects of other EMG classifiers were however not separable at the 95% confidence level; this is interestingly in keeping with their impact on Unimodal EMG accuracy as seen in Table 5.19 above. In the Generalist case meanwhile, systems using EMG-GNBs were again consistently the weakest classifiers, those with EMG-LDAs outperformed all but the use of EMG-QDAs. This likewise is coherent with earlier observations from Generalist single-mode EMG (Table 5.21 and indeed Feature-Level Fusion systems (Table 5.22).

Curiously, when considering the influence of EEG classifier choice on Bespoke Decision-Level Fusion, the only significant differences were systems using EEG-RFs or EEG-SVMs both outperforming those with EEG-GNBs. Among Generalist Decision-Level fusion systems the EEG-GNB, much as the EMG-GNB did, appeared the weakest choice of classifier. The EEG-RF, while appearing in the single most accurate system, contributed to Decision-Level Fusion systems only significantly stronger than those using the EEG-QDA.

Evidently it was not the case here that the "best-in-class" unimodal classifiers (EMG-SVM & EEG-LDA in Bespoke, and LDAs for both datatypes in Generalist, per Tables 5.3 & 5.4) were necessarily always best suited as components of a Decision-Level Fusion system. This supporting the design decision that had been made to not take this assumption, and instead optimise for all three choices (EMG classifier, EEG classifier, and Fusion Algorithm) simultaneously. While the potential influence of random effects cannot be disregarded, this distinction suggests the various classifier types as having properties which made them more or less suited to certain late fusion approaches. A rule-based algorithm which combines probability distributions mathematically for example may be better served by an EEG classifier which when incorrect produces flatter, less peaked distributions rather than one which is “confidently wrong’. This could reduce the risk of the incorrect EEG distribution drastically hindering the typically more reliable EMG distribution. Likewise a fusion algorithms whose operation is more akin to a selection between the EMG & EEG models’ outputs could perhaps, as speculated above, be best served by an EEG model which even if itself suboptimal, provided correct classifications localised in parts of the dataset where EMG models were less accurate. This possibility could motivate exploring alternate fusion strategies altogether such as by training an EEG model specifically on the residuals of a unimodal EMG system, though such investigations are left for future work.

5.5.7 Unimodal EEG performance & subject-independence

While the primary driving focus of this work is the multimodal fusion of both EMG & EEG data, as discussed in Chapter 3 the challenge of achieving subject-independence in unimodal EEG classification of motor activity remains somewhat under-researched and distinctly unsolved. The results in Table 5.6 indeed indicate that systems in this work using solely EEG were not able to reach usable classification accuracies. However, the accuracy of approximately 50% across four same-hand gestures is significantly above the chance level — which as noted in 5.5.2 is 29% at the $\alpha = 0.05$ confidence level for these Generalist systems — motivating further exploration and contextualisation here.

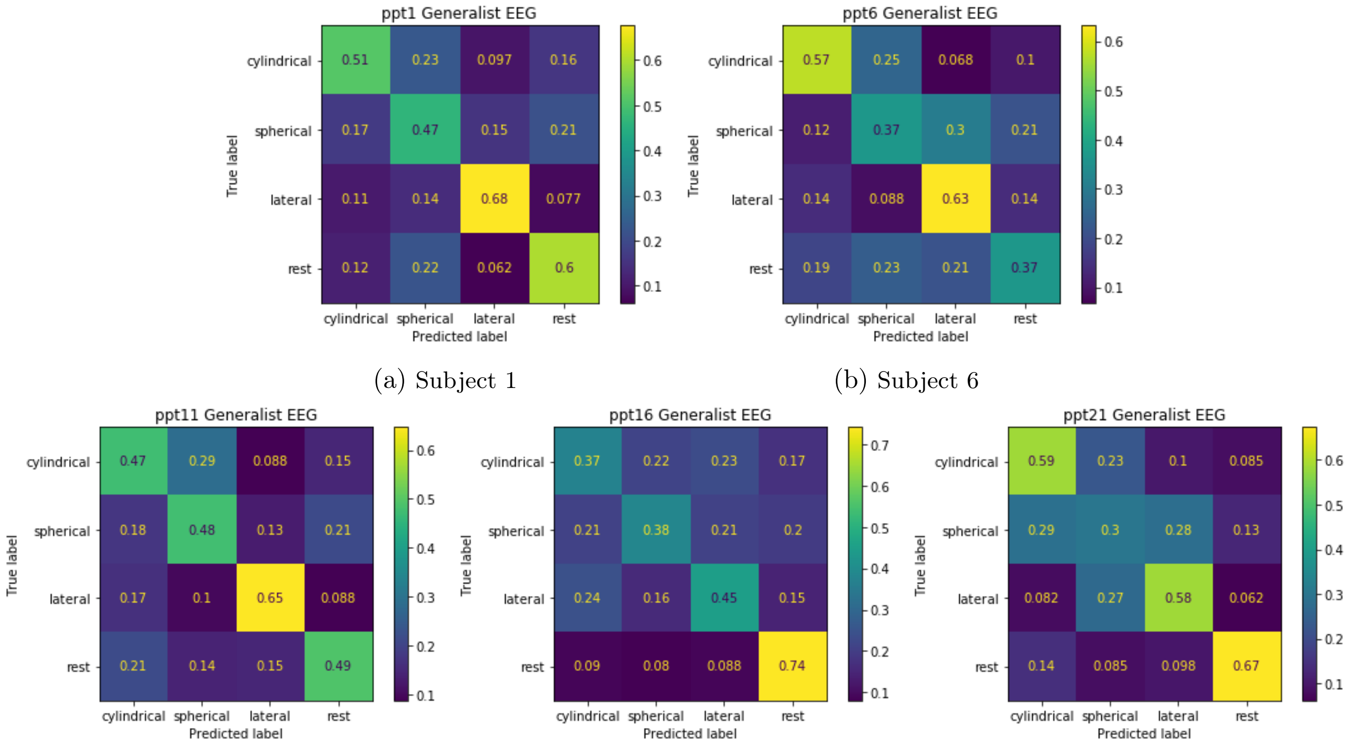
Subject	Generalist EEG Accuracy
1	0.5638
6	0.4850
11	0.5234
16	0.4871
21	0.5367
Mean	0.5192

Table 5.29: Holdout Set accuracy of CASH-optimisation-identified Generalist Unimodal EEG system

As seen in Table 5.29, this mean classification accuracy of approximately. 50% persisted when generalising to unseen held-out subjects. That is, when the optimal Generalist EEG configuration (identified in Table 5.6 as a Least Squares Solver LDA) was as outlined in 5.2.3 trained on the data of all 20 Development Set subjects, this trained model predicted each Holdout Subjects' gestures in turn at a mean accuracy of 52%. Among literature it has been observed that discrimination of specific gestures from EEG is often much more difficult a task than identifying the presence or absence of movement. Such can be inferred from the frequency with which EEG studies seek only to identify movement onset, or to distinguish between highly separable gestures such as movements of different sides of the body [58], and has been noted to impact intuitiveness of EEG-BCIs for users [152]. This trend has been explicitly encountered in studies such as [128], wherein multiclass accuracy was sufficiently low as to motivate simplifying the EEG classification problem to a move-vs-rest paradigm.

An accuracy of approximately 50% across the four gesture classes in this work (Figure 4.1) could *conceivably* be achieved by a system unable to actually discriminate between gestures, but highly accurate in identifying the rest class. To demonstrate with the logical extreme, consider a theoretical system which correctly identified all Rest gestures, and otherwise made a perfectly random guess between the three grasp types. With a balanced dataset such a system would have an expected classification accuracy of $\frac{25}{1} + \frac{25}{3} + \frac{25}{3} + \frac{25}{3} = 50\%$. Inspecting the per-subject confusion matrices of the optimiser-identified Generalist Unimodal EEG system's Holdout validation however, as presented in Figure 5.35, reveals this not to be the case here. Certainly in some subjects the Rest class was most reliably identifiable and the grasps more often confused, this is unsurprising given the greater similarity in hand-shape between the three grasps than between any one of them and the hand at rest. Misclassifications are far from exclusive to the grasp classes however, and even among poorer-performing subjects such as Participant 16 (Figure 5.35d), the attempted between-grasp classification is clearly at least somewhat more accurate than a random guess. The Generalist Unimodal EEG system's accuracy can thus be concluded **not** to be driven solely by its ability to distinguish movement from rest.

Additionally of some note is the frequently superior classwise accuracy of the Lateral grasp, in comparison to those of the Cylindrical and Spherical grasps. This can be plausibly assumed to be due to greater dissimilarity in the handshape of this gesture to those of the other grasps. As described in 4.2 & illustrated in Figure 4.1, the latter two gestures involve abduction of the thumb such that it is in opposition to the palm, while the Lateral grasp instead sets up *side* opposition between the thumb & forefinger with the thumb remaining adducted.



(a) Subject 1

(b) Subject 6

(c) Subject 11

(d) Subject 16

(e) Subject 21

Figure 5.35: Confusion matrices for optimal Generalist Unimodal EEG system in prediction of Holdout data

This subject-independent accuracy is notable in the context of that achieved in prior work. Gordleeva et al. [128] as mentioned ultimately simplified the EEG component of their Decision-Level EMG-EEG fusion to identify only the presence of movement. Their attempted three-class prediction between no movement, movement of the right leg, and movement of the left however achieved a mean accuracy of 51.31% (standard deviation 17.14%) across eight subjects, lower than that reached in a four-class problem here. Ofner et al. [75], while able to distinguish movement from rest at an accuracy of 85±5%, reached an average accuracy of 44±7% over 15 subjects in move-vs-move classification of six gestures of the same limb (elbow flexion/extension, wrist pronation/supination, & hand opening/closing).

Two key dissimilarities between such studies and this work should be recalled. Firstly, that movements of the right & left legs as in the case of Gordleeva et al., and indeed of the elbow, wrist, and hand as in Ofner et al., are more dissimilar than those gestures solely of the right hand used in this study. The right & left legs use wholly different muscle groups from one another — as do the elbow, wrist, and fingers — and accordingly are primarily controlled by geographically separable locations in the motor cortex. Indeed as can be seen in Figure 5 of Ofner et al.'s paper [75], misclassification in their work was more frequent between gestures involving the same muscle groups (e.g. a "hand open" was more likely to be mislabelled as a "hand close" than an "elbow extension"), suggesting performance was partly driven by their system's ability to distinguish the muscle groups being utilised.

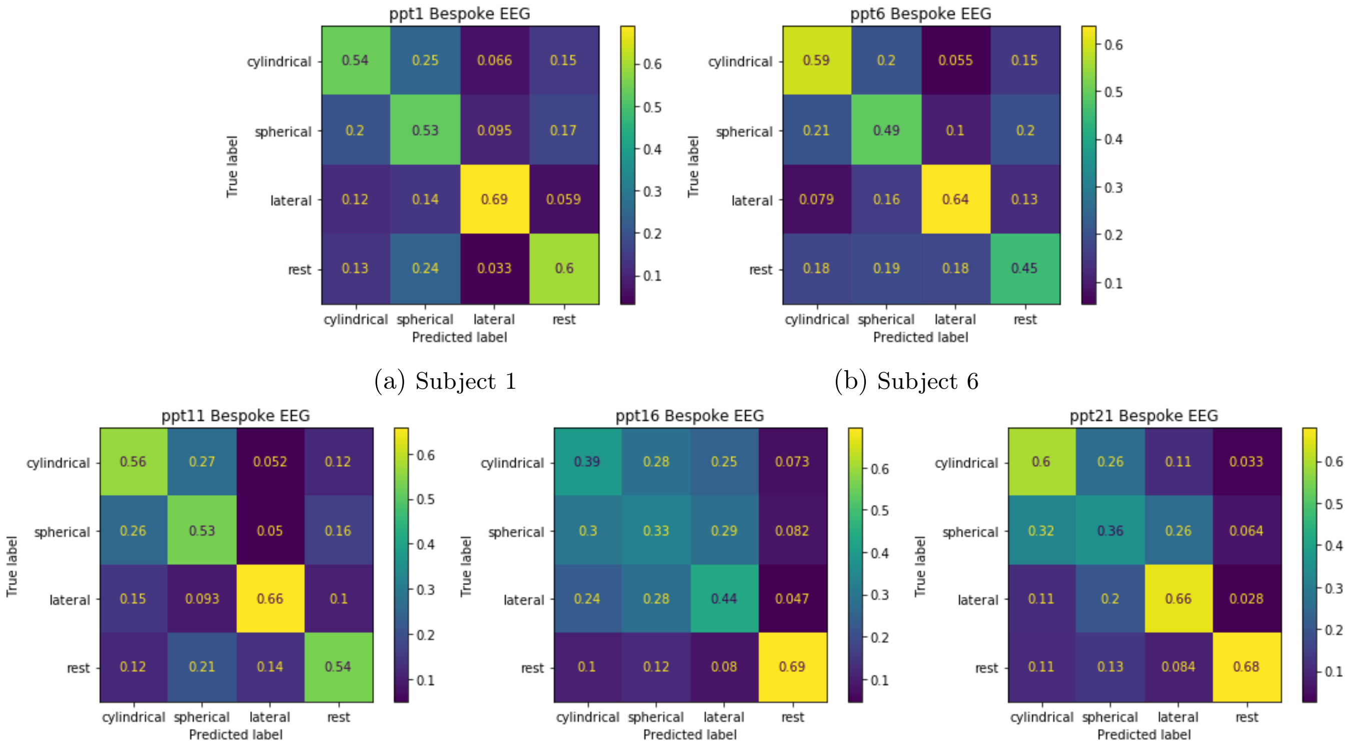
Secondly, the models in the mentioned studies were subject-specific in nature. Here, comparable or even greater accuracy levels were reached with a subject-independent model. This second distinction is perhaps the most crucial. In experiments by Jeong et al. [198] which accompanied their publishing of the dataset used in this work, multiclass accuracies in the order of 40-50% were achieved by LDAs in identifying grasp types from EEG data, but these were again on the basis of subject-specific models. Even the work of Iturrate et al. [162], notable for achieving a mean accuracy of 75.9% across 10 subjects in discriminating between same-hand "power" and "pinch" grasps (loose correlates of the Cylindrical and Lateral grasps in this work) from EEG data — which approaches the accuracies achieved in works using Electrocorticography [167, 168] for similar tasks — did so considering subjects separately on a single-trial [170] basis. Likewise Cho et al. [336] who classified four similar grasp gestures from EEG data at an average accuracy of 68% used subject-specific models. Their strategy is founded on using EMG data to supplement model training (albeit not requiring EMG at the testing stage) and so their results may not be a wholly fair comparison with the Unimodal EEG accuracies obtained here. It should also be noted that Cho et al.'s offline experiments do not appear to be tested on withheld data, limiting their use given the methodological concerns noted in 3.3.2 regarding data leakage; their later online tests while ostensibly more suitable comparators risk bias by being carried out only on those subjects whose offline accuracies were highest.

The work of Fazli et al. [64] was noted in 3.3.1 as one of few EEG studies employing a leave-one-subject-out approach as used by the Generalist system here. While their subject-independent accuracy reached an impressive 73%, exceeding that of this work, it should of course be noted that their model classified between movements of the left and right hands — a problem, as discussed, inherently less difficult than the four-gesture multiclass task of this work.

Subject	Bespoke EEG Accuracy
1	0.5916
6	0.5425
11	0.5681
16	0.4691
21	0.5730
Mean	0.5489

Table 5.30: Holdout performance of CASH-optimisation-identified Bespoke Unimodal EEG system (means of 100 trials)

It is acknowledged that the Bespoke Unimodal EEG accuracies achieved here (Table 5.30), while well above the chance level and outperforming some studies including the aforementioned works of Gordleeva et al., Ofner et al., & Jeong et al., do not reach such heights as some of those found among literature such as Iturrate et al., Cho et al., and others. What should be recalled however is that as outlined in 5.1.2, the Bespoke systems of this chapter while *trained* on a within-subject basis are not themselves *designed* in a subject-specific way. Rather they are intended to be "portable" configurations. By contrast to some studies, wherein model selection & hyperparameter tuning were done on a per-subject basis, or chosen according to their predictive power across all subjects, here the subjects held-out for validation were excluded from *all* stages of modelling. This means a configuration was selected for to its predictive power on the Development Subjects, then provided to novel users (the Holdout Set), to be trained and tested in a subject-specific manner as a true test of the configuration's generalisability. There is no means by which Bespoke models' configurations, in terms of the classification algorithms used or their corresponding hyperparameters, are tailored to the Holdout subjects at any point. This may account for some part of the difference in performance between this work and those studies where EEG system configurations were themselves tailored to be subject-specific.



(a) Subject 1

(b) Subject 6

(c) Subject 11

(d) Subject 16

(e) Subject 21

Figure 5.36: Confusion matrices for optimal Bespoke EEG system's Holdout Set predictions (over 100 trials)

Broadly similar trends in distribution of misclassifications, in terms of the classes more accurately identifiable for each participant, can be seen with Bespoke Unimodal EEG systems in Figure 5.36 as had been seen in the Generalist case above. That a subject's errors were similarly distributed when systems were trained on their own data as when trained on others' may suggest a key limiting factor in Unimodal EEG performance, beyond the suitability of the candidate models & their capacity to learn relevant information from this dataset, as being the dataset itself. This could be for example in the informativity of the feature ensemble used in the work (Table 4.2), or even the preprocessing applied to the data (4.2.5), if discriminative information was encoded in EEG at a frequency which was filtered out.

Given the multiclass accuracies reached by Jeong et al. [198] were of a similar level to those seen here, such a limitation could alternatively be simply in the in the amount of discriminative information carried in the unprocessed EEG data. Limitations in sensor fidelity or experimental protocol, or even subjects' inexperience with BCIs, could affect the richness of the information captured in the dataset. Subject compliance and consistency of behaviour could even be a factor. While Jeong et al.'s use of consistent objects as stimuli across participants will likely have helped to eliciting similar grasps, even properties such as the size of subjects' hands may have influenced the specific hand-shapes they formed when grasping the objects and hence potentially affected inter-subject consistency. Though Jeong et al. do not report any observed issues with participants' adherence to the task, neither do they mention in their paper any specific measures to monitor within-subject consistency of hand shapes & to consequently discard any dissimilar gestures. Of course in principle, a system would be best served by encompassing all possible subtle variations in the way by which a user may perform a gesture. With EEG being a coarser measurement, in comparison to invasive neuroimaging techniques, such slight variations may in fact be minimally distinct. The time-windowing of data outlined in 4.3 may also aid in robustness to any such minor gestural inconsistencies which were time-limited in nature. However, with only 150 performances of each gesture obtained from each participant it is unknown how much unexpected within-gesture variation could be tolerated. Outside the controlled experimental conditions of the lab, individuals cannot be reasonably expected to exactly repeat gestures with consistent precision. There would be value in future biosignal classification research specifically investigating the tolerance of models to minor variations in gesture performances.

5.5.8 Verifying usefulness of the CASH pipeline

Aim 5.3, to "*Establish a pipeline for the unbiased identifying of a performant multimodal system*", motivates assessing whether the CASH optimisation pipeline proposed in this work was a valuable technique for determining system configurations. To enable this, the optimal Bespoke & Generalist fusion systems established by the CASH pipeline are here competed against fusion systems derived from literature precedent, which are detailed fully in 5.3.4. The null hypothesis of such a comparison is that "*A fusion system established by the CASH optimisation pipeline will be no more accurate than one defined on the basis of synthesising biosignal literature*", or formally:

$$H_0 : \mu_{pipeline} - \mu_{literature} \leq 0. \tag{5.4}$$

The is hypothesis is tested in both Bespoke and Generalist contexts. Being similar in nature to hypothesis 5.2 above, a paired one-tailed t-test continues to be applied for statistical analysis here.

Subsequently, this chapter of the thesis is concluded by briefly identifying areas wherein the results presented have indicated that the hyperparameter search space for CASH optimisation can be reduced in complexity by eliminating ill-suited modelling options.

5.5.8.1 Comparing pipeline-identified systems to a "Literature-Informed default"

Table 5.31 presents accuracies achieved for each Holdout Subject by both the Bespoke Fusion system identified in Table 5.3 as the most accurate in optimisation on the Development Set, and the Literature–Informed "default" system defined in 5.3.4 above.

Subject	Bespoke System	
	Literature Default Fusion	Pipeline-derived Hierarchical Fusion
1	0.7551	0.8293
6	0.8072	0.8345
11	0.9280	0.9467
16	0.7996	0.8324
21	0.8071	0.8697
Mean	0.8194	0.8625

Table 5.31: Bespoke performance on Holdout of Literature Default & Pipeline Derived

As in 5.5.1.2, the one-tailed paired t-test's assumptions are virst verified. The Shapiro-Wilk test resulted in a W-statistic of 0.89702 at a p-value of 0.3936; failing to reject its null hypothesis thus indicating normality of paired differences. Comparing variances gave an F-statistic of 0.59767 at a p-value of 0.6303; again the null hypothesis is not rejected indicating the assumption of equality of variances holds true.

The paired one-tailed t-test between the CASH-pipeline-determined Fusion system and that derived from literature reports a t-statistic of 4.0222, at a p-value of 0.007918. The estimated mean difference in accuracy was 0.0431 (lower bound of the 95% confidence interval = 0.0203). The null hypothesis can hence be rejected; the pipeline enabled a subject-specific multimodal fusion system with a significantly higher accuracy than that reachable had the system been designed solely on the basis of inferences from biosignal literature.

Per-Holdout-subject accuracies achieved by the best-performing optimised Generalist fusion system (Table 5.4) and the literature-informed "default" Generalist defined in 5.3.4 are presented in Table 5.32.

Subject	Generalist System	
	Literature Default Fusion	Pipeline-derived Feature-level Fusion (Joint selection)
1	0.6829	0.6633
6	0.7338	0.7475
11	0.8317	0.8233
16	0.6983	0.7246
21	0.7038	0.7121
Mean	0.7301	0.7342

Table 5.32: Generalist performance on Holdout of Literature Default & Pipeline Derived

Reviewing again the one-tailed paired t-test assumptions, normality of paired differences is indicated by the Shapiro-Wilk test's W-statistic of 0.97264 & p-value of 0.8919, and a comparison of variances producing an F-statistic of 0.96258 at a p-value of 0.9714 indicates the assumption of equal variances is also valid.

Here however, the paired one-tailed t-test did not inidicate a significant difference in accuracy between the pipeline-determined subject-independent fusion system and that based on literature insights, with a t-statistic of 0.50315 at a p-value of 0.3207. Though an estimated mean difference of 0.0041 (lower 95% CI: -0.0132) was found, the high p-value indicates this was not statistically significant at the $\alpha = 0.05$ level. While CASH optimisation resulted in a fusion system more accurate than equivalently-optimised Unimodal EMG for subject-independent classification, it was no better than a fusion system with characteristics inferred from literature. Considering the equivalent accuracies of different Fusion architectures for Generalist systems discussed in 5.5.2 above this may indicate the most suitable design for a subject-independent EMG/EEG fusion remains to be found and lies outside the modelling space defined in this work. Alternatively, it may indeed be that despite there being very little literature precedent upon which to draw, the subject-independent system configuration defined in 5.3.4 was indeed particularly suitable. Modelling choices being synthesised from their performances on subjects outside the dataset used in this work could plausibly lead to their being well-suited for cross-subject classification. In Bespoke systems, which are advantaged by their ability to consider specific characteristics of subjects' data in optimisation, such synthesis may have come at the expense of this tailoring — though it should be recalled that the Bespoke systems of this chapter are designed for "portability" in their configurations, and their optimisation is done with no sight of Holdout data.

5.5.8.2 Learning from these results for a narrowed modelling space

Following from the findings outlined in 5.5.4 and 5.5.6, certain axes of the hyperparameter search space previously presented in Figures 5.6 and 5.7 can be adjusted or removed for experiments in the work's subsequent chapters. This is done in a subtractive way. Rather than allowing only the highest-performing options to persist, here only those cases where a hyperparameter value appeared to consistently lead to significantly worse accuracy are removed. Where competing options were equivalent or reasonably competitive they are not ruled out here, to allow for the possibility that alternative configurations may be optimal for the subtly different problems explored by the following chapters.

To enable further chapters' experimentation to be similarly verified with the unseen Holdout Set, as was done here in 5.5.1.2, these decisions are made on the basis of models' performance over the Development Subjects — as has been the case throughout Section 5.5 other than where results were explicitly noted as relating to the Holdout Set. This ensures the isolation of the Holdout data is preserved and it is not used to influence the modelling decisions. Had subsequent experiments been informed by knowledge of hyperparameters favoured by the Holdout Set, this would render such separation invalid [190] and be a significant source of data leakage [191]. As discussed throughout the thesis, this work takes particular effort to avoid the data leakage issues which Hosseini, Powell, et al. [31] among others note to be common among much biosignal research.

The resultant reduced search spaces can be seen in Figures 5.37 and 5.38 for Decision-Level Fusion Algorithms and for component EMG & EEG Classifiers respectively.

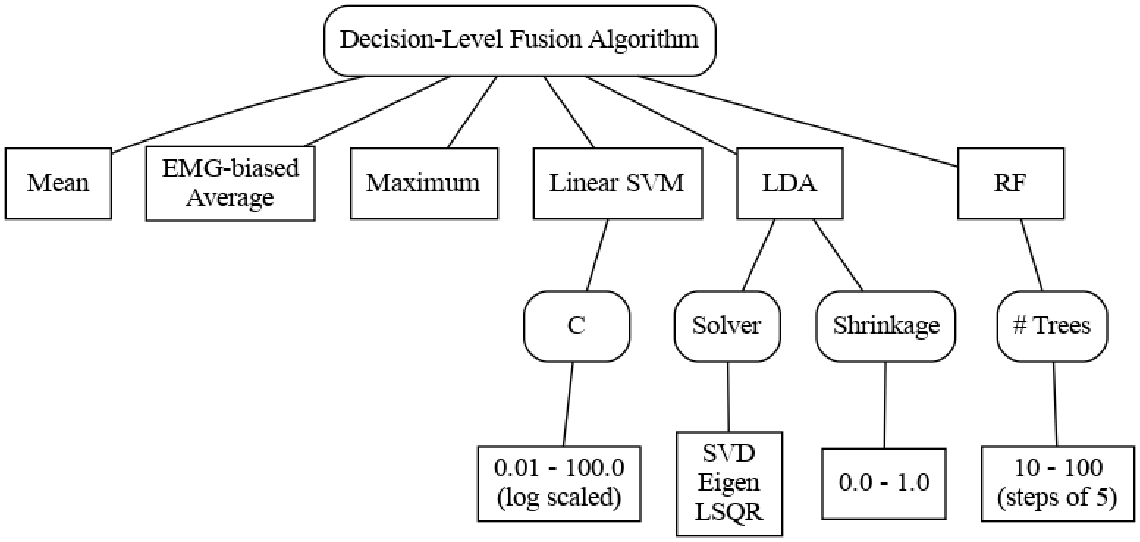


Figure 5.37: Subsection of the hyperparameter search space describing the algorithm used for Decision-Level Fusion, reduced from that of Figure 5.7 following findings described in 5.5.6.

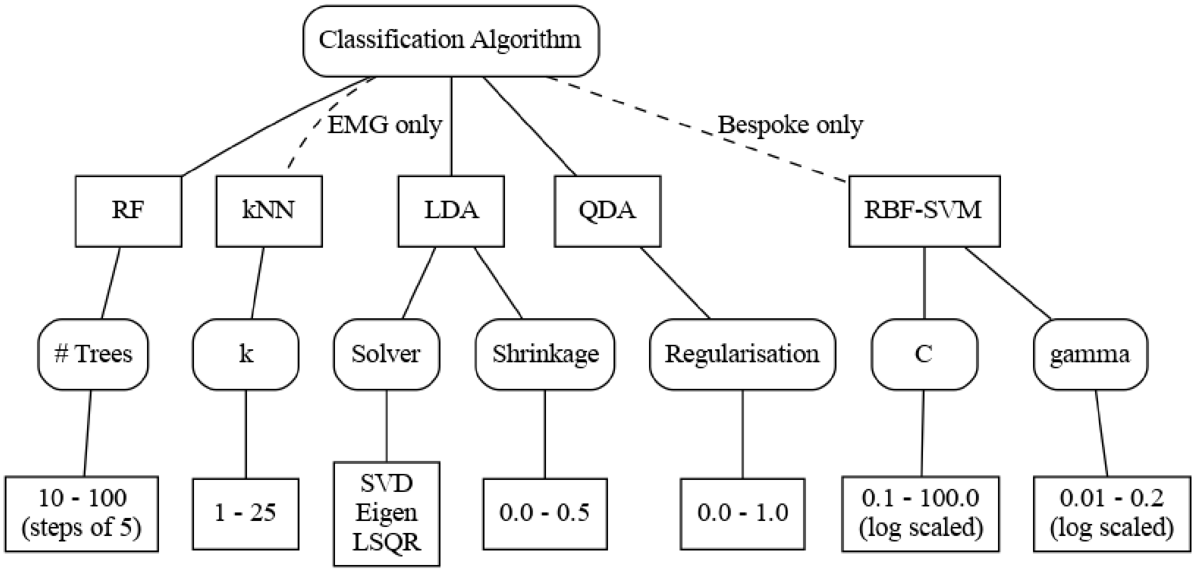


Figure 5.38: Subsection of the hyperparameter search space describing the nature of EMG and EEG classification models used & their conditional hyperparameters, reduced from that of Figure 5.6 following findings in 5.5.4.

5.5.9 Conclusions

The results of these experiments clearly demonstrate the potential of the early, late, and novel Hierarchical approaches to multimodal EMG-EEG fusion presented in 5.3.1 for both subject-specific ("Bespoke") and subject-independent ("Generalist") multiclass gesture classification.

The proposed Combined Algorithm Selection & Hyperparameter optimisation pipeline, novel to the domain, proved capable of identifying suitable model configurations for such systems, in the Bespoke case producing a more performant system — implementing a novel "Hierarchical" fusion architecture — than could be reasonably inferred from literature precedent as demonstrated in 5.5.8. The fusion of EMG & EEG was demonstrated to provide more accurate Generalist classification than the best-performing single-data-modality system given an equivalent optimisation budget. In a number of its findings, such as the suitability of the Linear Discriminant Analysis model for EEG-based gesture classification discussed in 5.5.4.1, this research provides a sound evidential basis to corroborate trends which are prominent but not always evidenced among the Brain-Computer Interface literature.

Given its successful use in determining "portable" system designs here — configurations suitable for bespoke modelling with novel users — a strategy based on this work's CASH approach could have a transformative impact on the process of deploying gesture-recognition based prostheses in the real world. It may be viable, for instance, to extend the method to allow some degree of automated personalisation. An offline CASH optimisation over a large multiple-subject pre-existing dataset could provide an initial baseline system which was then further customised according to an individual user's data. Computation could be offloaded from the prosthesis itself to a cloud — or a paired mobile device if privacy concerns motivated data being kept local — which would pick up the optimisation routine to continually fine-tune models' hyperparameters with longitudinal biosignal data collected over the device's use.

The use of CASH optimisation also enabled a fair comparison between the fusion architectures, with modelling choices configured through a systematic, transparent, unbiased evaluation of their potential. This places the work in stark contrast to much of the literature on BCIs, noted by Lotte et al. to frequently be weak in the rigour of its tests and to present unqualified modelling decisions which cannot be confidently said to be free from bias [30]. The care taken to validate results on wholly unseen data, rather than risk reporting classification accuracies artificially inflated by over-optimisation, also sets this work apart from many contemporaries — as discussed in Chapter 3 such issues, and problems of data leakage more broadly, have been found by Hosseini, Powell, et al. [31], Li et al. [192], and others to be prevalent among biosignal studies.

This research also goes beyond many prior investigations into multimodal EMG-EEG fusion in its application of fusion techniques to a more complex task than typically attempted before. Where others have sought to classify between movements of different limbs [112, 139, 146], different joints of the same limb [130, 137], or movements of different intensities [141–143, 145], here fusion was successfully used in a multiclass problem consisting of three similar right-hand grasps and a rest class — which has been seen only before in a few works such as [113] and does not appear to have been previously done on a subject-independent basis. This subject-independence was also notable in the Unimodal EEG classification performed here. Section 5.6 discussed that the mean Generalist accuracy achieved, while not yet at a usable level for real-world deployment, was competitive with many studies' attempts to classify similar problems on a subject-specific basis. This indicates promising steps towards a reduction in the subject-dependence of gesture classification systems, and the potential benefits in terms of cost & convenience to users which would follow.

Leveraging Cross-Subject Learning in Gesture Classification

6.1 Aims & Overview

Chapter 5 investigated two opposed ways by which data from a given user of a system (hereafter "Same-Subject data") and data from other individuals (hereafter "Other-Subject data") could be utilised, and found that a well-trained Bespoke system could not be outclassed by a Generalist allocated equivalent optimisation budget. This might initially be taken to indicate a wholly subject-specific approach as the universally more viable option for designing a biosignal-based gesture classification system. However, those presented in Chapter 5 are not the extent of possible approaches for combining these data. The motivation for further exploration of cross-subject generalisation of biosignal classification is strong — any mechanisms which may be able to reduce the need for subject-specific training data would naturally lessen the burden on an end user of a system. Making a system more convenient to use in this way is both a meritworthy goal in its own right, and could also have a potential beneficial impact on the rejection rate of devices such as robotic prostheses. It may also be possible for cross-subject transfer learning to improve a gesture classification system's accuracy beyond that which could be achieved with a given amount of subject-specific data.

This chapter essentially seeks to explore in further depth that potential benefit of leveraging data belonging to "other" subjects in the classification of a given individual's data, and to identify suitable strategies for doing so. The Aims being investigated in the chapter are hence as follows:

- **Aim 6.1** *Can inclusion of other-subject data boost performance above that achievable with same-subject data alone?*
- **Aim 6.2** *Can inclusion of other-subject data allow the same level of performance to be obtained with a reduced amount of same-subject data?*

While the majority of studies in the biosignal literature focus solely on within-subject classification (see 3.3.1), there have been a number of strategies explored for attempting to reduce subject-dependence. This chapter explores two possible ways by which data from a given subject and data from other individuals can be integrated:

- Dataset Augmentation — a subject's data is pooled with data from other individuals prior to learning, and models fit to that merged dataset
- Model Transfer — a subject's data is used to adapt models previously fit to data from other individuals

Previously in Chapter 5 the Generalist system was defined as being fully subject-independent, having no sight of a target user's data. The Bespoke system meanwhile was defined in a "portable" way: its modelling decisions were made on the basis of multiple subjects' data through the Combined Algorithm Selection & Hyperparameter Optimisation process, but the resultant system was trained solely on data belonging to each given subject-under-test in turn. In this chapter by contrast a "fully" bespoke system, in which modelling decisions are made through CASH optimisation performed solely on the basis of each individual subjects data, is considered as the baseline from which other-subject data may be incorporated. While increasing the computational load of the experiments, as per-subject optimisation is required, this may have the potential to enable systems to be more specialised to each subject-under-test.This increased computational expense further motivates exploration of the extent to which one could reduce the level of same-subject data required by a system to achieve an equivalent classification performance. Not only would this reduce the data collection burden on an end user of a theoretical deployed system, but the resultant decrease in size of the dataset would lead to faster convergence in the optimisation process. In addition to trialling various levels of subject-specific data, it is valuable to assess the impact of integrating varying amounts of supplementary other-subject data. Were a system deployed in the "real world" to rely upon the collection of data in advance from a number of individuals other than the end user, it would be beneficial to know how much data is required for this purpose to avoid unnecessary expense in terms of the time and cost required to recruit said individuals and collect & process their biosignal data.

6.2 Methodology

Deriving from the Aims outlined above, the purpose of Chapter 6's experiments are to assess whether:

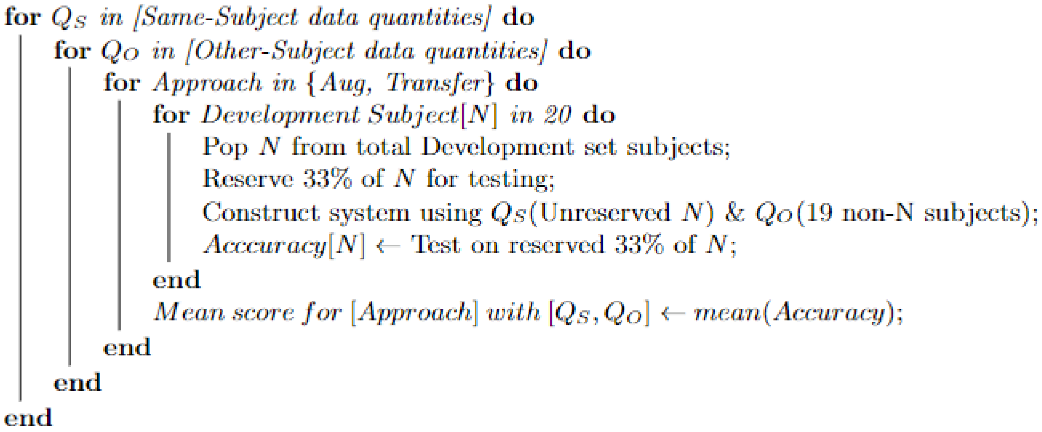
- Incorporating other-subject data improves a system's classification accuracy for a subject [Aim 6.1].
- Incorporating other-subject data allows a system with access to less same-subject data to achieve the same classification accuracy as one with more (i.e. whether it can enable a reduction in the data collection requirements for a subject) [Aim 6.2].
- Dataset Augmentation or Model Transfer appears a more suitable approach for such incorporation of other-subject data.

There are essentially three independent variables which hence arise: the level of same-subject data provided to a system, the level of other-subject data incorporated to it, and the approach used for combining those data. To explore their impact, the predictive power of systems with access to a range of quantities of Same-subject and Other-subject data are investigated for each of the two combinatorial approaches. As with Chapter 5's experiments, this initial exploration is conducted with five subjects (Subjects 1, 6, 11, 16, and

21 as outlined in 4.2.3 above) being again held out, to be used for validating the observations made from experimentation & the extent to which the findings generalise.

For given quantities Q_S and Q_O of Same- and Other- Subject data,with each of the Augmentation and Model Transfer approaches a subject-specific system is created and tested in turn for each subject N among the the 20 Development Set subjects. Considering Subject N as the "subject-under-test", 33% of N's data is reserved for testing the system. A portion Q_S of the remaining 67% unreserved same-subject data, and a portion Q_O of the data from the other 19 subjects, is then used for modelling. The resulting system is used to predict the reserved 33% of N's data and the accuracy of those predictions evaluated; the process is subsequently repeated with subject $N + 1$ *et cetera* in the same way. The mean of these subject-specific accuracies is computed to quantify the predictive ability of that given combinatorial approach at that particular combination of same- and other- subject data levels. This procedure is outlined in Algorithm 1. Systems' classification performances are then analysed to evaluate the extent of the impact of incorporating other-subject data to systems with access to varying degrees of subject-specific data, and to compare the performance of the two different approaches in doing so.

Algorithm 1 Procedure for exploration with the Development Set



```
for $Q_S$ in [Same-Subject data quantities] do
    for $Q_O$ in [Other-Subject data quantities] do
        for Approach in {Aug, Transfer} do
            for Development Subject[N] in 20 do
                Pop $N$ from total Development set subjects;
                Reserve 33% of $N$ for testing;
                Construct system using $Q_S$(Unreserved $N$) & $Q_O$(19 non-N subjects);
                $Acccuracy[N]$ ← Test on reserved 33% of $N$;
            end
            $Mean\ score\ for$ [Approach] with $[Q_S, Q_O]$ ← mean(Accuracy);
        end
    end
end
```

The Holdout Set is then used to verify the specific observations arising from this exploration, as outlined in Algorithm 2. Here each Holdout subject is considered in turn individually, with all 20 of the Development subjects treated as the "other-subject"s from whom data is available to supplement a Holdout subjects' system. That is, no data is shared between Holdout subjects; they are assessed in isolation and have an identical pool of other-subject data from which to draw, to enable the fair assessment. The mean subject-specific classification accuracy over the five Holdout subjects is calculated for each system under test.

Algorithm 2 Procedure for verification with the Holdout Set

```
for $each\ specific\ system\ being\ evaluated$ do
    for $Holdout\ Subject[N]\ in\ 5$ do
        Reserve 33% of $N$ for testing;
        Construct system using $Q_S$(Unreserved $N$) & $Q_O$(20 Development subjects);
        $Acccuracy[N] \leftarrow$ Test on reserved 33% of $N$;
    end
    $Mean\ score\ for\ system \leftarrow mean(Accuracy)$;
end
```

6.3 System Design

6.3.1 Overview

Figure 6.1 illustrates the generic structure of an Other-Subject-Supplemented system. In every subject-specific system a random 33% of the Subject's data is initially reserved for testing as previously noted, ensuring the final evaluation is an assessment of the system's performance on unseen data from the subject. This split is stratified by the type of gesture being performed, to ensure a balanced of classes in the resultant dataset. Data is split on the basis of whole-gesture performances to minimise data leakage, by ensuring time-adjacent samples are not distributed between training and testing sets as explained in 5.2.3.

6.3.1.1 Data Combination

The remaining unreserved 67% of the Same-Subject data is scaled according to Q_S, and combined with data from the other 19 Development Subjects scaled according to Q_O; both of these downsamplings are stratified by class. Dagois et al.'s work on Motor Imagery classification was able to estimate similarity between subjects' data, and use this to select specific "best match" subjects with whom to augment a given user's dataset [337]. The possible benefit however of such an approach over alternative selection methods, or a random selection of augmentation data, is not assessed. Additionally while Dagois et al.'s study is multimodal in nature, the modalities it makes use of alongside EEG is functional Transcranial Doppler ultrasonography (fTCD), a measurement of blood flow in the brain via ultrasound [338]; both data sources capture neural activity, by contrast to the use of muscular and neural signals in this work. Other work utilising similar approaches such as that of Azab et al. [184] wherein other-subjects' influence is weighted according to a metric of similarity between their Common-Spatial-Pattern-filtered EEG data based on Kullback-Leibler divergence [339], and Lotte & Guan [182] wherein the class covariance matrices used within the CSP & LDA algorithms incorporated data from a subset of additional subjects chosen on the basis of uncalibrated cross-subject classification performance, similarly make use only of neurological data. It is unlikely to be a valid assumption that similarity in neural data between certain subjects would necessarily indicate a similarity in their muscle data and vice-versa, thus to incorporate such a "screening" technique to this study would require EMG & EEG to be handled separately, and potentially each supplemented by different subjects. This could not only add significant additional computation but could pose various challenges to the functioning of the

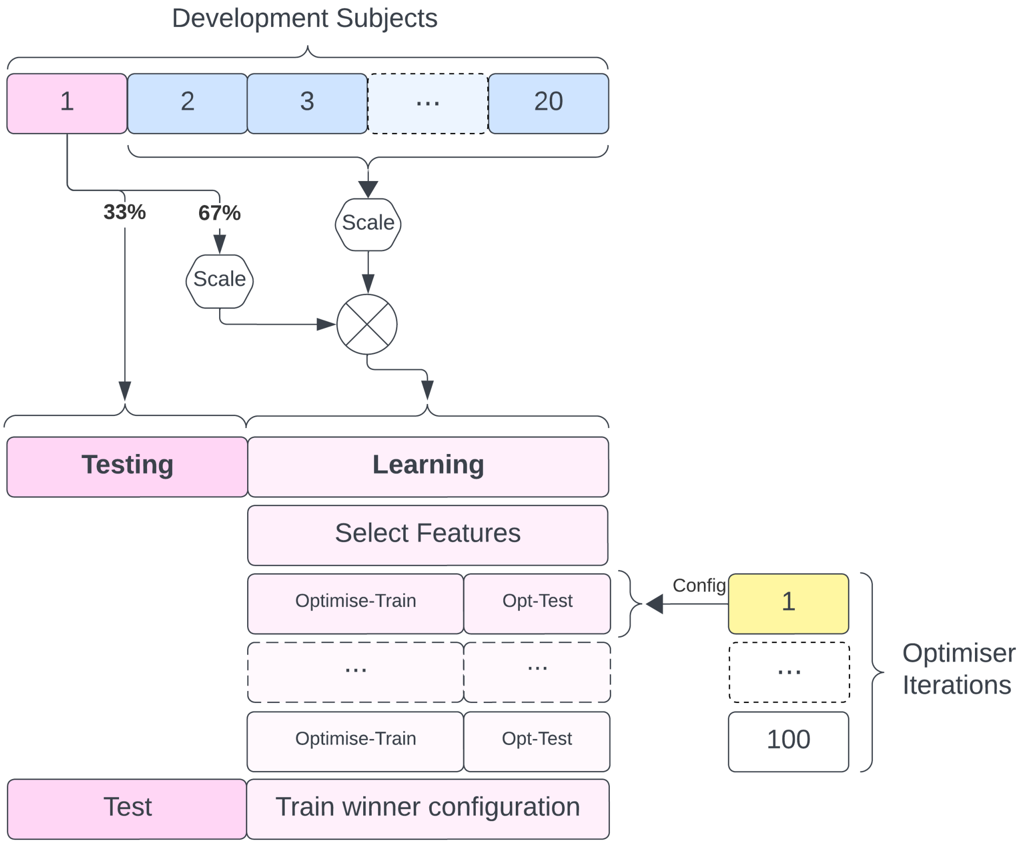


Figure 6.1: Illustration of data flow for a generic cross-subject learning system, of a given Development Set subject at a given combination of Same-Subject and Other-Subject data levels.

systems, most notably impeding meta-model based fusion algorithms which require training on synchronous EMG- & EEG- predicted class probability distributions (as described in 5.3.1.3).

It should additionally be noted that with a sample of just 20 individuals in the Development set, there is likely to be significant variation in the similarity between subjects' data and the data selected to augment them with; some subjects may have closer (and hence more beneficial) "best matches" than others. This factor could be difficult to assess and control for, so would present challenges in aggregating subjects' results for evaluation. Thus while future work may indeed take interest in exploring the merits of doing so, "screening" of the other-subject data was not considered to be of sufficient likely benefit here to justify the added system complexity which it would necessitate. The downsampling of Other-Subject data was therefore instead *stratified* by subject, such that the resulting Other-Subject dataset used to supplement a system held equal contributions from all 19 (or 20, in the case of Holdout verification) subjects not-under-test. This minimises any risk of bias arising from variations in the degree of similarity between subjects' biosignal data.

6.3.1.2 Feature Scaling & Selection

For practical implementation reasons, features were scaled according to the Same-Subject data in all cases. The Same-Subject data Q_S present after downsampling were standardised to have a mean of zero and standard deviation of 1, and the same scaling transformation applied to any Other-Subject data Q_O present which supplemented it. This transformation was also later applied to the reserved 33% Same-Subject testing split, ensuring the test set was scaled according to data which can be assumed likely to be similar to it, thus potentially reducing the risk noted in 5.3.2.1 of unseen data being scaled to novel values outside a model's expected range.

Informative features are selected from the merged Same-and-Other-Subject dataset using the methods described in 5.3.2.2 above, & the reserved Same-Subject test data reduced to that same feature ensemble. Here EEG data were reduced to 40 features as in a Bespoke system (see 5.3.2.2) — this figure was chosen for consistency with experiments in Chapter 5 and to ensure consistency across the experiments in this chapter rather than introduce a potential confounding variable, but otherwise arbitrary[1].

6.3.1.3 Optimisation & Modelling

The resulting combination of same- and other- subject data can then be used for learning, of which as in Chapter 5 there are essentially four stages: Feature Selection, Model Selection, Hyperparameter Optimisation, and Model Training. As in Chapter 5, Model Selection & Hyperparameter Optimisation are performed in parallel through a Combined Algorithm Selection and Hyperparameter optimisation (CASH) routine to identify a suitable system configuration.

The work in Chapter 5 found that in Bespoke systems, the Decision-Level and Hierarchical fusion algorithms proved significantly stronger than other approaches (see 5.5.2), albeit not able to exceed a Unimodal EMG approach. By contrast in Generalist systems these two algorithms were no better or worse than their competitors, but *did* outperform the Unimodal EMG system (5.5.1). While systems in this chapter draw on both same- and other- subject data and are hence neither strictly Bespoke nor Generalist in the same manner as those in Chapter 5, it can be inferred from these results that Decision-level and Hierarchical fusion algorithms are promising candidates. For simplicity, Decision-Level Fusion is adopted as the system architecture from hereon, though future work may find merit in exploring the problem of cross-subject generalisation with Hierarchical fusion systems.

It should be acknowledged explicitly here that this selection of Fusion Architecture is carried out on the basis of their performance on the Holdout Set. Section 5.5.8.2 noted the potential for such an evaluation to be problematic and indeed to risk jeapordising the isolation of the Holdout data if knowledge of its properties is used to inform experimental parameters & focus on areas of interest which ought not to be known [190]. What is distinct here however is that the choice of Fusion Architecture is no longer a part of the problem to be solved; it is neither a dependent variable being investigated in these experiments nor is it explored algorithmically (e.g. within the CASH optimisation), it is simply made static. This foreknowledge of the

[1] NB that even at the highest level of Other-Subject data investigated (see 6.3.4.2), merged datasets in these experiments were significantly smaller than those of the Generalist models which in Chapter 5 necessitated a wider feature array.

suitable Fusion Architecture thus does not risk undermining the separate investigation here into transfer learning mechanisms.

Section 5.5.8.2 also used the findings of Chapter 5 to allow a narrowing of the search space for CASH optimisation, ruling out less promising candidates to enable the optimisation process to converge its exploration more quickly and hence have greater potential capacity for exploitation. While the optimal model & hyperparamer configuration for a system is not the investigatory focus of this chapter — rather a means of enabling a fair comparison of systems, by affording each different subject-specific system an equivalent optimisation budget — it is nevertheless a significant part of the procedure. A hyperparameter space tailored to the Holdout set could perhaps be assumed to give all systems trialled in this chapter an equal artificial "boost" but would nonetheless boost them, and would certainly impede the extent to which Holdout performance could imply generalisation of the findings beyond this study. For this reason, as highlighted in 5.5.8.2, these evaluations were specifically on the basis of the Development Set performance explored in 5.5.4 & 5.5.6. Figure 5.37 presents the updated hyperparameter space for Decision-Fusion Algorithms and Figure 5.38 that for the component EMG & EEG classifiers. Not all of these candidate models are capable of being adapted post-training; some are hence incompatible with the Direct Model Transfer approach for merging same- and other- subject data. Subsection 6.3.3 below discusses this further and presents in Figure 6.4 the alternative hyperparameter space over which Model Transfer systems optimise.

After the removal of the 33% reserved for testing, the maximum amount of same-subject data available to a system to learn from is 67% (400 whole gestures) before any downsampling. There is meanwhile a maximum of 11400 other-subject gestures which could be made available to a Development Set subject's system (600 from each of the 19 remaining Development subjects) — which rises to 12000 when testing with a Holdout subject (which as noted above can draw from all 20 Development subjects). The levels of available same-subject and other-subject data evidently differ greatly and the same-subject data risks being dominated by other-subject data in a merged dataset. This dominance will be exacerbated by the downsampling of data as the ability of a systems to perform well with less subject-specific data is investigated. To ensure the same-subject data is used effectively even when only present in very low quantities, or only representing a small proportion of the merged dataset, in all cases the optimisation target of a CASH optimisation procedure is defined by its predictive power over same-subject data only. For a given CASH optimisation routine, in each of the optimiser's 100 iterations a random 33% of the Same-Subject data within the merged "Learning" dataset is held as the "optimise-test" split of that iteration. The rest of the merged dataset is available for training the candidate model which was configured according to the iteration's selected point in the hyperparameter search space. The manner in which this training takes place corresponds to the Approach used to learn from the two data sources, as outlined below in 6.3.2 & 6.3.3). The optimiser's loss function was the error rate of the candidate's predictions of the aforementioned "optimise-test" split.

Of those 100 optimisation iterations, the hyperparameter configuration which maximised Same-Subject classification accuracy in this way is taken as the winner. The winner system is then re-trained on the entirety of the "Learning" dataset in the manner prescribed in 6.3.2 or 6.3.3, and this trained model used to predict the gestures of the unseen 33% of Same-Subject data which was previously reserved. The accuracy of its predictions is evaluated and recorded as the accuracy of the given Approach at the particular given combination of Same- and Other-Subject data levels for the subject-under-test. The mean accuracy across each Development subject's subject-specific model for a given Approach at given data levels is taken to determine the overall performance of that Approach at that data level, as per Algorithm 1.

Other practical considerations, such as the training procedure for meta-model-based Decision Fusion algorithms & the coercion of probability estimates from Support Vector Machines, are carried out as they were described in Chapter 5.

6.3.2 Augmentation Approach

Under the approach defined here as Dataset Augmentation, data from the subject-under-test ("same-subject" data) and data from the remaining subjects ("other-subject" data) are merged together to form a new dataset before any learning takes place. This can be seen as a form of transductive transfer learning: the system has access to both the "source" (other-subject) and "target" (same-subject) data from which to model characteristics.

Dagois et al.'s aforementioned work [337] adopts augmentation as a form of transfer learning in the classification of Kinaesthetic Motor Imagery from EEG data, selecting five non-target subjects with whose data to supplement that of the subject-under-test based on assessing the similarity of their feature distributions, and subsequently classifying with QDAs, LDAs, and SVMs. Interestingly they find this augmentation to enable a potential decrease in the required amount of subject-specific data to achieve the same level of classification accuracy, strongly motivating the exploration of this approach in this work given the stated Aim 6.2. Azab et al. also quantified similarity in featurespaces between subjects, in their case with regard to the Common Spatial Pattern filters derived from subjects. Rather than use this to select subjects for inclusion however it instead determined the weighting assigned to each when modelled together with a Logistic Regression classifier. Their study found that where the degree of subject data present in a system was "sufficient" then a subject-specific approach was preferred (being variously either stronger or achieving equivalent accuracy with lesser complexity than the augmented systems), but when this was not the case the augmentation did indeed improve a weaker classifier's accuracy. Kobylarz et al. [230] used an augmentation approach in the context of classifying three hand gestures from EMG data with a Random Forest. In an inversion of the framing of those works previously discussed, they initially developed a subject-independent model trained on data from multiple individuals, and upon finding it to perform poorly with both novel users and repeat users during subsequent sessions, sought to address this by incorporating some calibration data. They found that including an additional 5 seconds of data per gesture class from the target subject & retraining a model on this augmented dataset could improve classification accuracies to usable levels. The extent to which these augmented models actually benefited from the presence of the initial other-subject data was not explored however. Benalcázar et al. [173] meanwhile found that a wholly subject-independent EMG classifier using a Leave-one-subject-out approach, akin to that of the Generalist presented in Chapter 5, was not able to achieve usable classification accuracies. An augmentated system however, in which an equal quantity of data from each subject was used to train the model, offered vastly improved performance.

In the optimising of Augmentation systems in this work, for each iteration, after one third of the same-subject data within the combined "Learning" dataset was held as the "optimise-test" set, one third of the other-subject data was temporarily discarded. The candidate model was being assessed was trained on the remaining two-thirds of the same-subject data along with two-thirds of the other-subject data. The proportion of same-subject to other-subject data within a candidate model during optimisation was hence the same as the proportion within the full "Learning" dataset later used to train the winner model, thus ensuring a closer match between the problem being optimised for and the problem to which the optimisation result was applied. Without such a consideration the dominance of other-subject data over same-subject in a candidate model would be exaggerated, potentially motivating the optimiser to find a configuration which paid undue attention to other-subject data. Figure 6.2 presents an Augmentation-specific derivation of the generic system overview seen in Figure 6.1 above, illustrating the nature of the data splitting during CASH optimisation.

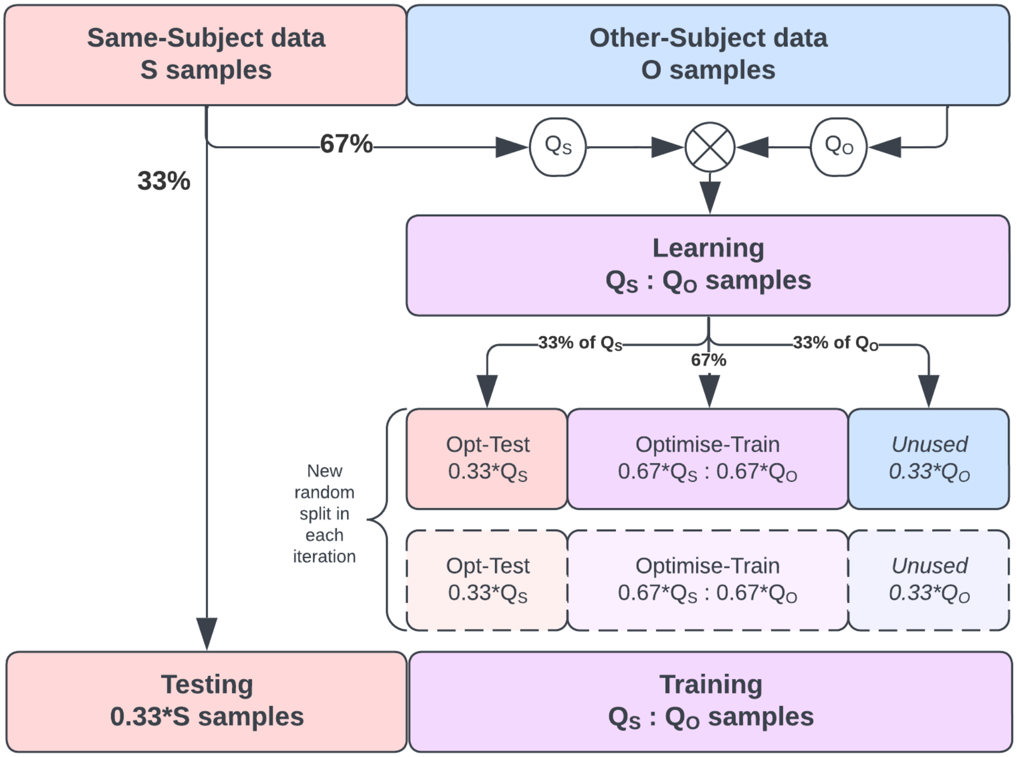


Figure 6.2: Illustration of the splitting of data within the optimisation of an Augmentation system

6.3.3 Model Transfer Approach

The approach defined here as Model Transfer makes use of inductive transfer learning principles: a system is trained initially on subjects other than the one under test which are treated as the source domain, and the trained model subsequently adapted the hitherto unseen domain of the subject-under-test. The system can thus potentially both reap the benefit of accessing a larger, more diverse source dataset while also being fine-tuned to specialise on the target subject.

While transfer learning of this more direct kind has been approached in various ways among the biosignal literature such works typically focus on enabling the transfer of information during the modelling of a dataset's feature encoding, rather than at the stage of model training itself. Kang et al. [180] classified binary Kinaesthetic Motor Imagery problems from the gold-standard "BCI Competition" EEG datasets using LDA models following Common Spatial Pattern feature extraction. These are both well-established domain techniques but Kang et al.'s work addressed the challenges presented by the highly session-specific nature of CSP filtering [152] (as discussed in Chapter 4) by a transfer learning scheme. CSP filters for each subject were found simultaneously on the basis of modelling subjects' data as sharing a latent space, thus allowing for CSP projections which were somewhat aligned across subjects — extending their prior work's attempting of a similar concept through clustering subjects' CSP spatial patterns, which facilitated information transfer within clusters (ie between subjects with similar spatial patterns) [181] but did not enable group-wide characteristics to be captured [180]. Guar et al. [178] in a similar vein used other-subjects' EEG data to directly influence a system's modelling of a subject's CSP through derivation of features from the data's tangent space for subsequent use with both linear [178] and logistic regression classifiers [179]. Such featurespace-focused transfer learning is not constrained to modelling of CSPs however. Min et al. [340] took a quite unique approach in motivating a Long-Short-Term-Memory (LSTM) Neural Network to explicitly model characteristics which did not directly provide class-relevant information but helped in predicting to which individual from among the dataset a datapoint belonged. Aspects of a model trained to identify the user were used to modify a gesture-predicting model, encouraging it to pay attention to such characteristics when classifying new data. Joadder et al. also eschewed the use of CSPs and instead explored the suitability of a range of statistical features to subject-independent KMI classification with a leave-one-subject-out method [183], essentially evaluating the extent to which the distributions of such features in the source and target domains were equivalent [341]. Gonzales-Huisa et al.'s recent work [185] is one example of similar techniques being applied in the classification of EMG data. They made use of correlational alignment (CORAL) to transform subjects' data such that their feature distributions matched, enabling SVM models to be applied across both source and target domains, and used the direct model adaptation approach, akin to that which is applied in this work, with their LSTM classifiers. Ketykó et al. [171] similarly domain-shifted their EMG data, applying transformations to account for the variation between subjects & subsequently classifying the data via a 2-stage Recurrent Neural Network. 50% of a target-subject's data was used for this domain adaptation, and they also investigated the impact of using that 50% instead a pre-trained RNN, though the most suitable modelling choices for the latter technique were underexplored.

By contrast to the Augmentation systems outlined previously, the training procedure for a Model Transfer system in this work intrinsically incorporates a stage of specialisation to the subject-under-test's data; same- and other- subject data are not pooled at the point of model training. The risk of an imbalance between same-subject and other-subject data causing a system to pay insufficient attention to the same-subject data is thus much reduced. Hence, during the optimisation of Model Transfer systems it was not necessary to discard a third of the other-subject data; candidate systems were trained initially on all the other-subject data available in the merged "Learning" dataset, and adapted using the two-thirds of the same-subject data which were not held for testing that iteration's candidate system configuration. This procedure can be seen in Figure 6.3.

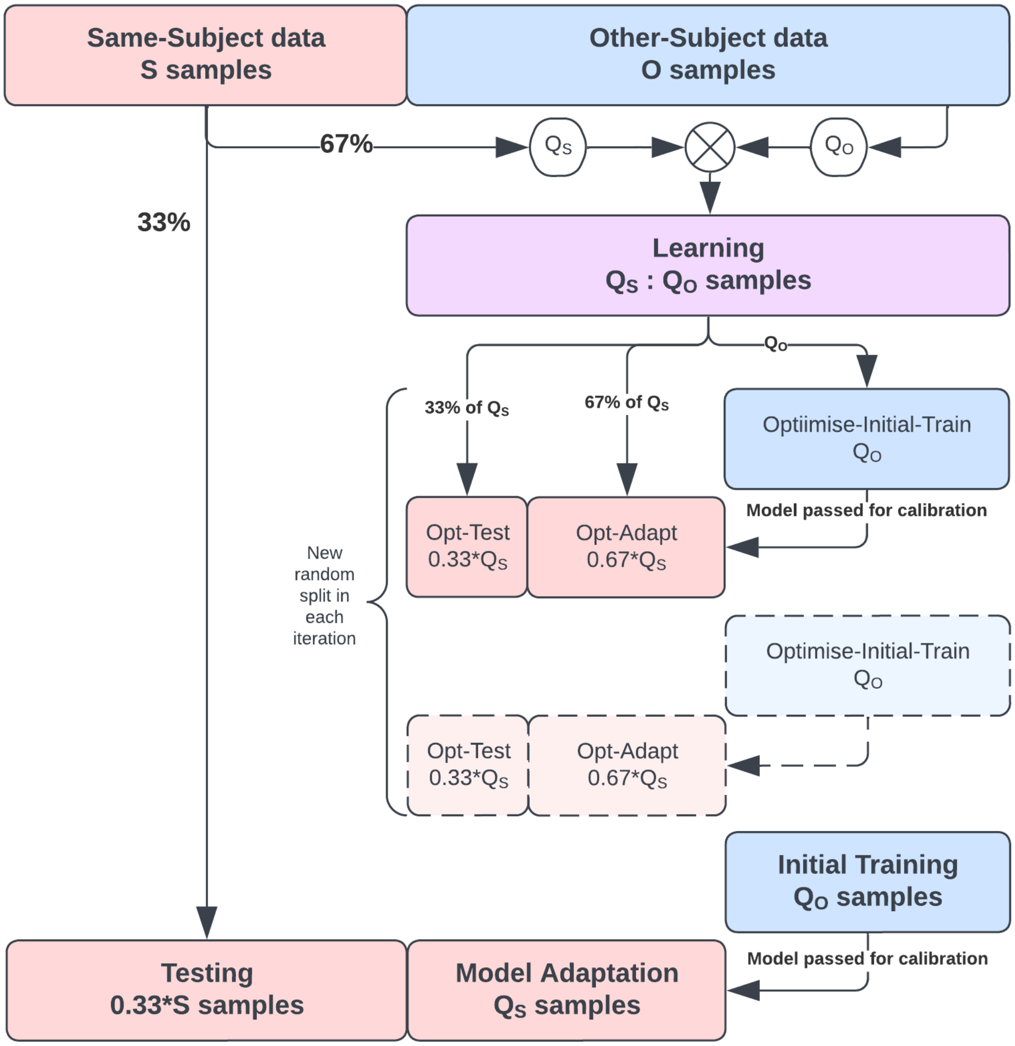


Figure 6.3: Illustration of the splitting of data within the optimisation of a Model Transfer system

Of those classification algorithms included as candidates in 5.3.3, only the Random Forest and Gaussian Naïve Bayes, in their *sickit-learn* implementations, have functionality enabling a pre-trained model to be updated. The Random Forest is capable of a "*warm_start*". An already-fit Forest can be provided new data upon which it will fit new Decision Trees, and add them to the pre-existing ensemble. In these experiments, the number of additional trees to be fitted to the same-subject data was arbitrarily fixed at 10, regardless of the size of the initial Forest. The Gaussian Naïve Bayes classifier implements a "*partial_fit*" method, with which the classwise feature means and variances of the GNB are updated. Chan et al.'s algorithm [342] for computing the joint mean and variance of two datasets from their respective means, variances, and occurrence rates is used to incorporate the same-subject data into the GNB's underlying gaussian likelihood estimations. It is noted that the GNB model had previously been eliminated from the hyperparameter space following weak classification performance in Chapter 5's experiments (see 5.5.8.2). Nevertheless the promising generalisation performance to unseen subjects observed in previous work [1], and the merit of widening the CASH search space in the interests of discouraging overfit, motivate its inclusion here. The search space for the GNB's Smoothing hyperparameter is adjusted for EMG models to be a logarithmic distribution between 1×10^{-9} & 0.5, rather than 1×10^{-9} & 1. This follows from the aforementioned experimental results which indicated a drop in Unimodal EMG-GNBs' accuracies at high Smoothing values as can be seen in full in Appendix A.1.1.3. This trend was much weaker in Generalist Unimodal EEG systems and was absent entirely from Bespoke Unimodal EEG systems (Figures A.7d & A.7c respectively), thus the boundaries of the Smoothing hyperparameter in EEG models is unchanged.

6.3.3.1 Logistic Regression

The limited range of transfer-capable models in the search space also motivates exploring the inclusion & merits of another model, to allow the potential of the Model Transfer approach to be better exploited and ensure a sufficiently wide search space as to discourage overfit. As a popular probabilistic linear classifier in various Machine Learning applications, here the Logistic Regression (LR) model was chosen. A Logistic Regression classifier fits logistic (sigmoid) curves [343] to model the relationships between the probabilities of a datapoint belonging to the various classes and the values of its features. While the Logistic Regression sees less frequent use among biosignal literature than some of the candidates covered in 5.3.3 [344] it nevertheless has precedent in both EMG [230,345–348] and EEG [179,184,249,349–352] applications, and has been observed to offer performance competitive with the domain-standard LDA model for certain EEG classification problems [350].

All the LR models in this work determined feature coefficients using the Stochastic Average Gradient descent [353] as the solver, an algorithm suitable for large datasets and for being applied to multiclass problems by fitting a multinomial regression; rather than treating each class as a binary problem, the cross-entropy loss over the whole probability distribution is fit & classwise probability predictions determined via the softmax function. The coefficients of a Logistic Regression classifier can be regularised to discourage the model from overfitting to its training data, not altogether dissimilar to the regularisaton of a Support Vector Machine as described in 5.3.3.6. Here the L2 norm was used as the penalty for regularisation with the primal formulation — that is, the sum of squares of the coefficients was suppressed — as per the default *scikit-learn* implementation.

The hyperparameter C defines the strength of this regularisation[2]. Various works among the EEG literature give no indication of including a regularisation term in their Logistic Regression models [249,351], and others such as that of Halme & Parkkonen [354] and Tomioka et al. [350] note it as having been determined empirically, such as through cross-validation over their training data, but without providing the resultant value. Cene et al.'s work applying Logistic Regression classifiers to EMG data does state the selected regularisation constant, stating that a value of approx. $1x10^{-5}$ was empirically found most suitable [346], but offers little detail on the precise mechanism by which it was determined or the range of choices explored. Lee et al.'s 2022 study by contrast notes that values of 1, 0.1, 0.01, 0.001, and 0.0001 were trialled across both L1- and L2- norm penalisation. In cases where the L2 norm was selected, the optimal regularisation constant was 1 [355]. In this work C was made a tunable hyperparameter of the CASH optimisation search space, with possible values distributed logarithmically between 0.01 & 10. As the prevention of overfit is naturally of great importance to the problem of cross-subject generalisation, lower C-values (and thus more heavily regularised models) were made a more frequently explored region of the search space.

The Logistic Regression classifiers enabled transfer learning by means of a "*warm_start*". A model is initially fit to all the other-subject data it is provided. The solution of this fit, i.e the learned attributes, are subsequently used as the initialisation point to fit a new multinomial regression to the same-subject data.

6.3.3.2 Decision-Fusion Algorithms

Of the metamodel-based Decision-Level Fusion Algorithms outlined in 5.3.3.7, only the Random Forest facilitates model transfer and thus it is the only candidate metamodel retained.

The results in Tables 5.26a & 5.34b demonstrated that in Chapter 5's experimental results there did not appear to be notable significant differences among the performances of the three candidate classifer-based Decision Fusion algorithms. For this reason it is not expected that the Logistic Regression model would be likely to offer notably different classification performance as a Decision-Level fusion algorithm; it is thus not implemented as such and is only a candidate model for the component EMG & EEG classifiers of a system as previously outlined.

Those decision algorithms which are rule-based i.e. involving no learning, and therefore for which the notion of transfer learning is inapplicable, are also retained (exclusive of the EEG-biased Average & Tunable Weighted Average which had been eliminated as discussed previously in 5.5.8.2).

The resultant search space for systems of the Model Transfer approach is thus as presented in Figure 6.4.

[2] Sometimes referred to as λ, as in [346,352].

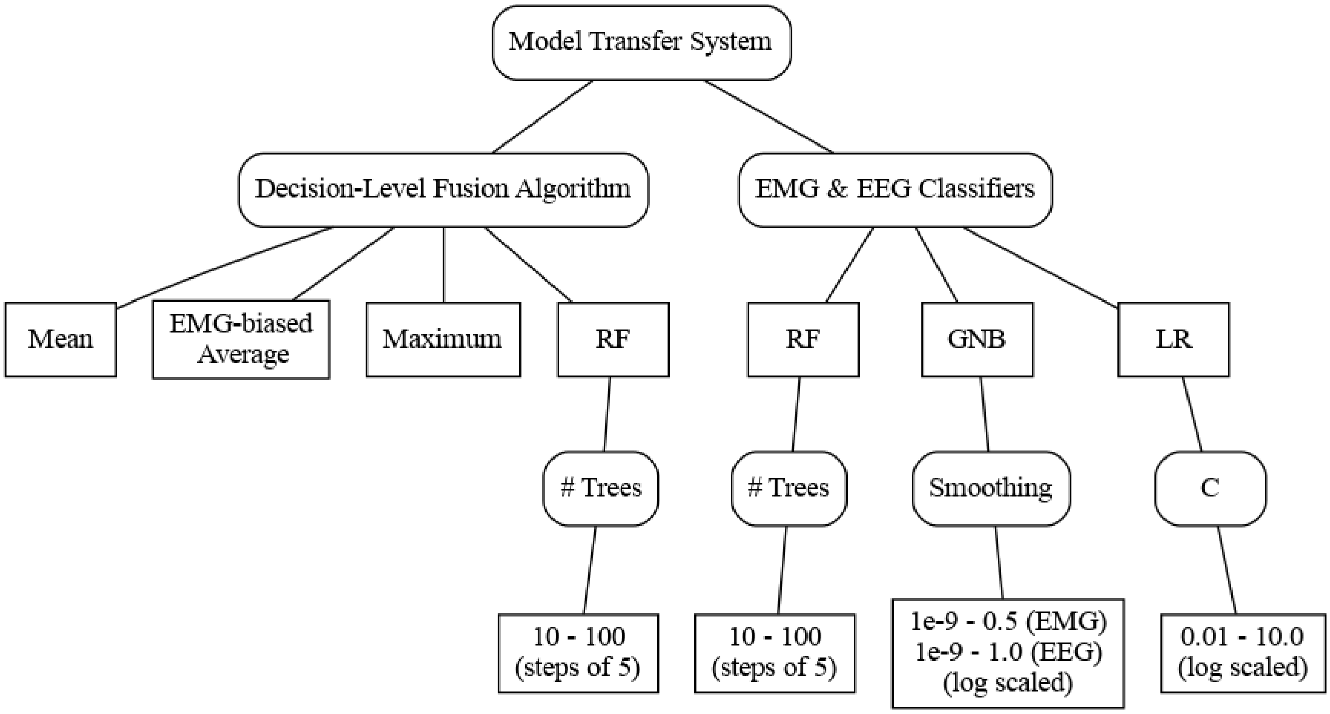


Figure 6.4: Hyperparameter optimisation space for systems using Model Transfer.
Note that as in Chapter 5, EMG and EEG classifiers did not share a hyperparameter space in optimisation but rather each held a copy of a near-identical space, presented together here for brevity.

6.3.4 Data Subsampling

6.3.4.1 Same-Subject Data

With a total of 600 individual gesture performances in each subject's dataset (per 4.2.5), following the reservation of 33% of a target subject's data for testing a given subject-specific system there remained a maximum of 400 gestures from that subject available to the system for modelling.

To explore Aim 6.2, the extent to which systems can perform well with access to lower levels of subject-specific training data (and thus hypothetically reduce the data collection requirements for a potential user of a deployed system) is investigated by downsampling this remaining same-subject data to a quantity Q_S as mentioned above.

The scaling factor for this downsampling was initially defined as a linear spacing of five values over the interval [0.01,1]. It was not however possible to use a scaling factor of 0.01; this downsampled to a Q_S of 4 gesture performances, one per class, which was insufficient to divide between the optimise-train and optimise-test data subsets within a CASH optimisation routine whilst representing all possible gesture types in each split. Instead to explore the performance of systems with access to very low levels of subject-specific data, the factors 0.05 & 0.1 were included. These scaling factors can be found in Table 6.1 along with the numbers of training gestures per class, and the total number of training gestures Q_S, for each subject, to which they correspond. This is additionally represented visually in Figure 6.5 to convey the distribution of these values.

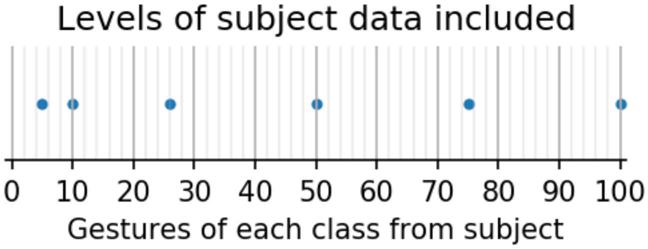


Figure 6.5: Visualisation of the levels of Same-Subject data tested

Scaling Factor	0.05	0.1	0.2575	0.505	0.7525	1
Approx Gestures per class	5	10	25	50	75	100
Approx Total Gestures Q_S	20	40	100	200	300	400

Table 6.1: Quantities of Same-Subject data & corresponding scaling factors

6.3.4.2 Other-Subject Data

Similarly, in any given case not all of the data belonging to subjects other than the target subject-under-test was used to supplement the subject's system; the other-subject data was downsampled to the quantity Q_O. Since none of the other-subject data were reserved for testing, the full 600 gesture performances from each other-subject were available; a maximum of 11400 gestures in the case of a Development subject wherein 19 not-under-test subjects could be drawn from, and of 12000 for a Holdout subject which could make use of the entire Development set. In either case, the range of Q_O values was chosen primarily on the basis of the number of gestures per class per other-subject to which they corresponded, rather than the sum total of other-subject gestures, to ensure the supplementing other-subject data could be appropriately balanced according to subject. These values were initially [0, 1, 3, 11, 15, 25, and 50] as in Figure 6.6 which corresponded to scaling factors of [0, 0.00666, 0.02, 0.075, 0.1, 0.166, and 0.33]. An additional scaling factor of 0.05263 was included which, in experiments with the Development Set, would reduce the total quantity of other-subject data to approximately 600 gesture performances — as many as had been collected from any single subject individually.Table 6.2 details the scaling factors used, the number of gestures per class contributed by each non-target subject to which they corresponded, and the resultant quantity Q_O of other-subject data in a Development Set experiment (drawing from 19 subjects) and a Holdout Set test (drawing from 20).

It should be noted that the use of 50 supplementing gestures per class per non-target subject resulted in datasets sufficiently large so as to make Support Vector Machines an infeasibly slow option for classification of EMG or EEG data (as discussed above in 5.3.3.6). Rather than avoid attempting such high levels of Q_O, the hyperparameter search space for Augmentation systems[3] of this scale (Figure 5.38) was adjusted to not include SVMs; this inability to use a potentially highly performant model[4] being considered part of the cost of using such high levels of other-subject data to augment the subject-specific dataset.

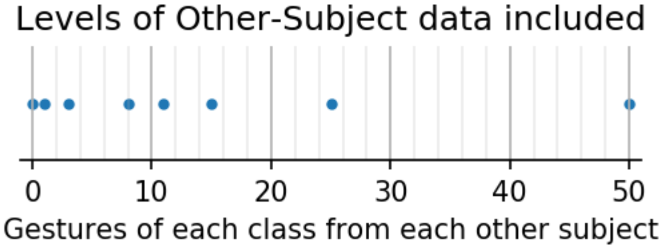


Figure 6.6: Visualisation of the levels of Other-Subject data tested

Scaling Factor	0	0.00666	0.02	0.05263	0.075	0.1	0.166	0.33
Approx Gestures per class per other-subject	0	1	3	8	11.579	15	25	50
Approx Total Gestures Q_O (Development)	0	76	228	608	836	1140	1900	3800
Approx Total Gestures Q_O (Holdout testing)	0	80	240	640	880	1200	2000	4000

Table 6.2: Quantities of Other-Subject data & corresponding scaling factors

[3] As noted in 6.3.3 the SVM is not a candidate classifier for Model Transfer systems.

[4] See Chapter 5 for exploration of the SVM's potential strengths in this problem.

6.4 Other Approaches

6.4.1 "Generalist"

Comparisons between systems in this chapter at times use a Generalist system as a point of reference. While a given Generalist model in Chapter 5 is itself subject-independent in nature — that is when tested on a subject N, the specific trained model has in no way accessed any of subject N's data nor used it in Feature Selection — the hyperparameter optimisation and selection of its component model configuration was done on the basis of maximising performance across all Development Set subjects. At each optimisation iteration the corresponding candidate Generalist system was assessed by a Leave-One-Subject-Out evaluation, being trained on $\mathbb{D} - \{N\}$ and tested on N for each N in turn; the aggregation of this assessment over the 20 Development subjects thus means that the Generalist's construction incorporated knowledge of the properties of all their data and they were not truly "unseen"[5].

By contrast, when a Generalist system was tested on Holdout subjects, their data had not contributed in any way to its development — in any discussion of tests using the Holdout data, the Generalist systems can indeed be truly said to be subject-independent.

Intuitively, it may be expected that the chosen Generalist system in Chapter 5 would hence be a configuration particularly suited to classifying Development Subjects' data, and that its performance on their data may be artificially boosted in comparison to unseen subjects. This was not however borne out in experimental results as can be seen by comparison of tables 5.4 and 5.10 and is discussed in 5.5.2. Indeed as noted in 5.5.8.1 the chosen optimised Generalist system was not significantly more performant than one chosen solely from literature inferences (and hence with no sight of any subjects' data); the optimisation appeared to provide no such benefit.

It would in principle be possible, for a fairer comparison, to develop a "true" Generalist system — akin to the "truly" Bespoke nature of the subject-specific systems of this Chapter as discussed in 6.1. This could be measured by taking the mean performance of 20 Generalist systems: a unique system for each Development Subject which did *not* consider that subject's data during optimisation, but instead found an optimal by a Leave-One-Subject-Out validation over the *remaining* 19, as in Algorithm 4. Indeed such systems could be thought of as subject-specific after a fashion, but by omission — as each would use all data *except* that of the subject-under-test — and would thus perhaps be a more direct comparator to the subject-specific systems which this chapter's experiments otherwise consider. To develop such systems however would dramatically increase the required computation time over that of Chapter 5's Generalist, by a factor of approximately 20. Given that the advantage of subject-inclusive optimisation appears to be minimal, and that the Generalist is included in comparisons here largely as a reference point rather than a competitor system design unto itself, it was not considered that this would add sufficiently worthwhile nuance to the discussion, though it may well be an area future work could seek to explore in more depth.

The Generalist system used here is thus the winning Generalist configuration with respect to Development set accuracy per Table 5.4; Feature-Level Fusion with features selected jointly from EMG & EEG.

[5] See 5.2.3 for a more complete explanation.

Algorithm 3 Generalist optimisation procedure as performed in Chapter 5, incorporating data of all Development Subjects into the selection & optimisation of system configuration.

```
for $iteration[i]$ in $100$ do
    Select candidate models & hyperparameters of $i$ via Tree-Structured Parzen Estimator search;
    for $subject[N]$ in $20$ do
        Pop $subject[N]$ from total subjects;
        Select features from 19 non-N subjects;
        Train on 19 non-N subjects;
        $Acccuracy \leftarrow$ Test on $subject[N]$;
        $Iteration\_subject\_scores[N] \leftarrow Accuracy$;
    end
    $Iteration\_scores[i] \leftarrow mean(Iteration\_subject\_scores)$;
    $Iteration\_scores\_per\_subject[i] \leftarrow Iteration\_subject\_scores$;
end
$Winner \leftarrow argmax(Iteration\_scores)$;
Chosen system $\leftarrow opt\_iterations[Winner]$;
$Per\_subject\_scores \leftarrow Iteration\_scores\_per\_subject[Winner]$;
Mean Generalist score $\leftarrow Iteration\_scores[Winner]$;
```

Algorithm 4 Optimisation procedure for developing unique Generalist systems for Development Subjects, each totally naïve to the given subjects' data at all stages including optimisation.
Note this requires a near twentyfold increase in computation by comparison to the procedure of Chapter 5 & therefore was not undertaken.

```
for $subject[N]$ in $20$ do
    Pop $subject[N]$ from total subjects;
    for $iteration[i]$ in $100$ do
        Select candidate models & hyperparameters of $i$ via Tree-Structured Parzen Estimator search;
        for $subject[K]$ in $19$ non-$N$ subjects do
            Pop $subject[K]$ from non-N subjects;
            Select features from 18 remaining subjects;
            Train on 18 remaining subjects;
            $Acccuracy \leftarrow$ Test on $subject[K]$;
            $Iteration\_subject\_scores[K] \leftarrow Accuracy$;
        end
        $Iteration\_scores[i] \leftarrow mean(Iteration\_subject\_scores)$;
    end
    $Winner \leftarrow argmax(Iteration\_scores)$;
    Chosen system $\leftarrow opt\_iterations[Winner]$;
    Select features from 19 non-N subjects;
    Train on 19 non-N subjects;
    $Acccuracy \leftarrow$ Test on $subject[N]$;
    $Per\_subject\_scores[N] \leftarrow Accuracy$;
end
Mean Generalist score $\leftarrow mean(Per\_subject\_scores)$;
```

6.4.2 Synthetic Augmentation

In this chapter transfer learning is discussed in the context of supplementing a system with data collected from other human subjects. Undergoing such data collection naturally carries a time implication, financial costs, and indeed the burden of actually recruiting individuals from whom to collect said data — while in this work these factors were not of direct consequence as the data had been already collected by Jeong et al. [198], their impact should not be ignored. Reductions in these costs which avoid negatively affecting the performance of developed systems would naturally be of great interest not only to future research but to those seeking to deploy biosignal gesture classification systems in the "real world".

The author's prior work [1] evidenced that Generative A.I. models (specifically *OpenAI*'s *Generative Pretrained Transformer 2* [356]) could produce artificial raw, continuous biosignal data of sufficient apparent quality to be of demonstrable value in augmenting both EMG and EEG classification tasks. This work was notable as being believed the first study to evidence GPT-2's capability for this application (subsequently corroborated by [357]), and the only demonstration of GPT-2-generated fake EEG which verifies the time- and frequency- domain characteristics of the synthetic signals [358], The classification tasks to which it applied the synthetic augmentation however were coarser, more easily separable problems than the same-hand grasp types explored in this work. It remains to be found whether generative AI can learn characteristics of biosignal data with such specificity that the synthetic data produced would be discriminable between such similar gestures as these grasps. Additionally, the application of the synthetic augmentation approach to both EMG and EEG data in [1] served to explore the possible capabilities of the method in distinct but not wholly dissimilar domains; the two classification problems were in themselves unrelated and the findings give no guarantee of its applicability to a multimodal problem such as in this work wherein EMG & EEG data need be synchronous.

It should also be noted Jeong et al.'s dataset which is used in this work comprises 25 human subjects [199] which, while certainly a small sample of the global human population, is of comparable scale to many standard datasets used in BCI research. The work of Bird, Pritchard, et al. [1] did not explore comparatively the extent to which models could be positively impacted by augmentation with GPT-generated data and by augmentation with additional human subjects. Plausibly it could be assumed that where real human data is available, this would offer greater benefit to a system, though the potential for generating large amounts of artificial subject-specific data should not be ignored.

There would evidently be great merit in future work investigating the applicability of GPT-generated synthetic data augmentation to multimodal systems and to those with more similar gestures than the ones explored in [1]. Indeed there would likewise be significant benfit of reviewing the relative merits of augmentation by synethesisation of data against those of augmentation by subject recruitment and collection and the cost-benefit analyses involved therein. Considering the various differences between the classification tasks trialled in [1] and that of this work however, and the ready availability of a good number of additional human subjects in [198] with whose data a system can be supplemented, these lines of investigation are considered out-of-scope of the current work; the synthetic augmentation approach was not applied here[6].

For the reader's interest however, given its clear relevance to the topic of biosignal augmentation discussed in this chapter, portions of the aforementioned are included with this work as Appendix C. In accordance with the co-first-authors' mutual agreement on accreditation of the work (see *Collaboration Acknowledgements* above) only those parts related to its application with EMG data are covered; the remainder can be found in the work in full [1].

[6] It should perhaps also be considered that while the participants in Jeong et al.'s study gave their clear and full consent for their biosignal data to be published and re-used in future research [198], the ethical climate with regard to generative AI models has shifted rapidly in recent years. At the time of the dataset's publication the possibility of their data being used to train generative AI was unlikely to have been a prominent consideration of the participants. Indeed even at the time of the author's prior work in Bird, Pritchard, et al. [1] being carried out, the topic was significantly underdiscussed. In 2024 however there is much recent discussion, and indeed many ongoing legal cases, regarding the ethical implications in sourcing of the data used to train generative AI. A study applying this method in the current climate, or indeed any which makes its participants' data publicly available for use by others in the research community, ought perhaps to consider carefully the need to seek express permission from its participants regarding use of their data in this way.

6.5 Results

6.5.1 Experimentation with Development set

Figure 6.7 presents, for all combinations of same-subject and other-subject data quantities Q_S & Q_O, the mean classification accuracy across Development set subjects obtained by both the Dataset Augmentation and Direct Model Transfer approaches.

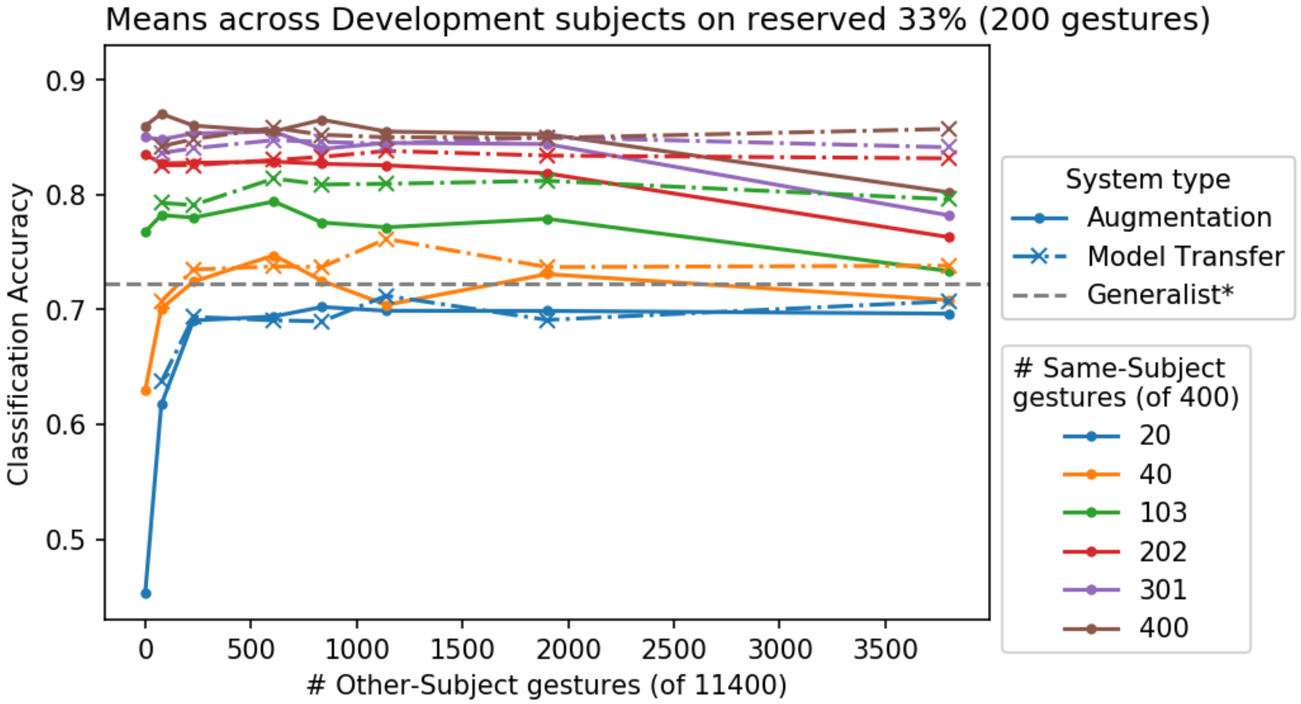


Figure 6.7: Mean classification accuracies achieved with different levels of Same-Subject and Other-Subject data by Augmentation and Model Transfer systems.
NB that as discussed in 6.4.1 the "Generalist" seen here was not fully subject-independent, as it incorporated Same-Subject data into the model selection & optimisation process.

A number of trends are immediately evident. Aim 6.1 sought to identify whether learning from other-subject data could boost a subject-specific system's performance. The results in Figure 6.7 certainly indicate that, provided a system has access to a sufficient quantity Q_S of same-subject data, the inclusion of data from other subjects by either of the two approaches does not appear to result in a meaningfully improved classification performance. Those systems with access to 300 or greater same-subject gestures were in fact largely degraded by incorporating other-subject data. It may be noted here that this finding is reinforced in [180] which, while using a distinct technique for transfer learning to those trialled here, similarly found that systems with sufficient data belonging to a target subject were not notably improved by inclusion of data from other subjects. Systems with access very low levels of data belonging to the target subject however were indeed improved by the presence of other subjects' data, though this impact appeared to rapidly reach a saturation point beyond which further other-subject data was of little to no additional benefit, and does not seem to frequently achieve accuracies notably higher than a baseline Generalist model (albeit noting the caveats to this comparison outlined in 6.4.1 above).

It should also be noted that those Augmentation systems with datasets sufficiently large as to preclude the use of SVMs as noted above in 6.3.4.2 were generally weaker than their equivalents with fewer data; given the apparent strength of the SVM for this problem seen in Chapter 5, particularly with regard to the component EMG classifier of a system, this is not altogether surprising. In the majority of cases other than this exception however, the difference between the score of a Dataset Augmentation system and a Model Transfer one with access to equal levels of subject-specific and other-subject data was minimal; the two approaches performed equivalently well on the whole, though there appears a slight preference for Model Transfer among systems with minimal subject-specific data and for Augmentation among those with large quantities.

Aim 6.2 asked if "*inclusion of other-subject data* [can] *allow the same level of performance to be obtained with a reduced amount of same-subject data*". At no point in Figure 6.7 does the plot of a system with access to a lesser quantity of same-subject data surpass that of one with more, suggesting that cross-subject supplementation was not able to meaningfully reduce the data requirements for a subject in this way. Figure 6.8 presents the same data as Figure 6.7 but with the x and *colour* axes essentially swapped; systems are grouped by the quantity of other-subject data Q_O to which they had access & thus the trends related to the same-subject data levels are more easily visible. This highlights that the level of same-subject data Q_S is in fact the dominant factor in a system's performance by a significant degree; the level of data used to supplement the system has a much lesser observable impact. The response of systems to increasing levels of same-subject data in distinctly similar in nearly all cases, with the notable exceptions of Augmentation by one-third of the available other-subject data (the aforementioned case wherein dataset size disqualified SVMs) and those systems which were wholly subject-specific with no form of transfer learning. The latter were notably weaker than their competitors when given access to very low quantities of same-subject data; this is hypothesised to be in part due to their total amount of modelling data being simply too low for the system to learn patterns which generalised well beyond that training data. It does appear however that such a reduction of same-subject data can actually be achieved *without* transfer learning. The beneficial effect of increasing the level of subject-specific data provided to systems appears to begin to saturate at approximately 300 same-subject gestures.

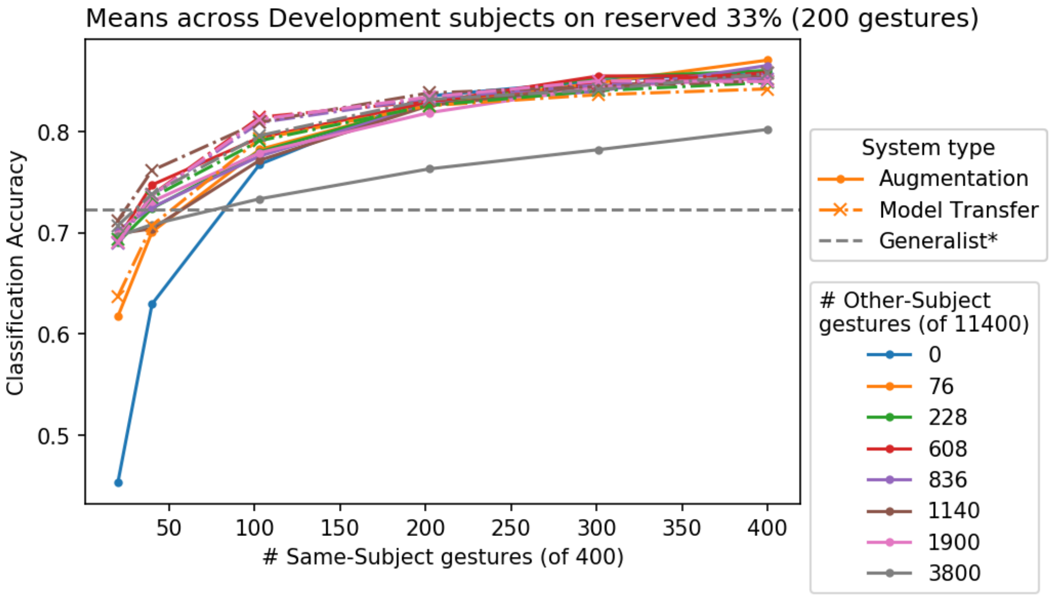


Figure 6.8: Mean classification accuracies achieved with different data levels by Augmentation and Model Transfer systems, as in Figure 6.7, presented instead by level of Same-Subject data.

The key observations from these results which can help inform design of gesture classification systems are formalised as follows, to enable their validation upon unseen data from the Held-out subjects:

- The benefit to system accuracy of increasing the amount of same-subject data available for learning saturates once a system has access to approximately 300 same-subject gestures
- Where the benefit of additional same-subject data has saturated, inclusion of other-subject data will not improve system accuracy
- At such levels of same-subject data, direct use of Model Transfer is no stronger a method of transfer learning from other-subject data than simply using such data to Augment the dataset (provided the resultant dataset is not infeasibly large), though neither are more performant than a subject-specific system
- Where the level of same-subject data accessible to a system is very low, incorporation of other-subject data will improve classification accuracy
  - Such improvement will reach accuracies no better than a Generalist system (i.e. one with no access to a subject's data) would

6.5.2 Validation of findings with Holdout data

The above outlined observations are validated in turn using the Holdout dataset. For any given system discussed, every Holdout Subject is tested in isolation, with all 20 Development Subjects considered other-subjects from which to draw data.

6.5.2.1 Same-Subject Saturation

Verifying the saturation in accuracy of adding more same-subject data can be done simply by assessing the response of Holdout-subject-specific systems to increasing levels of same-subject data, the result of which is presented in Figure 6.9. These results are taken as the mean of five trials at each quantity of same-subject data Q_S, in efforts to reduce the risk of the variation in the random selection of 33% of a subject's data, or the randomness of the subsequent downsampling of the subject-specific learning data to Q_S, unduly influencing the results.

It should be noted that classification accuracy trends lower in Holdout subjects than in Development. Chapter 5 had similarly found Holdout subjects' Bespoke accuracy to be lower than that of Development subjects (see 5.5.2), suggesting this perhaps to be a facet of the inherent seperability of the underlying data. It should also be noted though that as discussed in 6.3.1.3 above, the narrowing of the optimisation hyperparameter space in 5.5.8.2 was on the basis of considering the various options' performance with regard to accurate classification of Development subject data. It is plausible that in doing so the search space had specialised to those subjects, thus accounting for a slight artificial boost in performance. Nevertheless Figure 6.9 clearly demonstrates that the trend here was distinctly similar to that observed with the Development Set; in both cases the classification accuracy is a monotonic increasing function of the quantity of same-subject data available, and likewise the saturation of each is clearly similar.

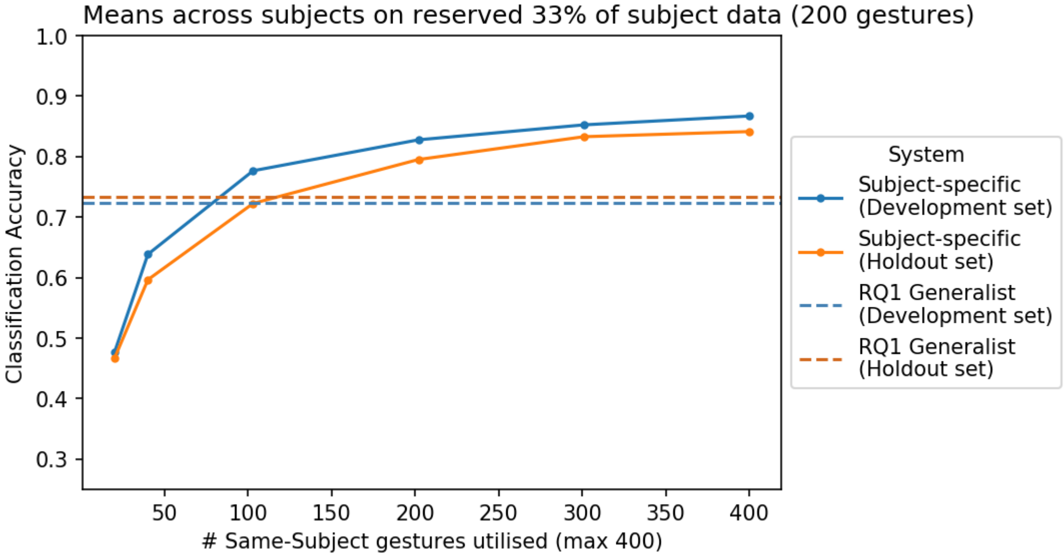


Figure 6.9: Classification accuracies of subject-specific systems with access to varying amounts of same-subject data.

6.5.2.2 Merit of inclusion of Other-Subject Data

To verify the lack of benefit from including other-subject data at this saturation point, a candidate augmentation system & model transfer system need be put forward to compete against the unaugmented subject-specific system. To ensure a valid comparison this selection cannot be made with prior knowledge of the Holdout set's performance. In the Development set, the Augmentation approach offered greatest performance with 608 other-subject gestures and the strongest Model Transfer system made use of 1900 other-subject gestures (Figure 6.7). The equivalents of these Q_O values (as per Table 6.2) in the context of a Holdout subject are 640 & 2000 gestures respectively. Table 6.3 presents the classification accuracies of such configurations with the Holdout data.

Subject	Candidate System under test (& quantity of Other-Subject data)		
	Saturated Subject-Specific	Augmentation (640 Other-Subject gestures)	Model Transfer (2000 Other-Subject gestures)
1	0.729798	0.780303	0.768939
6	0.819444	0.813131	0.747475
11	0.939394	0.904040	0.901515
16	0.741162	0.804293	0.789141
21	0.862374	0.825758	0.816919
Mean	0.818434	0.825505	0.804798

Table 6.3: Classification accuracy per Holdout subject of winner candidate Augmentation & Model Transfer systems, and of the wholly subject-specific approach, where systems have access to 301 subject-specific gestures for modelling.

Both the Augmentation and Model Transfer systems are of greater complexity than the Subject-Specific approach, and both rely on the advance collection of biosignal data from a significant number of individuals with which to supplement the same-subject data. This additional expense means that such approaches only merit use if they outperform the subject-specific system. If they are only equivalently accurate their adoption would not be worthwhile; were these systems able to obtain equivalent with fewer subject data that reduced burden to an end-user may justify them, but as noted above this was not the case. The following null hypotheses are thus derived from Aim 6.1 (whether "*inclusion of other-subject data* [can] *boost performance above that achievable with same-subject data alone*") and a paired one-tailed t-test used to test each:

- "the candidate Augmentation system will not provide classification accuracy significantly greater than that achieved with the subject-specific approach (6.1)"
- "the candidate Model Transfer system will not provide classification accuracy significantly greater than that achieved with the subject-specific approach (6.2)".

$$H_0 : \mu_{augmentation} - \mu_{subject-specific} \leq 0. \tag{6.1}$$

$$H_0 : \mu_{model\,transfer} - \mu_{subject-specifc} \leq 0. \tag{6.2}$$

Considering first the system Augmented by 640 Other-Subject gestures, the t-test's assumptions of normality of differences and equality of variances are initially verified with the Shapiro-Wilk test ($W = 0.84921$, $p = 0.192$) and an F-test ($F = 3.4479$, $p = 0.2578$) respectively. The result of the t-test between the Augmentation & Subject-Specific systems in Table 6.3 is $t = 0.33486$, $p = 0.3773$. The test's null hypothesis is not rejected, indicating that Augmentation was not significantly better than the un-supplemented system.

Subsequently testing the Model Transfer system, again the t-test's assumptions are verified with a Shapiro-Wilk result of $W = 0.86563$, $p = 0.2492$ indicating normality of differences and the F-test resulting in $F = 2.1225$, $p{-}value = 0.484$ suggesting there is indeed equality of variances. The result of the t-test between the candidate Model Transfer system & the Saturated Subject-Specific is $t = 0.5666$, $p - value = 0.6994$. This null hypothesis is also not rejected; the candidate Model Transfer system was also not significantly more accurate than the Subject-Specific approach.

6.5.2.3 Model Transfer vs Augmentation

The subsequent finding, that "direct use of Model Transfer is no stronger a method of transfer learning from other-subject data than simply using such data to Augment the dataset", provided sufficient same-subject data is made available to a system, is assessed by comparing the classification accuracies over the Holdout subjects of both approaches in at the saturation point of $Q_S \approx 300$. That the Model Transfer approach does not outperform the Augmentation approach is then verifiable simply by observation of Figure 6.10 which plots these results[7].

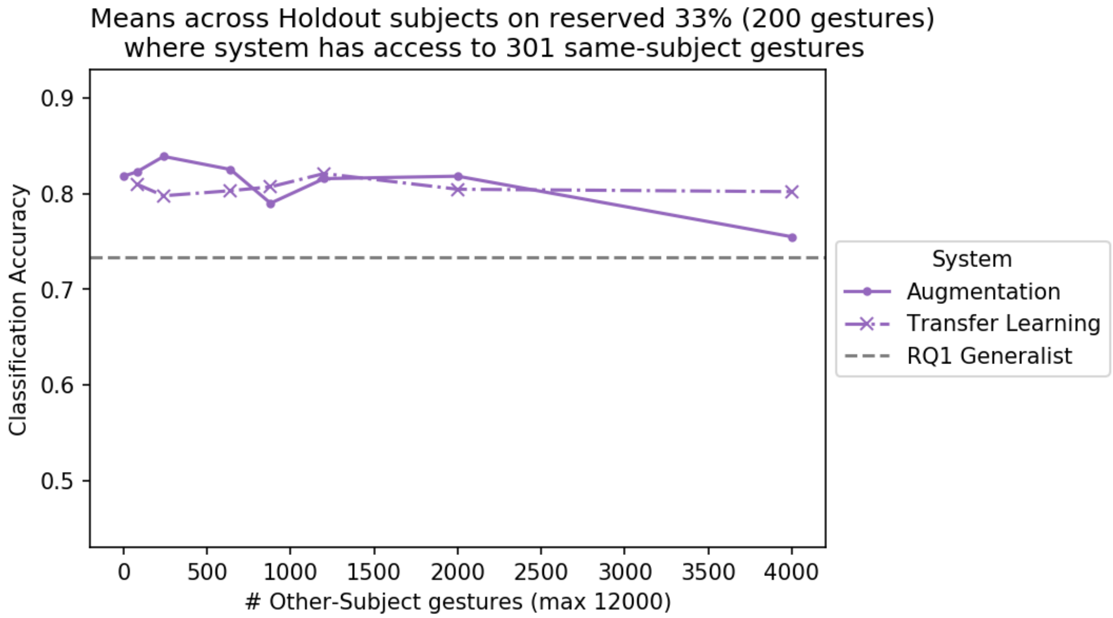


Figure 6.10: Augmentation and Model Transfer systems with 301 Same-Subject gestures & varying levels of Other-Subject data

[7] With the exception of the case where $Q_O = 3800$, preventing the use of SVMs in Augmentation systems as previously described (6.3.4.2), which is included in Figure 6.10 for completeness but otherwise discounted from this comparison of approaches.

6.5.2.4 Supplementation in cases of scarce Same-Subject data

Experiments with the Development Set gave rise to two observations regarding systems with access to low levels of subject-specific data (i.e. those with 10 or fewer same-subject instances per gesture). The first of these, that in such cases the incorporation of data from other subjects will improve classification accuracy, is verified by observation of Figures 6.11a and 6.11b. These present respectively the results of Augmentation and Model Transfer systems with access to various levels of other-subject data wherein the quantity of same-subject data $Q_S \leq 40$. It can be clearly seen that all supplemented systems outperformed those which were unaugmented[8].

The second aspect of this observation, that such supplemented systems would nevertheless fail to outperform a subject-independent Generalist, appears likely from a simple comparison of means — none of these systems' mean classification accuracies across the Holdout subjects surpassed that of the Generalist. It should be recalled here that as discussed above in 6.4.1 the Generalist of Chapter 5 is indeed legitimately subject-independent with respect to the Held out subjects.

Subject	System		
	Chapter 5 Generalist	Augmentation (640 Other-Subject gestures)	Model Transfer (880 Other-Subject gestures)
1	0.66333	0.67298	0.63131
6	0.74750	0.68687	0.72980
11	0.82333	0.82828	0.84217
16	0.72458	0.70707	0.68056
21	0.71208	0.72854	0.67046
Mean	0.73416	0.72475	0.71086

Table 6.4: Classification accuracy per Holdout subject of winner candidate Augmentation & Model Transfer systems with access to 40 Same-Subject gestures

To ensure the veracity of this apparent lack of superiority however the claim can be tested statistically, with the null hypotheses that:

- "The best attempt at an Augmentation system with access to 40 or fewer subject-specific gestures will not reach classification accuracy greater than that achieved by a Generalist model (6.3)"
- "The best attempt at a Model Transfer system with access to 40 or fewer subject-specific gestures will not reach classification accuracy greater than that achieved by a Generalist model (6.4)".

$$H_0 : \mu_{augmentation} - \mu_{generalist} \leq 0. \tag{6.3}$$

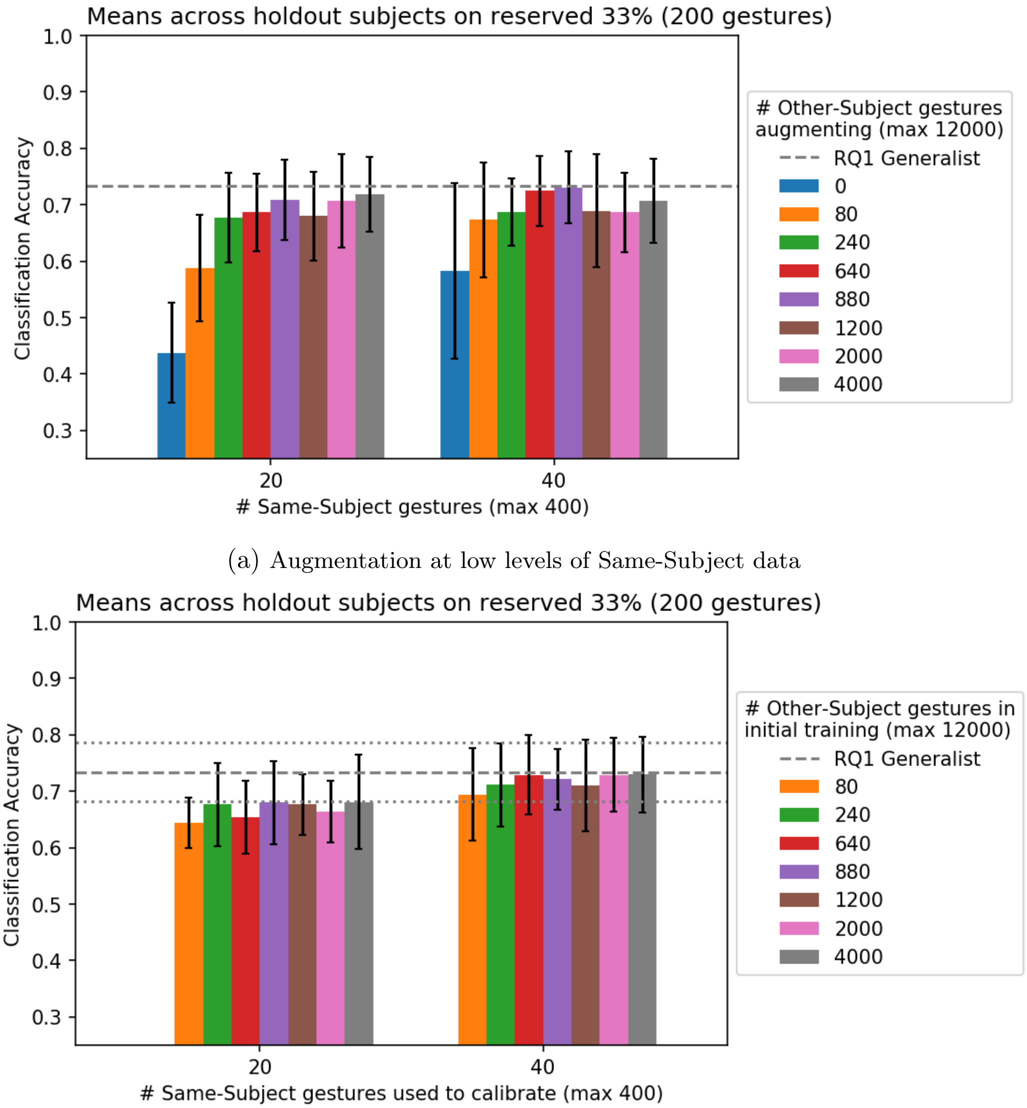
$$H_0 : \mu_{model\,transfer} - \mu_{generalist} \leq 0. \tag{6.4}$$

As before, for a valid test the candidate systems are selected on the basis of their Development Set results as seen in Figure 6.7; the strongest Augmentation system of those where $Q_S \leq 40$ being one utilising 40 same-

[8] NB that while Model Transfer systems by definition cannot make use of no other-subject data, as some is required for constructing the initial models which the same-subject data is used to adapt, the scores in Figure 6.11b can be compared to those of the subject-specific unaugmented systems in Figure 6.11a, which they readily exceed

subject gestures and 608 other-subject (which corresponds to 8 per subject per class and thus $Q_O = 40$ for a Holdout subject, per Table 6.2), and the strongest Model Transfer system drawing similarly on 40 subject-specific gestures but 1140 other-subject gestures (15 per subject per class thus 1200 total for a Holdout subject). The subject-wise classification accuracies of these candidate systems for the held out subjects can be seen in Table 6.4.

Following the same process as in 6.5.2.2 the hypotheses 6.3 & 6.4 are similarly tested with paired one-tailed t-tests. In the case of Augmentation, following confirmation of differences being normally distributed with a Shapiro-Wilk test ($W = 0.85033$, $p - value = 0.1956$) and of variances being equivalent with an F-test ($F = 0.90535$, $p - value = 0.9255$), the t-test between the candidate Augmentation system & the Generalist of Chapter 5 resulted in a t-statistic of -0.6719 at a p-value of 0.7308. This is well above the $alpha = 0.05$ significance level and thus the null hypothesis is not rejected; the Augmentation system's accuracy was no greater than that of the Generalist. Likewise with the candidate Model Transfer system, after normality of differences & equality of variances are confirmed (Shapiro-Wilk $W = 0.85394$, $p - value = 0.2072$; F-test $F = 0.51834$, $p - value = 0.5401$), the t-test is performed with a resultant t-statistic of -2.0252 and p-value of 0.9436. This again indicates the null hypothesis not to be rejected and that the Model Transfer approach was not significantly more accurate that the subject-independent Generalist system.



(a) Augmentation at low levels of Same-Subject data

(b) Model Transfer at low levels of Same-Subject data

Figure 6.11: Response of Augmentation and Model Transfer systems with low quantities of subject-specific data to increasing levels of other-subject data with which to supplement

6.5.3 Conclusions

It appears from these results that in this gesture classification task the incorporation of data from other individuals through transfer learning, either by augmentation of a dataset or a direct model transfer, is of minimal benefit to a gesture classification system — provided that sufficient data from the subject for whom the system is being developed is available. In systems with only a very small quantity of subject-specific data, no approach (including that of a wholly subject-specific model) was able to make effective use of such a low amount of data; a system which did not use it was equivalently good. If the most user-specific data that can be collected for tailoring a system is such a small amount, there seems to be no merit in doing so.

Given the interest among literature in cross-subject classification, these results are perhaps surprising. Among works attempting cross-subject transfer learning however it is very rare for both cross-subject and subject-specific approaches to be directly compared [359]. The omission of such tests makes it hard to confidently state that the cross-subject learning in such works meaningfully contributed to their performances. By contrast this chapter's experiments assessed the impact of transfer learning on systems permitted different levels of subject-specific data, observing that it was indeed beneficial, but only when a subject's data were so sparse that greater accuracies could be obtained by forgoing subject-specificity altogether — corroborating findings that systems with sufficient subject-specific data may be unaffected by transfer learning [180].

While in Chapter 5 reasonable subject-independent performance could be achieved, this seemingly did not reliably translate into effective cross-subject calibration here. This may indicate that the cross-population trends learnable from the biosignal data were very distinct from those trends which were most predictive within any given individual's data; an algorithm could learn to model either sample-wide patterns or subject-localised ones, but they did not complement one another well. It may even be that some algorithms when trained on an augmented dataset simply model each subject's data separately in distinct parts of the modelling space, and thus gain little from the inclusion of multiple. This may suggest some stratification is needed for cross-subject transfer learning to be of maximal use. As discussed in 6.3.1.1 & 6.3.3, certain studies have selected sources of augmentation data on the basis of empirical similarity to a target subject's biosignals [182, 184, 337], and others have used preprocessing techniques to align data across subjects [178, 180, 185]. Both methods however have limited suitability for a real-world system; applying such selection or manipulation to novel real-time data would be likely infeasible. Demographic similarity could plausibly be a viable proxy for selecting data with which to augment a system, but to investigate this robustly would require much larger datasets than are currently available to the biosignal research community.

These findings thus have implications for the design of biosignal-based gesture recognition systems, in that the suitable choice of approach is highly dependent on the anticipated use case. Where it is expected that a significant quantity of user-specific data will be able to be collected, such as in the case of a prosthesis user receiving specialist care during the rehabilitation process, a wholly subject-specific system appears more appropriate. By contrast in a scenario where such support is unavailable, and a system is required to operate "off-the-shelf" and a user may be unwilling or unable to undergo extensive data collection routines for its calibration, the collection of a small quantity of calibration data would be of insignificant benefit and the system could be simplified by being made a fully subject-independent Generalist.

Cross-Session Classification: An Exploration of Strategies for Reducing Calibration Burden

7.1 Aims & Overview

Having established the necessity of subject-specificity in Chapter 6 it is important to consider the perspective of a deployed system's end user. A user would naturally expect a device such to classify their intended gestures accurately on each occasion they used it, while requiring minimal effort on their part to elicit such levels of performance; an accessibility device should not in itself present unnecessary barriers to use. Biosignal data collected on different occasions however may differ in distribution [360]. While variation in data obtained from the same subject is generally lesser than it is between individuals [171], a number of factors can affect within-subject consistency. Environmental conditions such as the level of background electromagnetic activity can add noise, to which EEG is particularly sensitive. Changes in a subjects' physiology can also be impactful — in the longer term changes in muscle mass and fat distribution among other factors can affect EMG data, and even shorter term phenomena like perspiration levels can alter the electrical impedance between the skin and electrodes placed upon it, degrading the quality of recorded signals. Unintentional inconsistencies in the placement of electrodes, or adjustments to them once fit, likewise have an effect [361].

Data obtained within a given session is hence likely to be the most predictive of further data from that same session. This could logically motivate ensuring a system's per-session accuracy by way of per-session modelling: requiring, on each occasion it were utilised, an individual to supply sufficient labelled session-specific data to train a model then used to classify data of only that specific usage session. Thus arises a potential conflict between the desired goals of maximising a system's accuracy and minimising the burden placed upon its user. An ideal system would be one better able to generalise to novel sessions while minimising the required amount of training data needed from each — or eliminating such a need entirely by classifying on a wholly cross-session basis. It may even be that per-session training is not necessarily the most performant mechanism by which to use a subject's session-specific data, if data sourced by other means could be leveraged effectively.

Literature on cross-session learning is limited; whether by deliberate choice or a necessity resulting from constraints on available data, many studies operate on a single-session basis [147, 184], including the overwhelming majority of notable multimodal biosignal studies discussed in 3.1 [129, 130, 139, 141, 145, 146]. Ozdenizci et al. [113] is one rare exception, having trialled their cascaded approach to EEG-EMG fusion on an uncalibrated cross-session basis by attempting to predict data of a target session using a model trained on all data from a given subject's preceding sessions. Though results were highly variable between participants, this reached only a mean accuracy of 28.8% for multimodal classification. Techniques for model adaptation or other forms of calibration were not explored; thus highlighting the challenges in naïve cross-session classification. Seeland et al. [295] did explore an adaptive strategy for cross-session classification. A model trained on out-of-session EEG data used EEG data of the designated test session to predict whether an individual was moving their left hand, right hand, or remaining at rest. An EMG-based thresholding mechanism was incorporated to detect movement onset and used as a source of "ground truth" to supervise the EEG classifier. Their incoming EEG data at testing time were thus labelled by this EMG algorithm, such that they could be subsequently used to retrain the EEG model. This strategy was found capable of improved classification accuracies over unadapted systems, by a greater margin in those cases where baseline cross-session accuracy was lower. However, such an approach hinges on the reliability of the EMG model as a source of accurate labels. In their case since only movement onset was being detected this was naturally very high; the approach would unlikely be well-suited to a multi-gesture problem wherein EMG-based classification were itself more challenging. These studies it should be noted were severely limited in sample size, with Ozdenizci et al.'s sample consisting of three subjects [113] and Seeland et al's only one [295]; this work by contrast uses Jeong et al.'s dataset of twenty-five subjects [198]. While only three sessions are available from each of the twenty-five — compared to five sessions in [113] and fourteen in [295] — by assessing a wider range of strategies over a larger population, this work offer a more extensive and thorough exploration of cross-session adaptation in multimodal biosignal gesture classification than those which have gone before.

Some insights into within-subject cross-session classification approaches can be gleaned from single-data-modality studies, though again this is a somewhat underexplored area of literature [182]. As was the case for the cross-subject problem of Chapter 6 some proposed techniques focus on the adaptation of Common Spatial Pattern filters, such as the development of of "prototypical" spatial filters ostensibly requiring little-to-no calibration in new sessions by Krauledat et al. [362]; as CSPs are not used in this work, per the discussion in 4.2.5.1, such strategies not of great relevance here. Li et al. [363] proposed a method of semi-supervised SVM modelling, training initially on a subset of EEG data corresponding to the first three characters of a subject's attempted use of a "P300 speller" Brain-Computer-Interface[1] and using subsequent characters to successively retrain, a form of direct model adaptation. Du et al. [175] applied a somewhat similar principle to inter-session transfer learning in the continual unsupervised adaptation of a Convolutional Neural Network for classifying a range of gestures from EMG data, noting a significant increase in cross-session accuracy over an unadapted CNN. Raza et al. [213]'s work explored unsupervised adaptation in classifying Motor Imagery from EEG. Their system used Covariate Shift Estimation to identify drift in the data incoming to a model

[1] see 3.1.2.1 for a description of the P300 speller

pretrained on out-of-session data, and subsequently a Probabilistic-Weighted KNN to determine if such novel data & its predicted label should be incorporated to the training dataset — if so, a new algorithm modelled on the expanded dataset was added to their classification ensemble. In this way previous learning did not strictly need to be modified but rather embellished or in effect augmented by the novel information. Abu-Rmileh et al. [364] saw success in updating EEG classifiers with a "batch" approach: retraining LDAs regularly on only subjects' most recent attempts at a motor imagery task performed over four consecutive days, finding this led to improved accuracies over the continued use of models trained on data of the first day. In subjects' earlier trials of each session, models were trained on both data of that day and the previous — this can be seen as akin to a dataset augmentation approach using both out-of-session and within-sesion data. In all these cases the caveat should of course be noted, as it was in 5.3.4, that strategies' success in unimodal systems may not necessarily indicate their suitability for multimodal ones.

The dataset used in this work comprises biosignals recorded from twenty-five subjects each on three separate occasions, each a week apart [198].The Bespoke systems seen in Chapters 5 & 6 shuffled all data belonging to a given user, selecting a stratified random third to be tested upon, with the remainder available for modelling; both their training and testing datasets contained data collected from each of a subject's three recording sessions. Such offline classification as performed earlier in the work therefore does not fully align with a deployment setting, wherein models would classify data of only one session at any given point in time.

Here, a range of possible options for cross-session learning are instead considered. One such approach (detailed in 7.3.3.1) does indeed involve learning directly from data of multiple sessions, akin to the modelling of previous chapters' Bespoke systems. Other strategies are introduced however, their choice particularly informed by hypothetical use-cases of a deployed system & characteristics which could potentially be desirable in such scenarios. Certain approaches for example use no session-specific data for training or calibration at all, and instead model on data collected from the subject in prior sessions. They thus allow for a "pick-up-and-play" mode of use at the expense of requiring the user to provide training data in advance of to a usable system being deployable. Others avoid drawing on such out-of-session data, relying only on that which is session-specific or was contributed by other individuals entirely, eliminating the need for advance collection of a subject's data before deployment. Systems are evaluated across a range of possible levels of session-specific data, to explore both explore their ability to perform with minimal calibration, and the upper extent of their achievable classification accuracies should the quantity of calibration data not be a barrier.

The chapter's key focus is the aforementioned desire to reduce the burden, in terms of time and effort, placed on the user of such a potential system. This burden manifests most obviously as the time spent providing the amount of labelled session-specific calibration data required for training or adapting a given system. Certainly this is a critical element, considering that a new "session" for a user would constitute to each "wear" of a prosthesis & hence could well mean multiple daily calibration sessions. The data collection time however should not be viewed as the only contributing factor. A system's need to determine suitable modelling choices (e.g via data-driven Combined Algorithm and Hyperparameter (CASH) optimisation) on a per-session basis will naturally also result in a lengthier modelling process over one which does not carry out this step, a potential additional delay to the system being ready for use. For this reason both approaches which undertake session-specific CASH optimisation, and those which predetermine a “static” model configuration by other means and “port” it to new sessions, are explored here.

That Jeong et al.’s dataset includes three repeat data collection sessions for each participant [198] enables an ancillary investigation into uncalibrated cross-session classification. The respective accuracies of systems trained on data gathered two weeks before a designated testing session and those using data collected one week in advance are assessed. Should a notable difference in performance be found this could imply limitations in the longevity of cross-session classification systems. Likewise explored is the extent if any to which learning from data collected on multiple prior occasions, rather than a single one, improves models’ generalisation to a novel session. Having access to a greater diversity of data could result in models more robust to further variation in factors such as sensor fit & environmental conditions that can affect biosignal properties — which may be indicative of appropriate training strategies for future gesture classifier development. These experiments are conducted first; their findings assumed a suitable proxy indicator of the appropriate out-of-session data on which to draw in calibrated cross-session systems.

The work in this Chapter is more exploratory in nature than that of previous, as is reflected in its Aims outlined below:

- **Aim 7.1** *Investigate the impacts of data diversity and source–target time delay on uncalibrated cross-session biosignal gesture classification.*
- **Aim 7.2** *Explore strategies for session-specific gesture classification which do not require prior collection of subject-specific biosignal data.*
- **Aim 7.3** *Explore the impact of the amount of session-specific biosignal data available to gesture classification systems on the accuracy of their gesture classifications of a target session’s data.*
- **Aim 7.4** *Identify suitable approaches for session-specific gesture classification with access to different quantities of target-session biosignal data.*

7.2 Methodology

7.2.1 Overview

The broad structure of the experiments in this chapter is not altogether dissimilar to that of Chapter 6. Classification systems, each implementing one of a number of approaches and which are both subject- and session-specific in nature, are created for each subject N in turn. These systems are designed with the subject's third and final data collection session as the session-under-test, with one third of this target session's data being reserved for testing.

Each system is provided with one or more of the following categories of data for modelling, according to its associated learning strategy (defined fully in 7.3):

- Session-specific / "calibration" data: Unreserved Session 3 data belonging to the subject in one of a range of quantities;
- Out-of-session data: Data collected from the subject in recording sessions other than the session-under-test (i.e. Sessions 1 and/or 2);
- Other-subject data: Data collected from subjects other than N.

A portion of the unreserved Session 3 data (i.e. that which is remaining after the aforementioned designation of one-third as test data) is defined as the "Calibration" data Q_C [2]. This is made available to those systems which learn from session-specific data in varying quantities, to allow exploration of the necessary amount of such data — and hence burden placed on a user — for accurate target-session classification.Various systems draw also, or instead, on data from other sources. A number of approaches make use of data collected from the subject N in one or both of the sessions prior to the session-under-test (Session 3), and some learn from data belonging to subjects other than N.

Their sources of data are not all that distinguish systems; the different modelling approaches presented in 7.3 also vary in the ways in which they learn from such data & the aspects of modelling for which they use data of different sources. Fundamentally however, each carries out the same multi-stage learning process with which the reader will be familiar from Chapters 5 & 6. From its modelling data, a system identifies an array of informative features, selects a model configuration and tunes its hyperparameters through Combined Algorithm Selection & Hyperparameter optimisation, and finally trains a model of the determined configuration with which to predict the reserved 33% of the subject N's Session 3 data. Those except Model Transfer systems use the search space defined in Figures 5.37 & 5.38. Model Transfer approaches optimise instead over hyperparameter space in Figure 6.4.

Each system (at each given level Q_C of Session 3 data, where applicable) is trialled for all subjects, with the mean session-specific classification accuracy across the subjects taken as the given system's accuracy (at that level of session-specific data Q_C). As per the methodology of Chapters 5 and 6, approaches are then

[2] This data is labelled Q_C for "Calibration" throughout this chapter including when it is in fact the only data seen by a given system. This is both for consistency and to avoid potential confusion arising from referring to "Subject-specific" data as Q_S in Chapter 6 and "Session-specific" data as Q_S here.

compared on the basis of these scores as evaluated over the Development Set, and the findings of these results subsequently verified through application of relevant systems to the unseen Holdout Set (as described in 5.2.3.2) and thus an evaluation of the extent to which those findings generalise.

7.2.2 Data Splitting & Sampling

The third data collection session was determined as the "target" session on the basis of avoiding temporal data leakage, a flaw in machine learning research which can arise when models are trained offline using data from a later point in time than the data they are asked to make predictions about, giving them an undue insight into the "future" which would not be replicated in a real-world case [191]. Strictly speaking there is little reason to expect such temporal leakage to be problematic in this work. There is not an inherent temporal association in the biosignal data nor the experimental procedure which would plausibly lead to, for example, a subject's Session 3 data to be more predictive of their Session 2 data than the inverse. Rather, temporal leakage has been identified as an issue in biosignal classification at the within-session level. Li et al. [192] noted that where prompts given to participants were grouped in blocks rather than randomised, classifiers learned temporal correlations from EEG data rather than patterns genuinely related to different states of neural activity. The stimuli used in collecting the multimodal hand gesture dataset by Jeong et al. [198] *were* presented in a random order however, & regardless there is no mechanism by which temporal trends in brain state that persisted between sessions and were discriminable with respect to the class (i.e. the physical gesture being performed at a given point in time) could arise. Nevertheless, selecting the chronologically last session as the target is in keeping with convention and certainly best replicates a real-world case wherein a model at the point of gesture prediction could of course not have access to data from a future point in time.

Each data collection session of any subject in Jeong et al.'s dataset [198] contributes 50 performances of each of the three same-hand grasp types, to which are added 50 "rest" gestures as described in 4.2.5 for a total of 200 gestures. In keeping with the 67/33 proportion used for train/test splits elsewhere in the work, here a random 33% of gestures in the session-under-test (Session 3) are reserved for testing to evaluate a system's accuracy, which equates to 66 gestures leaving a total of 134 session-specific gestures remaining to be learned from. Whenever performed, this is a pseudorandom split using *scikit-learn*'s *train_test_split* function, stratified by class. However, as per 4.2 there are four defined gesture classes in the dataset used in this work; because $4 \nmid 66$ and $4 \nmid 134$, the training and testing datasets split in this way will not be exactly balanced. This has the result that the testing dataset of any given trial will contain one "extra" gesture from each of two random classes; they will thus differ very slightly in distribution according to class. This variation will only be 2/66, or approximately 3%, of the dataset[3], and more importantly will not be systematic. The random 67/33 split of the Session 3 data is performed afresh for each given system being tested, at each calibration level Q_C to which the unreserved 67% will be downsampled, and of course for each subject N. It is hence anticipated that by the law of large numbers such minor unevenness in classwise distributions of the reserved testing data will ultimately balance out when aggregated results are considered. Even should there

[3] The class imbalance may be larger, albeit only slightly, in the event the two "extra" gestures are of the same type

be inherent differences in the ease of classifying each gesture type which could cause the individual accuracy of a given set of classifications of a slightly class-imbalanced dataset to misrepresent a system's true accuracy, which itself is not a guarantee, the impact of this on the variation in systems' mean predictive power across subjects and across calibration levels is expected to be very low.

To correct for this oversight and compensate for any risk posed by imbalance issues in the unreserved Session 3 data, in all cases the quantity of session-specific data made available to any given system for modelling is a multiple of 4. The downsampled target-session learning (or "calibration") data Q_C is thus always balanced. This means that in practice the, maximum number of target-session gestures which can be made available to a classification system is 132. The other chosen values trialled for Q_C are primarily linear spacings of 20 gestures (5 of each class): 20, 40, 60, 80, 100, and 120 in total, with an additional value of 72 gestures included to give greater precision to a region indicated as being of potential interest by early provisional experiments. Extreme cases of 1 and 2 gestures per class (thus 4 and 8 target-session gestures in total respectively) were additionally included, to enable a measure of systems' ability to perform with the absolute minimum session-specific data, thus reducing their potential burden on a user as far as physically possible. The range of values for Q_C trialled, and the number of gestures per class to which they correspond, are captured in Table 7.1, and visualised in Figure 7.1.

Gestures per class	1	2	5	10	15	18	20	25	30	33
Total Gestures Q_C	4	8	20	40	60	72	80	100	120	132

Table 7.1: Quantities of Session 3 calibration data

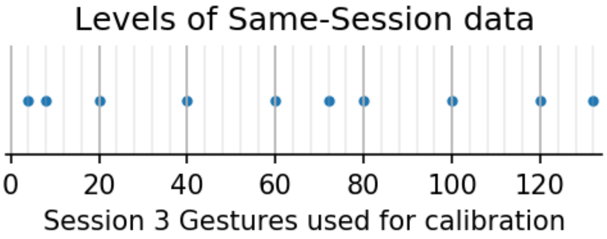


Figure 7.1: Visualisation of degrees of calibration tested

As described further in the explanation of each approach below, in many cases the CASH optimisation procedure's target is typically defined by the accuracy of a candidate configuration's predictions over one third of the total quantity Q_C of target-session data which has been made available for modelling. In the aforementioned extreme cases however where a system has access to only a very low amount of session-specific data, this cannot be split 67-33 while preserving the grouping of data by whole gesture performances to avoid leakage and ensuring each gesture type is represented in either portion of the split.

Where only 8 of the Session 3 gestures in total are made visible to a system, within the CASH optimisation iterations these are split 50–50 rather than 67–33; one unique session-specific performance of each gesture type is used for optimisation-training and one for optimisation-testing.

Where the Same-Session data is reduced to a total of just 4 gestures, only one of each class, these 4 are all used as the optimisation-test data and no target-session data is incorporated into the optimisation-training split. It should be noted that in where the optimisation process includes a transfer learning stage, as in 7.3.3.2, a system would be unable to perform such model calibration with no target-session data in the optimisation-training split, and therefore this value of Q_C is not trialled in such systems.

7.3 Approaches

7.3.1 Baselines

The set of approaches presented here are "zero-calibration", i.e. making use of no session-specific data in any way. A system with knowledge of session-specific data which is no more accurate than one without such data is unlikely to be of any great merit; the burden on a user of providing session-specific training data would not be worthwhile. These uncalibrated systems are hence intended primarily as point of reference by which to contrast other approaches.

7.3.1.1 Out-of-session learning with no adaptation

On the ground of Chapter 6's finding that, provided sufficient subject-specific data is available, there is minimal benefit to including other-subject data, the most pertinent choice of zero-calibration system would be one which makes use of data collected from the subject prior to its intended deployment. This baseline approach thus draws on a subject's out-of-session data for all aspects of learning, to construct a system intended to classify their Session 3 data.

Figure 7.2 presents the structure of such a system, the fundamental sequence of which follows the familiar pattern broadly reminiscent of those in Chapter 6 with a three-stage modelling process: selection of informative features, determination of system configuration via Combined Algorithm Selection and Hyperparameter (CASH) optimisation, and the training of a final model to be evaluated.

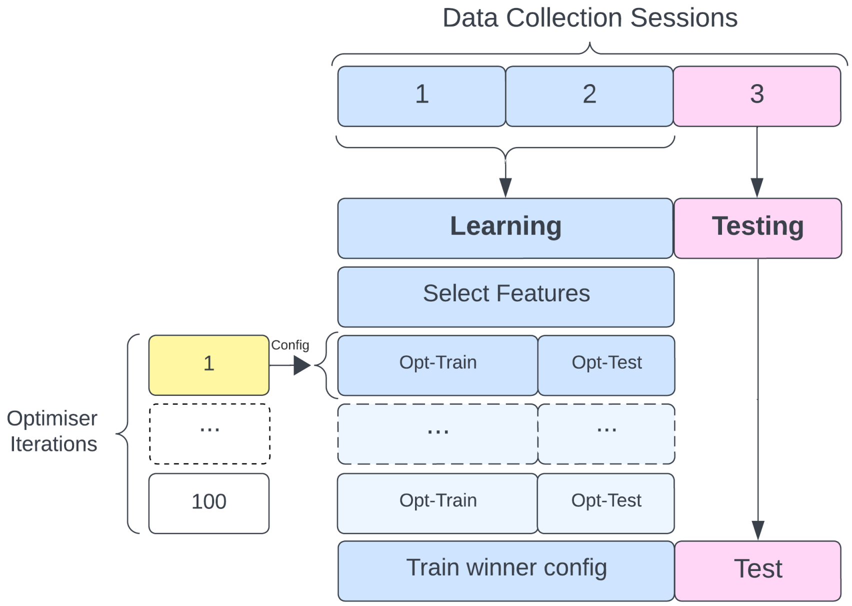


Figure 7.2: Structure of zero-calibration, out-of-session system

Feature Selection is carried out on the basis of all the out-of-session data which is present. The CASH optimisation is then performed on the hyperparameter search space defined in Figures 5.37 and 5.38 for 100 iterations, with a random 33% of the out-of-session data being held for testing in each iteration. This determines the optimisation target; that configuration achieving the highest accuracy is selected as the winner. A system of this winning configuration is subsequently trained on all the present out-of-session data, and used to classify one third of the Session 3 data[4]. This is repeated for each Development Set subject N in turn, and the mean classification accuracy across subjects computed as a measure of that variation of this approach's performance.

Ancillary Investigation: Effects of Time Delay & Data Diversity on Cross-Session Classification In the version of this approach illustrated in Figure 7.2, all of a subject's prior data (Sessions 1 and 2) are used for these purposes. In the dataset from [198] however there are a total of three sessions each of 200 gestures (including the rest class and following balancing, see 4.2.5), each collected a week apart. It is of interest to a system designer to know whether all this data is necessary. Under the framing of the 3rd session as our "deployment" scenario for the system to be used in, training on all of this out-of-session data would require carrying out individual data collection sessions with any new user on two separate occasions, and would incur the corresponding time & monetary costs to both developer and user. If a system could be equivalently performant when based on only one prior recording session, or where these sessions were reduced in size, that would certainly be an attractive property of the system. In other words, it would be useful to understand whether the diversity or quantity of out-of-session data used in a system of this approach is particularly influential.

The temporal separation of the data collection sessions also allows us to consider whether the timing is impactful. Dissimilarity in biosignal data recorded from the same subject in different sessions may be caused by a number of factors. While many of these such as the exact placement and fit of sensors and the level of background electromagnetic activity in an environment are likely to be largely random over time, some such as changes in the subject's physiology and musculature are more plausibly time-variant. Research typically finds such time-correlated variations to be relevant on the order of years or longer [365] — characteristics like the buildup or loss of fatty tissue or muscle mass, which can both affect electromyographic signals measured at the skin's surface (see 2.2.1), do not often see pronounced changes in the course of a week. Shorter-term temporally-dependent drift in biosignals has generally been found as a result of physiological changes at the site of surgically implanted sensors [43, 78, 79]. Nevertheless should there be such time-correlated effects on a system's performance this could affect the longevity of a system, especially one which does not incorporate any session-specific learning, and thus have implications for their design. The presence of two potential sources of training data, each recorded one week apart, provides an opportunity to explore any potential relationship between classification accuracy and the time delay between the recording of training and testing data.

To control for and enable assessment of these effects of data quantity, diversity, and time, four variations

[4] The remainder of the Session 3 data goes unused; for parity with those systems which do learn from Session 3 data these unreserved two thirds are **not** used to test the model.

of the cross-session pretrained baseline system were hence trialled for each subject:

- System has access to all data previously collected from the subject across both prior recording sessions (400 gestures total)
- System has access to all data recorded in the subject's Session 1, two weeks prior to the target session (200 gestures total)
- System has access to all data recorded in the subject's Session 2, one week prior to the target session (200 gestures total)
- System has access to random 50% subsample, stratified by class, of the gestures from each of a given subject's Session 1 and Session 2 data. (200 gestures total, balanced equal contributions from both prior data collection sessions)

7.3.1.2 Generalist (proxy)

The opposing extreme in specificity is a Generalist system; one which is wholly subject-independent in nature. The Generalist as presented in Chapter 5 was for each Development Set subject tested on all of their data across all three recording sessions, and its accuracy reported on this basis. Systems in this chapter, where Session 3 is specifically designated as the target session of interest, are by contrast tested on a reserved random 33% of a subject's Session 3 data as previously described. The accuracies of Chapter 5's Generalist systems are thus determined by their predictive power over a test dataset different in nature to those used in evaluating this chapter's systems; it would not be fair to draw a direct comparison between them.

However neither would it be appropriate to simply re-test the Chapter 5 Generalist system on a random third of each Development Subject's Session 3 data. Section 6.4.1 above discussed extensively the caveats of the Generalist's subject-independence with regard to the Development Set: while for a given subject N the final Generalist system was trained only on non-N subjects, and indeed feature selection performed without any of N's data, the CASH optimisation of the system *configuration* was guided by a loss function averaged over *all* Development Set subjects, and in that way some knowledge of their data was "baked in" to the overall learning process. Using this Generalist to predict a portion of each subject's Session 3 data would thus not be free of cross-session data leakage and not be a wholly like-for-like comparison with the other systems of this chapter. In the context of Development Set results, the "Generalist " results presented should thus be taken only as a proxy.

With regard to the Holdout Set however, as again outlined in 6.4.1 the Generalist system as it is defined in Chapter 5 has at no point had sight of any Holdout data. 5.2.3.2 describes that a Generalist system whose configuration had been selected on the basis of Development Set performance and which was trained, using features selected on the sole basis of Development Set data, on the data of the entire Development Set, was then used to predict data belonging the the Holdout Subjects. There is thus no risk of leakage of subject- or session- specific data and it was possible to re-test the Generalist on a random 33% of each Holdout subject's Session 3 data, providing a consistent performance measure suitable for comparison with the other approaches in this chapter at the Holdout validation stage.

7.3.2 Session-specific learning

The approaches here proposed for systems which *are* tailored in some form to the specific target session in which they're asked to classify data are grouped into two categories.

In this first category are approaches wherein the entirety of a system's direct modelling — that is, the Training stage of the learning process — is done on the basis of data from the target session. The two subtypes presented here differ however in their strategies for undertaking the other stages of learning: defining the configuration of the system & selecting the features which will be used to train it.

7.3.2.1 Within-Session Learning

Under the Within-Session learning scheme, all aspects of modelling draw solely on data from same session as that which the model will be tasked to classify, i.e. a subject's third data collection session. Data from the other two recording sessions data go entirely unused in this approach; a Within-Session system thus allows for a potential "immediate" deployment, wherein no additional data need be collected from a user in advance of the system's intended use session.

The quantity Q_C of a subject's Session 3 data which was made available for modelling is used for Feature Selection, for CASH optimisation, and for Training of the chosen system, as illustrated in Figure 7.3. Within each iteration of the CASH optimisation a random 67% of Q_C used to train a model of the candidate configuration, which is then tested on the remaining 33% — with the error rate of these predictions being defined as the loss function for the optimisation algorithm to minimise. The configuration determined as optimal on this basis is then retrained on the entirety of Q_C, and used to predict the reserved 33% of the Session 3 data. For each value of Q_C, this procedure is trialled across all Development Set subjects, and their mean accuracy taken as the accuracy of this approach at the given level of session-specific data Q_C.

This is as previously discussed perhaps the most common form for a subject-specific classification to take among the literature; the challenges of recruiting participants to attend the lab on multiple repeat occasions evidently precluding many studies from doing otherwise. It is however not ideal for a typical deployment scenario; the requirement for modelling decisions to be made on a per-session basis would not be conducive to use unsupervised by a developer, without the inclusion of a CASH optimisation procedure such as that of this work (or some equivalent process) to the software shipped on a device. Such optimisation would need to both be largely automated with little need for user input beyond the provision of the session-specific data, and sufficiently fast so as to avoid undue inconvenience or delays to a user's ability to operate the system.

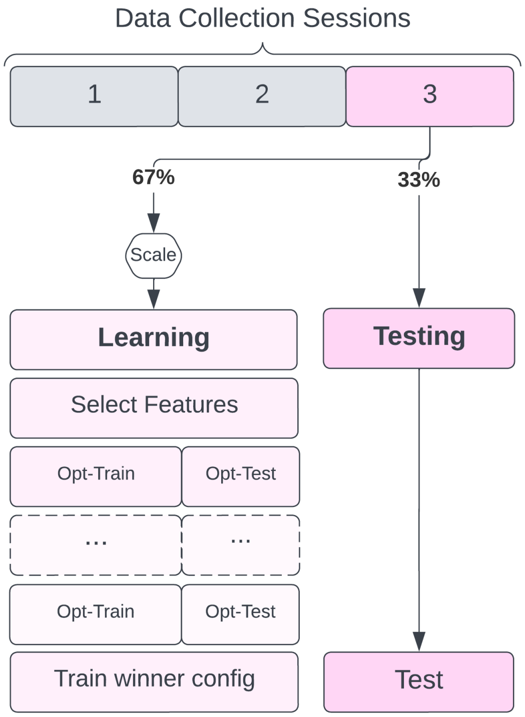


Figure 7.3: Structure of within-session system

On the challenges of within-session optimisation

It should be noted that in this approach, very low levels of session-specific data Q_C present challenges to the CASH optimisation process beyond those discussed in 7.2.2. Not only is there a likely potential for increased susceptibility to overfit, as the “optimise-test” split of the modelling data used as the optimisation target will naturally be very small, but the total quantity of data available to the system is simply too low to be compatible with certain combinations of hyperparameters.

In particular, where $Q_C = 8$ Support Vector Machines are excluded from the optimsation search space. As noted in 7.2.2 at $Q_C = 8$, data is divided 50/50 for the optimise-train and optimise-test splits of each iteration of the CASH optimisation, rather than the usual 67/33, such that both portions contain a single gesture of each class. The strategy for coercing probabilistic predictions from SVMs outlined in 5.3.3.6 however relies on a five-fold validation process; while this split, unlike most in the work, is on the basis of individual time-windowed instances (see 4.3) in the dataset, rather than on whole gesture performances, a single performance of each gesture does not contribute sufficient total datapoints to divide into five while ensuring all classes are represented in each split.

Additionally, at this level the upper bound of the kNN’s hyperparameter k is reduced to 10 (from its usual 25 as per Figure 5.38). While there are indeed be more than 10 datapoints in the training set, the size of the neighbourhood is so reduced to enable evaluation of metamodel-based Decision-Level Fusion schemes during the CASH optimisation. As outlined in 5.3.1.3 such metamodels are trained on predictions of lower-level EMG & EEG classifiers generated through three-fold cross evaluation. If the CASH optimisation process trials a metamodel fusion algorithm in combination with a kNN EMG model[5], the temporary kNNs set up to enable metamodel training could have a theoretical maximum neighbourhood size of 11 (4 total gestures in the optimise-train split, each contributing 4 instances for a total of 16 datapoints, two-thirds of which as per the three-fold cross-evaluation is $10.67 \approx 11$ instances for training the EMG-kNN).

Additionally, this approach is simply unviable at $Q_C = 4$. A training set comprising only four gestures is insufficient to be able to split, in keeping with the practice of splitting on the basis of whole gestures, into “optimise-train” and “optimise-test” subsets during CASH optimisation while ensuring each split contained data belonging to all possible classes.

7.3.2.2 Within-Session Learning with Ported Configuration (from prior data)

These limitations on the capabilities of the CASH optimisation routine which arise in a wholly session-specific system when attempting to minimise the required level of session-specific training data Q_C motivate efforts to find alternative means by which to determine system configuration.

The “Ported” Within-Session strategy draws on a subject’s out-of-session data for this purpose. Where the previously outlined within-session approach had optimised for target-session accuracy, this approach instead explores the supposition that a configuration which performs well when trained and tested on a given individual’s data at one point in time may well be suited to learning from their data at another point in time, even though the final trained model is not itself the same across those domains. It identifies a single configuration using certain subject-specific data & views this as “portable” to other data belonging to the same subject. This can be seen as a parallel to the subject-agnostic nature of the configuration of the type of “Bespoke” systems seen in Chapter 5 systems, but considering the extent of the data universe as data provided by a specific individual, and the unit of generalisation as the recording sessions in which those data were gathered, rather than looking across subjects.

Out-of-session data belonging to a subject is used for CASH optimisation to determine the system’s configuration. As seen in Figure 7.4 this data is also used for the Feature Selection stage; that same array of attributes is later taken from the session-specific data Q_C to ensure the resultant system is trained on the same features as its configuration was optimised using. It would in principle be possible to reverse this: to find the most informative features of the session-specific data Q_C and reduce the out-of-session data to those features prior to CASH optimisation, potentially allowing for greater specialisation to the target session. However this would hinder the portability of the configuration & require a discrete optimisation routine to be performed for each novel usage session, undermining one of the key potential benefits of the approach to usability. The selection of features on a dataset larger than those which would be seen at very low quantities of Q_C may also mitigate somewhat the risk of overfit presented by learning from a very small dataset.

At each stage of the CASH optimisation, a random 67% of the out-of-session data, stratified by class, is

[5] The kNN has already been excluded for EEG classification, as per 5.5.8.

used to train a system of the candidate configuration, and that system's accuracy in predicting the remaining third is used at the quality metric for the optimiser. The winning configuration is the one which maximised this accuracy. This winning configuration is then used on a within-session basis: copies of the system are trained on varying amounts of session-specific data Q_C, & each used to predict the reserved 33% of Session 3 data. In this way the approach allows the training stage itself to be session-specific & is able to reap any potential benefits of that, while addressing the challenges to attempting CASH optimisation at reduces quantities of provided target-session data. The mean accuracy of these predictions across all Development Set subjects for each possible value of Q_C is taken as the approach's accuracy at that level.

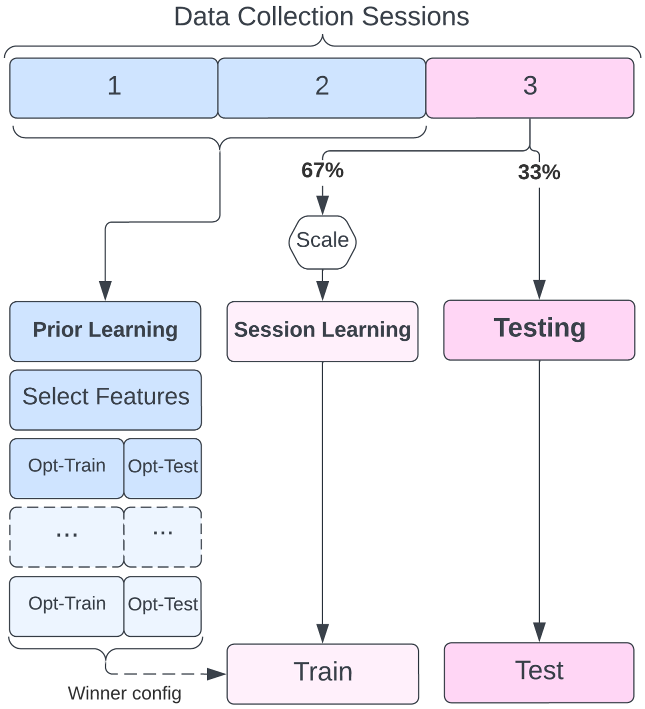


Figure 7.4: Structure of Within-Session system with "ported" configuration

Potential effects of the diversity & quantity of the out-of-session data used for feature selection and CASH optimisation were explored by trialling the following three strategies for sourcing this data:

- All data from both of the subject's prior sessions;
- All data from a single one of the subject's prior sessions[6];
- A stratified 50% subsample of each of the subject's prior sessions, such that the total out-of-session data was the equivalent of a single recording session (200 gestures).

[6] Specifically Session 2, on the basis of it on average providing greater uncalibrated cross-session predictivity over a subject's Session 3 data than their Session 1 data did, as covered in 7.4.1 below.

Contrasting the first and last of these options allows investigation of the impact of data quantity, and comparison of the latter two options enables investigating the benefit of undertaking both data collection sessions while controlling for quantity.

In addition, it was considered that there may in fact be a benefit to including target-session data in the Feature Selection & CASH Optimisation stages over carrying these out with other data, but that this benefit could be overshadowed in 7.3.2.1 by the potential insufficiency of optimisation data where Q_C is low. Therefore a fourth strategy was trialled wherein the session-specific data was "topped up": at each level of Q_C, a stratified random subsample of the subject's out-of-session data were added of a size inversely proportionate to Q_C. This meant that there were a consistent total of 200 gestures with which to perform Feature Selection and CASH optimisation, with Q_C controlling the proportion of those data which belonged to Session 3. The winner configuration was then retrained on solely Q_C for a more equivalent comparison to the other aforementioned variations of this approach (and to maintain a distinction in nature from the approach presented below in 7.3.3.1). The direct modelling was still on the basis of target-session data and the optimiser was also able to learn from it, but the problems of insufficient data with which to optimise encountered in 7.3.2.1 were thus addressed.

7.3.3 Cross-session learning

This second category of the two covers approaches wherein both session-specific and out-of-session data are directly used in the actual modelling of a classifier, and are thus referred to as incorporating "cross-session" learning. These two data sources are drawn on in a variety of ways, including strategies which parallel those previously used to merge same- and other- *subject* data in the experiments of Chapter 6.

7.3.3.1 Augmentation

The Augmentation approach, much akin to that seen in in Chapter 6 (6.3.2), simply uses data collected from the subject prior to their third session to supplement the dataset and thus allow a system to model on a greater quantity of data. In this way it relies on the expectation that a subject's out-of-session data will bear sufficient similarity to that of the target session so as to be of use in modelling.

Here the downsampled target-session data available for modelling, Q_C, is augmented with the entirety of the data collected from the subject prior in Session 1 & 2, and this joint cross-session dataset is then used for all stages of the learning process. Informative EMG & EEG features are first selected from the merged data in accordance with the Feature Selection strategies outlined in 5.3.2.2. For each iteration in the CASH optimisation routine, one third of the Session 3 data present within the merged set is held for testing. Correspondingly, one third of the Sessions 1 & 2 data are temporarily discarded for that iteration, to preserve the relative proportions of session-specific and out-of-session data within the optimise-train split and the overall learning dataset[7]. The iteration's candidate model (constructed according to its point in the hyperparameter search space) is hence trained on the remaining two-thirds of the merged dataset, and then used to make predictions over the held third of the merged set's Session 3 data. Of the 100 optimisation

[7] see 6.3.2

iterations, the combination of models and hyperparameters which made those predictions most accurately is taken as the winner; thus the determined system configuration is optimised for maximising classification accuracy of Session 3 data.

A system of this winning configuration is then retrained on the entirety of the cross-session learning dataset, and is finally used to predict the reserved 33% of the subject's Session 3 data. The accuracy of these predictions is computed; the mean accuracy across the 20 Development Set subjects is taken as the Augmentation approach's accuracy at a given level of session-specific data Q_C. Figure 7.5 presents a visual outline of this approach.

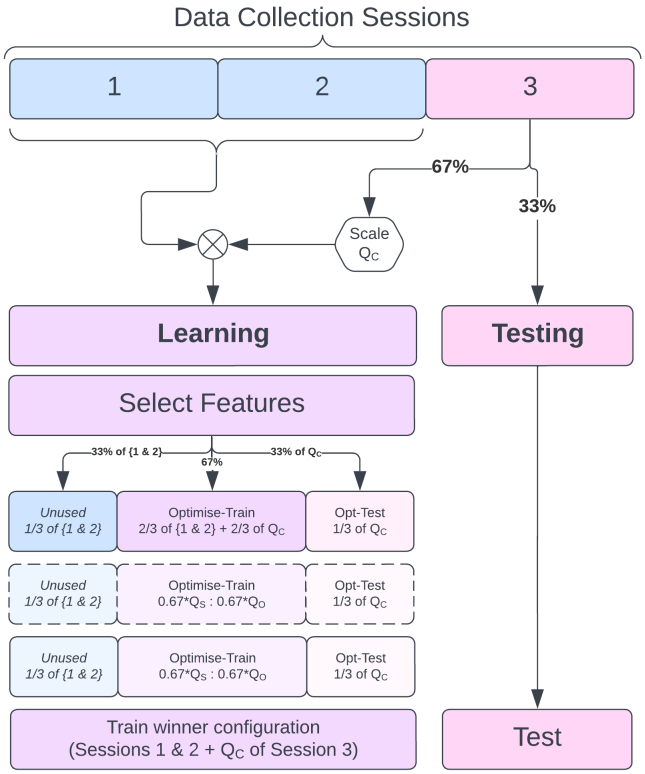


Figure 7.5: Structure of augmentation system

7.3.3.2 Model Transfer

Under the Model Transfer strategy, the training of a system is itself a two-step process: models are initially fit to data collected from a subject in sessions other than the target Session 3, and subsequently adapted to the portion of their Session 3 data which was provided to the system Q_C. This again is somewhat similar in nature to Chapter 6's Model Transfer strategy described in 6.3.3, with the scope of the data universe reduced to that of a single individual. Much as the Augmentation above does, this approach operates on the basis that a subject's out-of-session data may be assumed to provide some level of class-discriminable information to a machine learning model, but that a discrete process of specialisation on data from the session-under-test will enable models' refinement and lead to higher classification accuracy. By forcing a model to pay deliberate attention to the target-session data, this strategy may avoid such session-specific information being deprioritised which may be a particular risk in, for instance, an augmentation system with significant imbalance between session-specific and out-of-session data.

As depicted in Figure 7.6, here features were again selected from the join of a subject's out-of-session data and the amount of their Session 3 data Q_C made available to the system for calibration. The system's configuration was determined through exploration of the hyperparameter search space suitable for direct model transfer applications defined in Figure 6.4. At each CASH optimisation iteration, the configured model was initially trained on solely the out-of-session data. Two-thirds of the session-specific data present Q_C were then used to adapt this trained model to the target session[8], and the adapted model was used to predict the remaining third of Q_C, with the accuracy of those predictions defining the optimisation algorithm's loss function.

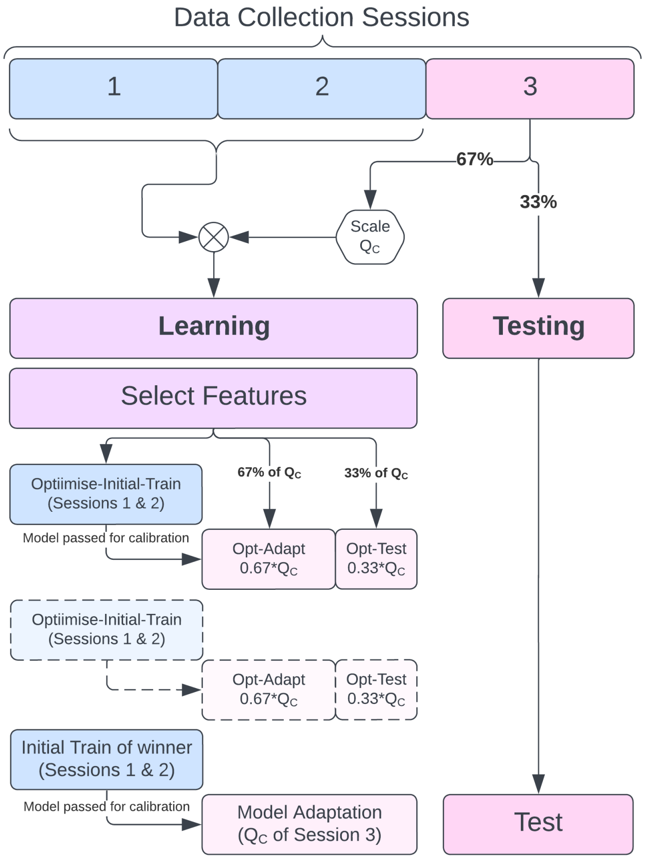


Figure 7.6: Structure of model transfer system including session-specific optimisation. Note parallels to Fig.6.3

[8] See 6.3.3 for a description of the mechanisms of domain adaptation of each eligible classification algorithm.

A system of the “winner” hyperparameter configuration which minimised the error rate in its predictions of this 33% of Q_C was then trained afresh on the out-of-session data, and adapted using this time all of Q_C. This calibrated model’s accuracy in predicting the reserved 33% of the subject’s Session 3 data was taken as the system’s accuracy for subject N at the given quantity for Q_C. The mean of these accuracies over Development Set subjects was determined, as previously described, for each possible value of Q_C defined in 7.2.2 to measure the performance of this approach.

7.3.3.3 Model Transfer with static configuration ported from prior user data

As discussed above, in 7.3.2.1 and elsewhere, the inclusion of a Combined Algorithm Selection & Hyperparameter optimisation stage to determine a suitable configuration on a per-session basis may present distinct inconveniences in a deployed system.

This variation on the Model Transfer strategy therefore, by contrast to 7.3.3.2, performs model adaptation following a single, one-time optimisation routine the outcome of which is applied across all levels of calibration data Q_C. While the dataset used in this work contains data of only three sessions for each subject, the principle of this approach is that the resultant chosen configuration would continue to be ported to any further sessions, without additional session-inclusive optimisation. If achievable this could suggest a much streamlined experience for the end user of a system, and a more lightweight software it. Rather than requiring a system’s component models and their hyperparameters to be determined uniquely for each given session, here these stages might be for example a part of the pre-deployment “setup” of a device — a customisation process carried out by the supplier & prosthesetist for each novel user before it is put into use — and yet the system would be able to benefit from bespoke tailoring to each new session through adaptation of the pretrained model to some amount (Q_C) of session-specific calibration data.

Much as outlined in 7.3.2.2, which applies the same notion of portability to a within-session system, this operates on the expectation that the suitability of a given system configuration is a property transferable, at least in some large part, between different sessions of a given user’s data — and thus that sufficiently appropriate models & hyperparameter values can be selected on the basis of data collected outside the target session.

Additionally it should be noted that the approach in 7.3.3.2 optimises, in principle, for a configuration well-suited to undergoing model adaptation. Here no such efforts are made. The system instead operates on the belief that a model which performs well in classifying data in the source domain, and is capable of undergoing domain adaptation, will be good at that domain adaptation — and will thus perform well in classifying data in the target domain following the transfer learning. While this is very likely a naïve assumption it is necessary to enable the elimination of session-specific CASH optimisation from the process. The divergence between the metric is being optimised for (performance of an unadapted model on data of the same session(s) upon which it was trained), and that which the system is being used for (performance of an adapted model on data of the session to which it was adapted) is certainly a potential weakness of the technique and one which may be interesting to investigate the impact of.

As can be seen in Figure 7.7, under this approach features are selected on the sole basis of the subject’s

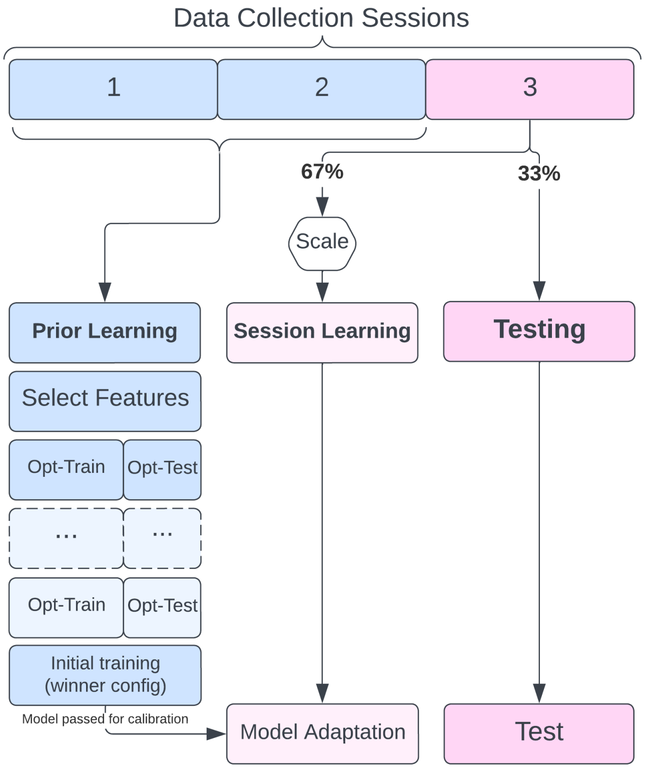


Figure 7.7: Structure of Model Transfer system with "ported" configuration

Session 1 and Session 2 data. The same are then used for performing the CASH optimisation, with a stratified random 33% of this out-of-session data used to evaluate the quality of the prospective hyperparameter configuration at each of the 100 optimiser iterations, and that which provided the greatest accuracy being considered the "winner". Both feature selection & CASH optimisation are hence each performed only once.

Varying levels of calibration data Q_C from the session-under-test are then trialled in turn. A system of the chosen configuration is trained initially on all the out-of-session data, then adapted using the session-specific Q_C, and finally used to predict the reserved 33% of the subject's Session 3 data. As with those approaches outlined previously, this is repeated for each subject N in turn at every defined quantity of calibration data presented in 7.2.2, and the mean accuracy of these predictions over all Development Set subjects is taken as this approach's accuracy for the given level of Q_C.

This approach's avoidance of a session-specific optimisation stage in conjunction with its ability to adapt a previously trained model not only allows for smoother use but sets up the possibility of an "online" or "adaptive" calibration strategy. While such a strategy was not explored here experimentally, an approach of this nature could be extended to give a user some input interface with which to signal that the system's previous classification attempt was incorrect, and thus trigger it to further adapt its classification algorithms "on-the-fly" on the basis of that residual, not altogether dissimilar to the principles motivating Gradient Boosting algorithms. The adaptation mechanism of the Gaussian Naïve Bayes classifier as implemented in *scikit-learn*, described in 6.3.3, lends itself particularly well to such incremental training. Such a method could perhaps even be automated: Förster et al. [115] had success in adaptive correction of an EMG-based classifier by using EEG measurements to detect a pattern of brain activity known as the error-related potential ($ErrP$) which presents when an individual recognises an erroneous or unexpected response to a situation, and Seeland et al. [295] were able to incrementally retrain an EEG classifier by repeatedly introducing "buffered" EEG data which were labelled according to the cotemporaneous predictions of an EMG classifier.

7.3.3.4 Model Transfer from Generalist

In the Model Transfer approaches above, given that each system is adapted to data from the specific target session, it is essentially expected that it primarily learns "coarse" properties from the source domain data, with the "fine" patterns derived from session-specific data. While this is a simplified framing, an interesting question can be found in extending this logic further — by exploring whether that initial learning could be adequately done using data not only of a different session, but which is not even specific to the subject N under test.

While Chapter 6 did not find incorporation of other-subject data to be of aid to a system which had access to sufficient subject-specific data, Chapter 5's findings in the success of both its Generalist and its "portable" Bespoke approaches suggest there to be much learning which could be generalised across individuals to at least a certain degree, even if less beneficial than subject-specific learning as Chapter 6 would imply.

There is certainly a clear motivation for exploring this if it were to prove viable. Eliminating the need for the user-specific setup of a system prior to its deployment could have disproportionately great accessibility implications than simply the time and cost reduction. The impact of such factors in terms of barriers to access can perhaps be best thought of as them imposing a high "activation energy" , in that a system with no pre-deployment requirements will be vastly more accessible than one with minimal. Even the shortest imaginable data collection session, for example, requires the availability of suitable space, facilities, labour, and expertise to carry it out.

In this strategy therefore the Feature Selection, CASH optimisation, and initial (or "cold") training of the system's accordingly selected model & hyperparameter configuration is carried out using data collected exclusively from subjects other than N.

This other-subject data, for pragmatic reasons[9] and for greater equivalence with the other systems of this chapter, was downsampled to be of equivalent quantity to the amount of out-of-session subject-specific data which would be available to other cross-session learning systems, and which this approach avoids the need to collect. This totals 400 gestures (200 from a subject's Session 1 and 200 from their Session 2), which are correspondingly taken from the pooled Sessions 1 and 2 of the non-N subjects. This downsampling was stratified by class, and as far as was possible by participant such that a similar quantity of data was

[9] Primarily training & testing speed, but see also e.g. 5.3.3.6 & 6.5.1 on the impact of using datasets too large for certain algorithm choices to be viable options.

contributed by each non-N subject. Where this approach was applied for tests with Holdout Set subjects, all 20 Development Set members were available as non-N subjects to draw from, and so an exact balance was possible (as $400 \mid 20$). In the experimentation with the Development Set however, where there were 19 subjects not-under-test at any given time, there was a slight imbalance of contributions. This imbalance is assumed to be random.

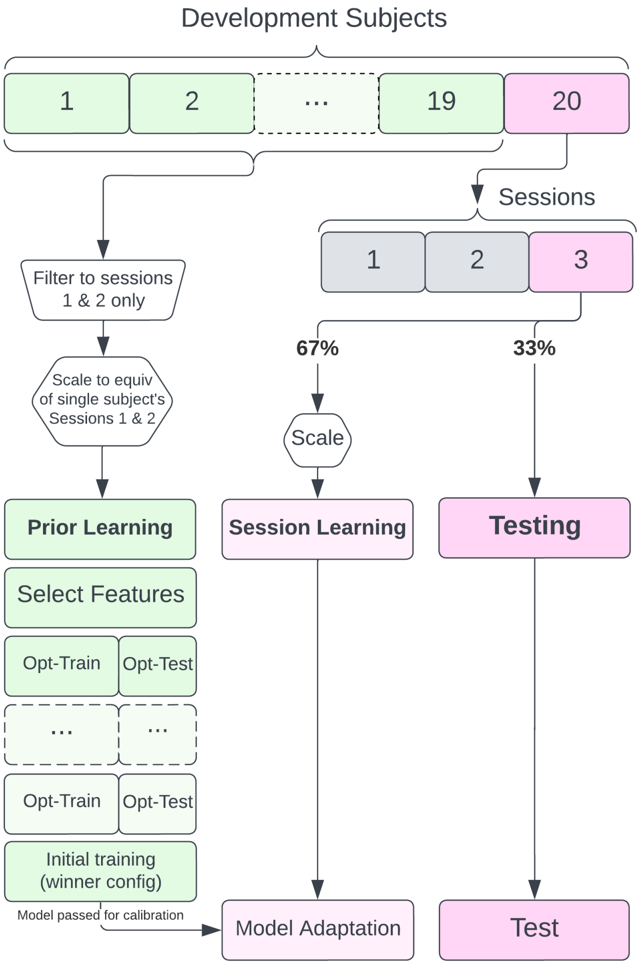


Figure 7.8: Structure of a Transfer From Generalist system. NB depicted here is the system when used with a Development Set subject thus leaving 19 "other" subjects available.

As depicted in Figure 7.8, all of these 400 other-subject gestures were used for Feature Selection. In each iteration of the CASH optimisation process, a random 33% of them were used as the test data by which the optimisation target was determined. It should be highlighted here that this random splitting into optimise-train and optimise-test was performed across all of the downsampled other-subject data; the optimisation target was not determined by any separation of the data belonging to different non-N subjects from one another. Thus while as shorthand this approach is referred to as transferring "from a Generalist", the hyperparameter configuration was not itself optimised on the basis of cross-subject classification performances as it was in the "Generalist" systems elsewhere in the work. Akin to the strategy described in 7.3.3.2 above, this optimisation was done only once for a given subject (though as noted, on the basis of data which did **not** include their own). A system of the determined configuration is then, for each value of Q_C, trained on the 400 other-subject gestures with the array of informative features found from them. The trained model is subsequently adapted to the subject-under-test's Session 3 using Q_C, and used to classify the reserved 33% of the subject's Session 3 data. Again, the mean classification accuracy over all Development Set subjects for each defined quantity of calibration data Q_C was computed to determine the accuracy of this approach at that level of calibration.

7.4 Results

The performances of these various approaches are assessed and comparisons drawn between them. Initially, the Development Set data is used to identify the most promising & interesting candidates from among the strategies outlined. Observations made of these systems' performances when provided with varying quantities of calibration data Q_C are subsequently tested by their application to the Holdout dataset.

It should be recalled here that as noted above (7.1), this part of the work was exploratory in nature and the investigation of some of the systems here motivated the inclusion of variations thereof which were not initially planned. Certain parts of these results are thus presented in a fashion not strictly linear in terms of the chronology of the experimentation. For example, 7.4.1 finds Session 2 to provide superior uncalibrated cross-session classification ability than Session 1, and this result informs the choice to use Session 2 in any subsequent systems which sought to draw on only one of a subject's non-target data recording sessions.

Any such iterative, result-informed *a posteriori* design choices occurred only at the exploration stage of these experiments with the Development Set. Any and all verification of trends and results with Holdout subjects, and the determination of particular findings to assess in this way, were done subsequently. The design of the experiments and tests discussed in 7.4.3 was in no way shaped by any advance knowledge of Holdout Set performance. As has been discussed in 5.2.1 and elsewhere throughout the work, care has been taken to avoid the pitfalls common to much biosignal research [30,152], and indeed machine learning research more broadly [191] of introducing bias or leakage to the selection of systems, comparisons between them, or any other stage of experimentation.

7.4.1 Uncalibrated cross-session baselines

Aim 7.1 of this chapter was to "*Investigate the impacts of data diversity and source–target time delay on uncalibrated cross-session biosignal gesture classification*". As described in 7.3.1.1 above, four variant strategies for developing a pre-trained, zero-calibration system were investigated, defined by the data to which they had access: all of a subject's "Session 1" data, all of their "Session 2" data, all data collected across both these sessions, and a downsampled set containing a random 50% of their data from each recording session. Figure 7.9 presents the classification accuracies achieved by these variations of the pre-trained baseline on the 20 Development Set subjects.

7.4.1.1 Impact of data diversity & quantity

It can be immediately seen that systems with access to both sessions' data achieved greater accuracies than those modelled on only one. That the "downsampled" system had a higher mean and median classification accuracy across subjects than either of the single-session approaches suggests this effect cannot be wholly ascribed to an increase in the total data seen by a system. Rather the presence of this effect after controlling for data quantity suggests that systems did indeed benefit from having a more diverse training dataset. Such diversity is likely to have led models to identify trends among the biosignal data which generalised across both recording sessions and were thus more likely to remain robustly generalisable to a third session, where

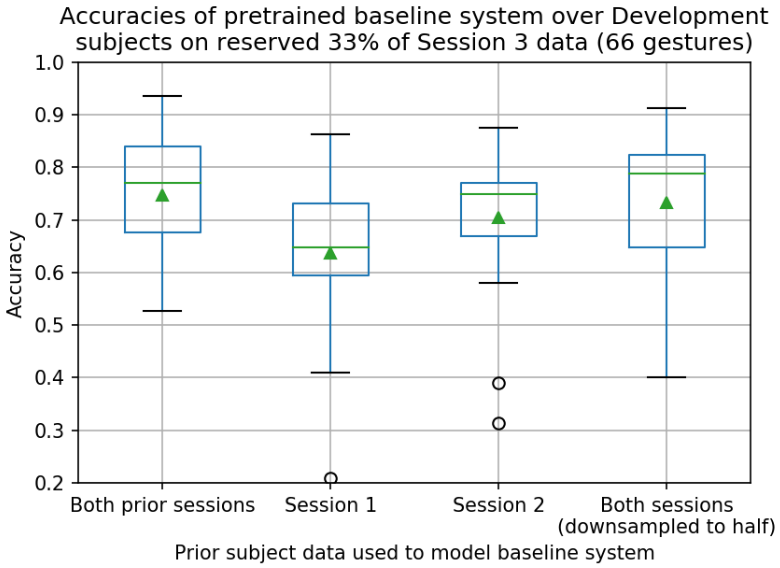


Figure 7.9: Comparison of pre-trained baseline systems with access to no session-specific data

by contrast a single-session model may have been more susceptible to overfit.

The impact of the overall data quantity appears less pronounced; comparing the baseline with access to all prior user data and that modelled on only half of it, the two reach similar mean accuracies and the latter's median accuracy is marginally greater. The all-prior-data baseline however was much more accurate for its weakest subject than the downsampled system, at 52.65% to 40.15% respectively. This increased robustness motivates its adoption as the *de facto* out-of-session pre-trained baseline system from hereon.

7.4.1.2 Impact of time delay

Interestingly, on average data from a subject's second recording session was a distinctly better predictor of their third session than data from their first session was. While impossible to verify with certainty, this is believed to be a random effect. In the longer term, time-correlated changes in the physiology of an individual may result in a "drift" in the properties of their biosignal data. As an individual ages for example their muscle tone and fat distribution among other factors, which can as noted in 2.2.1 affect electromyographic signals, will naturally change, as does the underlying behaviour of their body's Motor Units [365]. It would certainly be plausible for this to mean an individual's biosignal data on a given day was more closely correlated with that from the previous day than from the previous year (as an illustrative example). A subject's three recording sessions in this dataset however were spaced apart by only a week each. Even considering factors such as health and exercise level which could well lead to more rapid physiological changes than ageing, a time differential of a single week between Sessions 1 and 2 seems unlikely to lead to such notably weakened predictive power in the more distant session. Where a notable shift in the nature of recorded muscular biosignals has been observed over the course of just a week, this has been due to degradation in the quality of implanted sensors as a result of their prolonged presence in the body, rather than any fundamental change to the underlying bioelectrical activity [43].

Other temporal effects are perhaps more plausible, though none appear particularly compelling in the absence of further evidence. Modifications to the experimental procedure, measurement equipment, or environment of the data collection sessions could certainly be reasoned to affect the nature of the bioelectric signals recorded. While Jeong et al. [198] report no such deliberate changes it cannot be wholly discounted that conditions such as the level of background noise from ambient electronic equipment in the vicinity could have changed over time in ways which escaped the researchers' notice. It should be considered however that Jeong et al. do not imply the one-week time delay between sessions to be anything other than relative with respect to each individual subject. It is not stated that all subjects' first recording sessions were carried out in the same week and all completed a week before any subject's second; given the time and labour required for data collection & the total number of participants this seems ultimately unlikely. Were there any inadvertent changes to the experimental conditions over time these would almost certainly have been correlated with actual time and not occurred between each separate subject's second and third data collection sessions individually.

Potentially more plausible is the possibility of a gradual change in subjects' behaviour. Whilst the impact of user's progressively increasing experience with Brain-Computer Interface usage is more likely to have been relevant to the Kinaesthetic Motor Imagery data also recorded by Jeong et al. than the genuine Motor Execution data used here, what is perhaps possible is an experiential effect in terms of the physical gestures being carried out. Through repetition and familiarity, it could be posited that subjects' performances of the three grasping gestures became more consistent with regard to the hand shape during the grasp and the specific motions performed to reach it. This would hence result in less variable patterns of bioelectric activity which could lead to the observed greater forwards generalisation performance of the Session 2 data. That Jeong et al.'s paper makes no mention of an observable "learning curve" in participants' gesture performances certainly indicates this to be relatively unlikely to have occurred but it remains possible. Whether the apparent time-related increase in predictive power of the subjects' biosignal data is a non-random effect would not be possible to ascertain without data from a greater number of successive recording sessions, but if it were, then an increasing consistency over time in the precise hand shapes being performed seems the most plausible explanation.

There would be great merit in the further assessment of time-variant effects on systems' accuracy in future work. The challenges of recruitment & retention of participants in longitudinal studies means that within-subject, cross-session biosignal data is scarce, particularly in the order of months or years as would perhaps be most valuable in distinguishing physiological, environmental, behavioural, and other potential effects. Such work could however be of significant benefit to informing the design of cross-session or session-agnostic gesture recognition systems intended to be deployed for long-term use, and in quantifying or even reducing the extent of the need for continued maintenance.

7.4.2 Comparing session-targeted approaches

The mean classification accuracy across the 20 Development Set subjects of all the approaches trialled which have access to some level of session-specific data Q_C — those outlined in 7.3.2 and 7.3.3 — are presented in Figure 7.10. Additionally included as a reference are the out-of-session baseline (according to the conclusions reached in 7.4.1 above, only the variation with access to all of a subject's prior data is included here) and the Generalist system of Chapter 5, which as discussed above in 7.3.1.2 should be viewed only as a proxy measure with respect to Development Subjects.

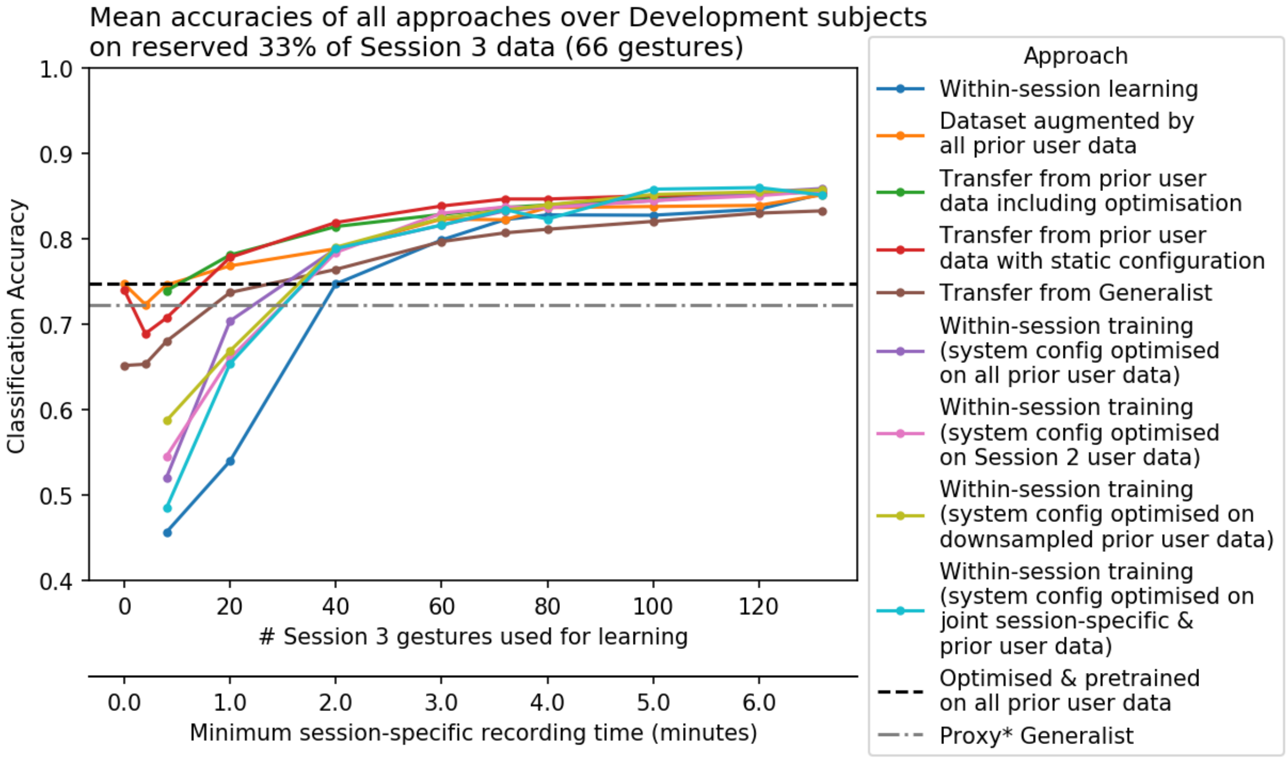


Figure 7.10: Mean Development Set accuracies of all approaches at increasing levels of session-specific data Q_C
NB Figures 7.11 & 7.12 separate these results according to their modelling approach.

Visible also in Figure 7.10 and others hereafter is an additional axis representing the quantity of Session 3 data used Q_C in units of time, with each discrete gesture performance corresponding to 3 seconds. Three seconds is the duration of each gesture performance extracted from the recorded data as outlined in Chapter 4; in actuality the participants of Jeong et al.'s study performed gestures for approximately 4 seconds each [198]. In addition the calibration routine of a real-world system may need to incorporate breaks between gesture inputs, and account for the possibility of an incorrect or mis-timed gesture from the user potentially requiring some repetition or correction. The time spent learning from this calibration data, whether that involves only model training or adaptation or includes the process of determining a system's configuration & tuning its hyperparameters, is also not captured here. This will not only vary between systems and their variations, but is likely to also be a function of the total quantity of data provided to the system. Estimates of training time taken from the offline experiments performed here would not be a good measure of the time required for such modelling by a deployed system with different hardware limitations, and the other potential sources of time delay mentioned are not quantifiable within the scope of this work. Rather than attempt unfounded estimations of such factors, the “time” axis is thus presented simply as the total duration of the session-specific data Q_C used, giving an indicative visualisation of only an theoretical absolute minimum required calibration time which in practice would likely be much greater.

The many approaches and versions thereof seen here differ in numerous ways not only on the means by which they learn from data but on the data to which they have access. To simply identify the most accurate system and present it as a singular universal recommendation would be a naïve comparison of their properties and performances. Rather, approaches of the two categories previously defined, those wherein model training is session-specific (7.3.2) & those in which it is not (7.3.3), are explored below in turn, with a view to refining this collection of strategies into a selection of viable “candidates” which differ in interesting, application-relevant ways, from which observations can be made and subsequently verified using the unseen Holdout dataset.

7.4.2.1 Session-Specific Training

Firstly those systems described in 7.3.2 which are trained solely on target-session data are assessed. Figure 7.11 displays the classification accuracies of these systems averaged over the 20 Development Subjects, at the various quantities of Session 3 data Q_C on which they were trialled.

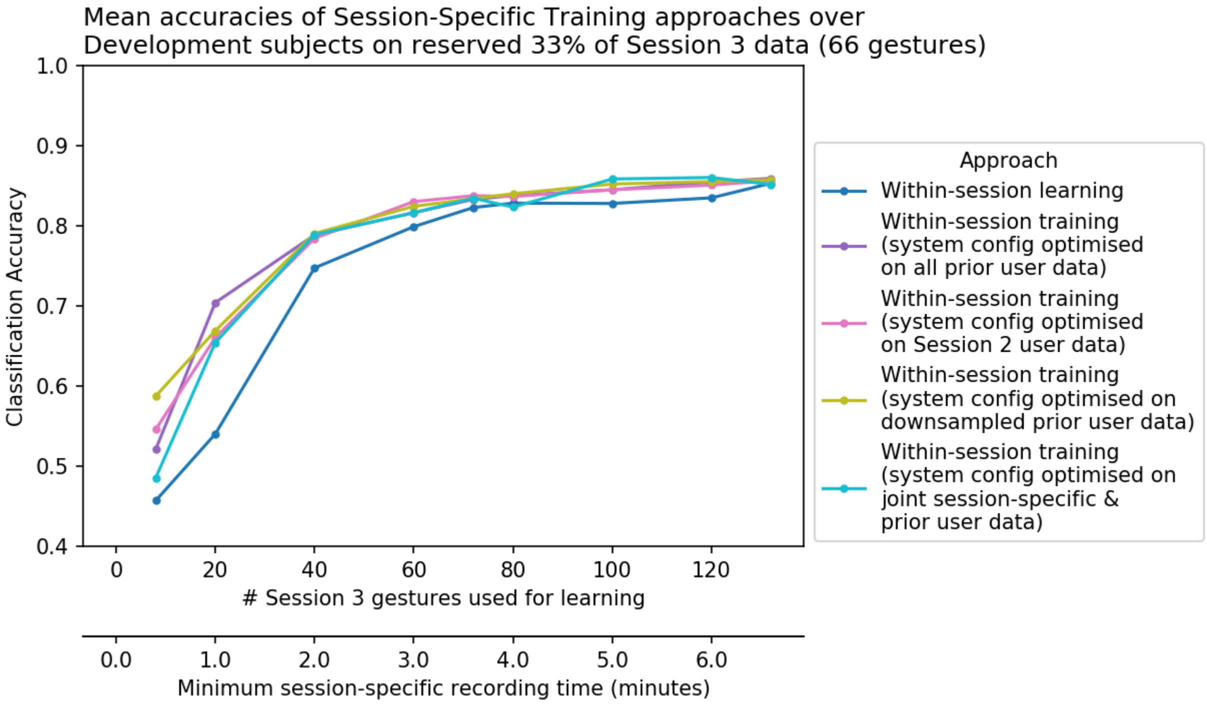


Figure 7.11: Mean Development Set performance of approaches wherein model training is based on session-specific data

Immediately visible is an almost logarithmic trend, consistent across all strategies, in their performance with respect to Q_C. While the benefit of additional session-specific data appears to begin rolling off, in many cases after approximately 80 total gestures (20 gestures of each class), in every approach there is a sharp decay in accuracy as the level of Session 3 data declines much below this point as the models most likely overfit to the few datapoints they have to learn from.

It is perhaps unsurprising that this presents most dramatically in the wholly within-session approach, and that this system consistently trends among the weakest overall. Under this particular strategy, wherein the classification system's modelling choices are determined bespoke using only the target-session data available, a severely reduced Q_C will be unlikely to provide sufficient data for a robust CASH optimisation — the chosen configuration is itself liable to be overfit to the specific Session 3 data used. Despite this underperformance, the within-session approach is, with sufficient data, seemingly able to reach predictive capabilities broadly commensurate with those of its peers. For this reason, and as the only strategy not reliant on any other sources of biosignal data, it is retained as a candidate of interest.

Those techniques outlined in 7.3.2.2 which avoid this pitfall by drawing on other data to determine their modelling choices all achieve remarkably similar mean performance across the various levels of Q_C upon which they are trained. That these largely exceed the accuracy reached by the pretrained baselines of 7.4.1 (provided sufficient training data), and that the variation which incorporates Q_C into the CASH optimisation procedure, suggests similarity between source and target domains of training data to be of greater importance here than that of the data used to determine a system's configuration. There seems no consistent benefit of including session-specific data at the optimisation stage and so this technique can be "disqualified" from the consideration of candidate systems.

This "elimination" has the additional benefit of discounting the second of the two approaches seen here which rely on a per-session optimisation of modelling hyperparameters; the remaining three to be discussed are alike in their use of a "static", predetermined system configuration and differ only in the data with which it is found. While as noted the distinction between these systems' scores is minimal, the variation which draws solely on Session 2 data for CASH optimisation trends marginally weaker. The prior exploration of out-of-session performance suggests a benefit to learning from diverse data, and while not testable with these data it stands that such diversity could be plausibly anticipated to lead models to generalise better to further novel sessions of a given subject. This is thus also disregarded as a potential candidate. Unlike in the case of the wholly out-of-session baselines however, here the reduction in quantity of joint Sessions 1 & 2 data to 200 total gestures does not appear to notably degrade performance in the main. The system drawing on all of a subject's prior data for optimisation did offer greater accuracy than its downsampled counterpart where $Q_C = 20$, but at such low levels of session-specific data neither variation exceeded the accuracy attained by the out-of-session baseline, indicating that with such low session-specific data the prior data is put to better use by being used for all stages of learning, making the superiority of the all-data approach moot and thus motivating its elimination in favour of the similarly performant strategy which requires only half the out-of-session data.

7.4.2.2 Cross-Session Learning

Subsequently explored are the approaches involving cross-session model training described in 7.3.3. The mean accuracies of these systems at varying levels of session-specific data Q_C are shown in Figure 7.12.

Again it can be seen that the systems of this category present relatively similar trends to one another in their response to increasing quantities of session-specific data. While these approaches do diminish in classification ability at very low levels of Q_C, comparison to Figure 7.11 highlights that this rolloff is much less steep than for those systems discussed above which relied solely on Q_C for training. That the cross-session systems, at such low quantities of calibration data, tend to outperform systems trained exclusively on target-session data regardless of whether the latter's configuration were optimised using that small dataset Q_C or using a subject's out-of-session data, suggests that both the training and optimisation stages contributed to session-specific systems' overfit in these cases. Both of these elements of the learning process were undermined by the low quantity of calibration data Q_C available in 7.4.2.1 and are here benefitted by being given access to more data, that benefit seemingly outweighing the impact of divergence of dataset's origin.

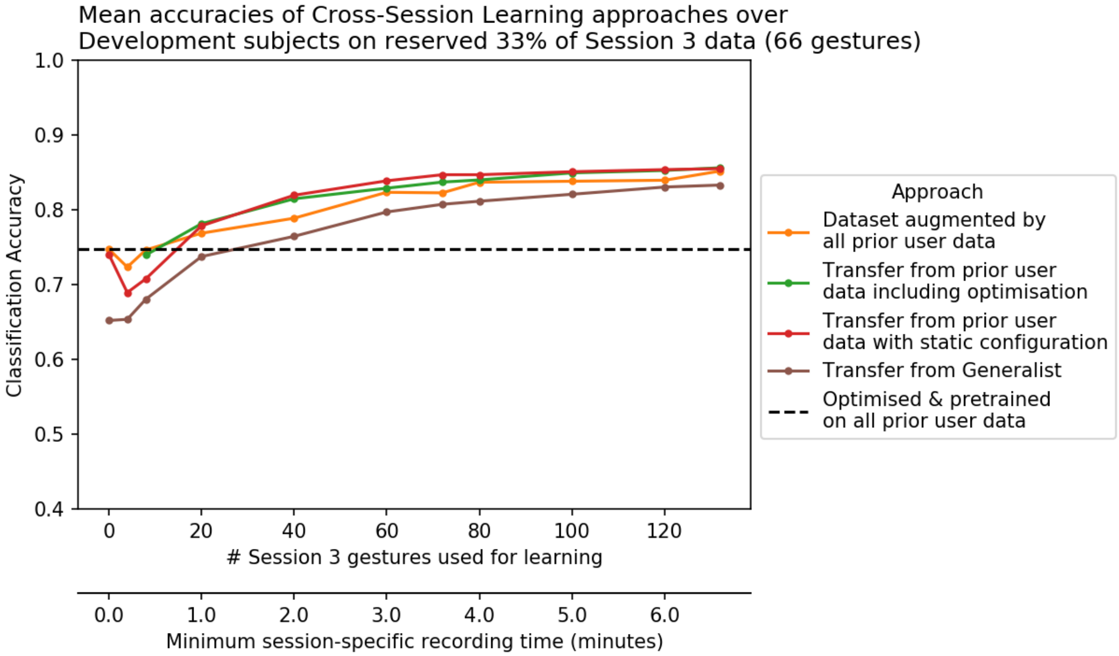


Figure 7.12: Performance of approaches incorporating some degree of direct cross-session modelling

Albeit only by a small degree, the Augmentation strategy routinely underperforms by comparison to those which use direct model transfer. The introduction of the classifier adaptation stage — the eponymous "transfer" step in a Model Transfer system as they are defined here — forcibly directs a system's component machine learning algorithms to pay dedicated attention to data belonging to the target session. Assuming that calibration data taken from the target session will be more similar to (and hence more predictive of) the reserved target-session testing data than out-of-session data will, a learning process tailored to such target-session data ought to reasonably be expected to lead to higher classification accuracy than one which is not, other things being equal. The Augmentation approach however does not distinguish between session-specific and out-of-session data at the training stage; its only mechanism for encouraging such focus is in the use of calibration data Q_C as its target during CASH optimisation. Not only is it thus less tailored to the target session, but the stage at which it is tailored, the determination of the system's hyperparameter configuration, is as previously discussed posited to be a more transferable aspect of the learning process — hence the viability of strategies which "port" their system configurations" — and thus one which may be only minimally benefited by giving specific attention to target-session data.

This interestingly stands in contrast to the findings of Chapter 6's experiments on cross-subject classification. As noted in 6.5.2.3 and elsewhere, in that context the direct adaptation of trained models to a novel subject was no better a strategy than the augmentation of a subject's data with that of other individuals. This is perhaps explainable by the expected greater homogeneity in data collected from a single subject on different occasions than that of data collected from different individuals. In the cross-session case of Chapter 6, provided there was not significant imbalance between source and target domain data, an augmented system wherein subjects' data were dissimilar may still have been able to fit the data in such a way that it modelled class-relevant patterns in the biosignal data of various subjects, even if quite distinct patterns were identified in each. Supposing such dissimilarity between subjects' data however, a model transfer system trained initially on data of subjects not under-test may struggle to adapt its fitted classifiers to an unseen and potentially unalike subject's data. In the cross-session context however, wherein the data universe is limited to that of a single subject, the "cold"-trained fit of a classifer in a Model Transfer system may be more readily adapted to the subject's target-session data; the greater similarity between source and target domains meaning it has less "distance to travel". Plausibly, it may even unadapted be a better predictor of the target-session data than an unadapted cross-subject model would be of the target-subject. In this way a Model Transfer approach appears more suited than an Augmentation strategy to the within-subject cross-session classification problem, where such strengths were not found in a cross-subject paradigm; Augmentation is thus not considered further as a "candidate" cross-session approach.

Of the two Model Transfer approaches which draw on prior subject-specific data, that which incorporates calibration data into the CASH optimisation process, and in doing so ostensibly optimises for a system configuration well-suited to calibration as discussed in 7.3.3.2 and 7.3.3.3 above, in fact appears generally no more accurate in classifying Session 3 data than that which uses a static configuration determined solely on out-of-session data. This indicates that, contrary to expectations, a system configuration selected on the basis of its post-adaptation performance in classifying target-session data was no more suitable than one selected on the basis of offline performance with out-of-session data. This gives further credence to the notion that, at least within the scope of a given subject's data, model selection & hyperparameter tuning may be reasonably transferable system properties, corroborating the related findings in 7.4.2.1 on optimisation strategies of systems trained solely on target-session data.

The only notable exception to this similarity in performance is at particularly low levels of calibration data Q_C. The static configuration variant performs unsurprisingly poorly here, presumably due to over-adaptation to such a small dataset resulting in classifiers which are in fact so overfit to Q_C that the model transfer process degraded their performance from that which they achieved unadapted (i.e. at $Q_C = 0$). The variant which incorporates per-session optimisation is seemingly able to compensate for this, perhaps by finding a system configuration which leads models to be less malleable in their adaptation and hence less susceptible to the domain transfer resulting in such overfit. Regardless, this distinction presents only in systems with access to such little session-specific calibration data that they fail to exceed the pre-trained baseline; there is no apparent benefit in these cases of performing model transfer at all over simply using the subject's out-of-session data for training. Given the lack of relevant distinction between the classification abilities of these two systems, the approach which utilises a preselected system configuration and hence avoids the noted inconveniences to a user of determining this on a per-session basis is clearly preferable, and will be retained as a candidate cross-session strategy.

It will no doubt be of interest in future research to expand on this approach to investigate appropriate techniques for its longer-term deployment. For example, it may be that a "clean" copy of the pretrained model ought simply to be adapted to every novel usage session. However alternatively, a single model which was successively adapted to each new session it encountered could potentially have a greater ability to adapt to gradual drift in the properties of the biosignal data, by maintaining a shorter "distance" between its learned properties and the new data, assuming some linearity to such drift. To explore this in adequate depth would require a multimodal dataset collected from subjects over a greater number of recording sessions, ideally spaced over a longer period of time — further reason that the expansion of the limited range of multimodal EMG/EEG gesture datasets made available to the biosignal research community is paramount.

Perhaps unsurprisingly the Model Transfer strategy which carries out the selection of features, optimisation, and initial training on data not collected from the subject-under-test (the "Transfer from Generalist" (7.3.3.4)) is consistently the weakest among these approaches. Nevertheless it demonstrates a clear ability to specialise on the target subject's Session 3 data and responds better when it has greater quantities of this data upon which to adapt its component classifiers. It even begins to narrow the gap in attainment between it and the approaches which access to subject-specific data for these stages of learning; sufficient target-session data Q_C enables it to compensate in part for the lack of prior subject data. Evidently this is an attractive property of the approach, and one the possibility of which motivated its inclusion in the experiments: that both the system configuration & initial training stages may be partially transferable across subjects through calibration to a target session, allowing for classification accuracies nearly competitive with the other investigated techniques to be reached without any need for a subject's involvement in advance of a system being put to use. Thus, while it does indeed achieve the lowest mean accuracy of these cross-session learning strategies, this apparent viability in a use-case of particular interest motivates it being retained as a "candidate" approach.

7.4.2.3 Resultant candidates

The above refinement of the array of potential session-specific gesture classification approaches has identified four candidates, selected on the bases of either their superior classification accuracies, or attractive use-case-relevant properties such as the source of the data which they leverage. The mean Development Set accuracies of these four, along with the most performant zero-calibration baseline as identified in 7.4.1 and the Generalist — which as noted in 7.3.1.2 should here be taken only as an indicative proxy measure of subject-independent performance — can be seen in Figure 7.13. Table 7.2 presents numerically both these mean accuracies[10] and the Standard Deviation in those accuracies across Development Subjects, at each trialled level Q_C of session-specific calibration data.

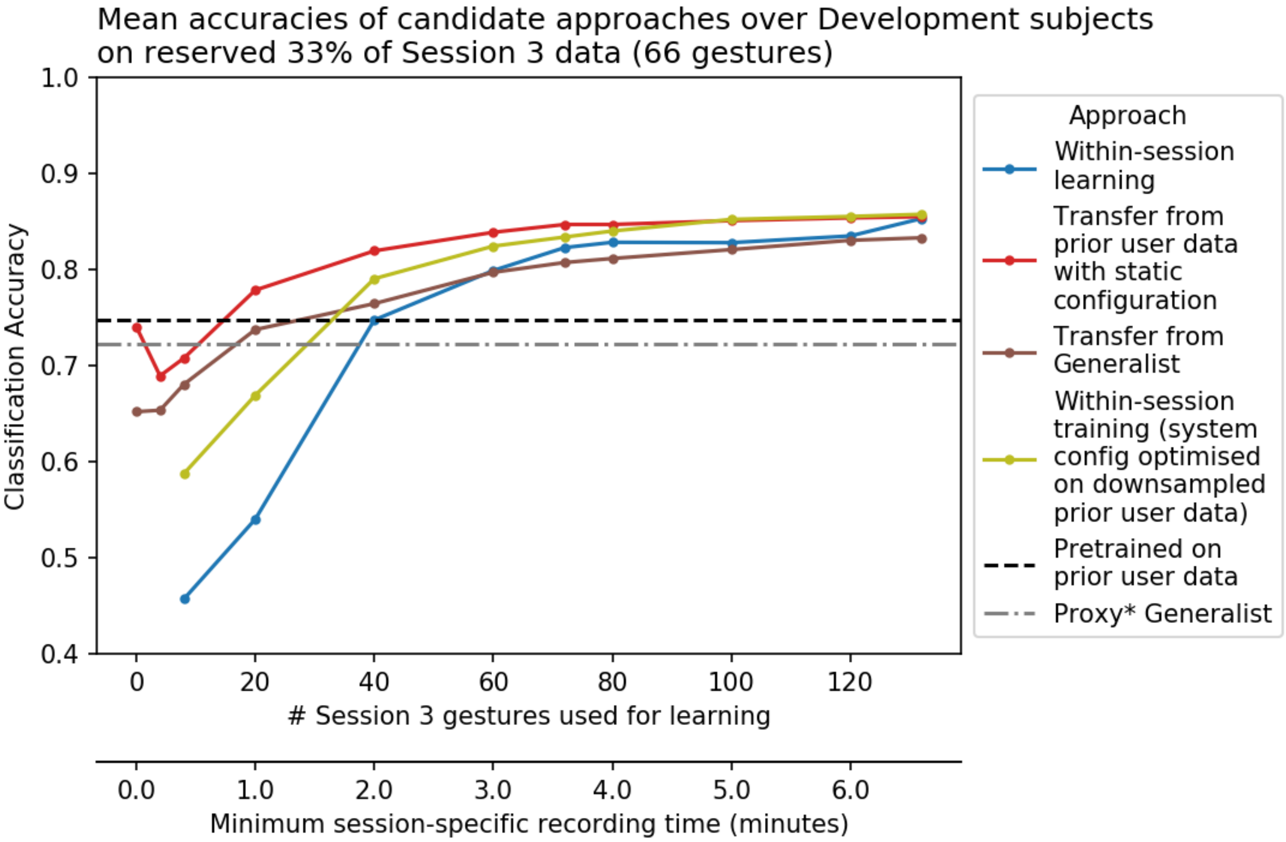


Figure 7.13: Mean accuracies over Development set of candidate session-targeted systems

[10] Recall that as outlined in 7.2.1, individual classifiers of each given approach (at each level of Q_C) were developed for every Development Subject — the mean of a given approach is calculated over those 20 unique systems' subject-specific accuracies, not a single model's attempt at classifying different subjects' data.

Qc	Within-Session		Transfer from user data		Transfer from Generalist		Within-Session (Ported config)	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
0	-	-	0.7407	0.1178	0.6521	0.1142	-	-
4	-	-	0.6894	0.1017	0.6536	0.1319	-	-
8	0.4570	0.1812	0.7078	0.1130	0.6805	0.1443	0.5875	0.1465
20	0.5402	0.1605	0.7786	0.1035	0.7375	0.1392	0.6691	0.1152
40	0.7475	0.1325	0.8195	0.0930	0.7646	0.1395	0.7905	0.0947
60	0.7991	0.1220	0.8388	0.0852	0.7972	0.1120	0.8244	0.0919
72	0.8229	0.1050	0.8470	0.0832	0.8074	0.1049	0.8339	0.0847
80	0.8284	0.1028	0.8470	0.0875	0.8116	0.1058	0.8402	0.0843
100	0.8280	0.1197	0.8509	0.0855	0.8208	0.0973	0.8523	0.0835
120	0.8350	0.1001	0.8538	0.0884	0.8305	0.0964	0.8553	0.0871
132	0.8532	0.0891	0.8551	0.0869	0.8331	0.0965	0.8576	0.0827
Prior Data?	No		Yes		No		Yes	

Table 7.2: Means and Standard Deviations in Development Set classification accuracies of candidate systems at differing levels of calibration data Q_C.
Proxy Generalist Baseline: Mean Accuracy = 0.7230, Std. Dev = 0.0733
Out-of-session Baseline: Mean Accuracy = 0.7475, Std. Dev = 0.1131

From these results a number of key observations can be made which warrant validation on held-out data:

- That the inclusion of session-specific calibration data improves systems' classification accuracy
- That of those systems trialled with access to no out-of-session subject data:
  - With only a small amount of calibration data, they will not surpass an uncalibrated Generalist system; there is no merit to attempting such calibration
  - With sufficient calibration, performance surpassing that of an uncalibrated subject-independent system can be achieved
- That of those systems trialled which do have access to data previously collected from a subject:
  - When only a small amount of target-session data is made available, it is better used to adapt a pre-trained model via transfer learning than to train a session-specific model of a configuration identified using out-of-session data
  - With sufficient target-session data, either one of these strategies can be applied to similar degrees of success
- There is merit to collection of user-specific data in advance of a system's intended use to be used as a basis for transfer learning, over collecting data from other individuals for this purpose.

7.4.3 Verifying findings with holdout data

Approaches' classification abilities when applied to the Holdout Dataset, consisting of data collected from subjects hitherto completely unseen by these systems, can be used to verify these findings in turn. The reader is advised that the candidate approaches presented above are hereafter sometimes referred to collectively as "**calibration**" systems, **including** those which do not strictly calibrate a pre-existing model but rather construct a model using exclusively target-session data. This is in part for convenience and consistency with the description of such data as "calibration data" as noted in 7.2.1, and to reduce the potential for confusion to arise from grouping them under terms such as "cross-session" or "session-specific" previously used in 7.3.3 and 7.3.2 to describe specific subcategories of approach.

7.4.3.1 Demonstrating the benefit of calibration

The first and most significant observation noted above is "*That the inclusion of session-specific calibration data improves systems' classification accuracy*" — directly addressing Aim 7.3: to "*Explore the impact of the amount of session-specific biosignal data available to gesture classification systems on the accuracy of their gesture classifications of a target session's data*".

This can be assessed by comparing systems' performances at low and high levels of session-specific calibration data Q_C. On the basis of those results in Figure 7.13 it can be seen that at levels of Q_C fewer than 20 gestures (i.e. $Q_C = \{0, 4, 8\}$), systems were unable to perform with any greater accuracy than the zero-calibration baseline modelled on prior subject-specific data. It should be noted that this is not intended to assert the threshold of 20 target-session gestures to be universally applicable. Rather it is posited that such a quantity of session-specific data which enables systems to surpass a zero-calibration baseline exists, and that of those quantities trialled here this appears to be no fewer than 20 for this dataset. Where systems were provided with the maximum amount of calibration data however (132 gestures, or 33 of each class), they all comfortably outperformed the baseline's accuracy. There appears remarkably little variation between their scores other than the "Transfer from Generalist" being weaker than the rest, suggesting that in cases where such a level of calibration data is deemed acceptable by a designer or user, the approaches are equivalently performant and could be selected on the basis of their other properties such as computational and data requirements.

The first aspect of this can be verified with a many-to-one test. At low levels of Q_C a session-targeted approach would only be of particular interest if it can outperform a zero-calibration system; if they reached the same accuracy, the simpler pre-trained model would be preferable. The null hypothesis of this test is that: *"there will be no difference in mean classification accuracy between the proposed calibration systems with access to 8 or fewer target-session gestures and the uncalibrated system modelled on out-of-session data"*[11], or more formally:

$$H_0 : \mu_{zero-calibration} - \mu_{calibrated\,(Q_C \leq 8)} = 0. \tag{7.1}$$

[11] The discrete values of Q_C trialled, as per 7.2.2, mean that an observation on the basis of "fewer than 20" gestures can only confidently be stated to apply at "8 or fewer"; $Q_C = 19$, for example, was not assessed.

The candidate approaches' classification accuracies with regard to the 5 Holdout Subjects are evaluated for quantities of calibration data $Q_C \leq 8$.

The impact of the choice of approach is then modelled in *R*, as a linear mixed effects model using the *lme4* package. The effects of variation in performance between subjects, and of the level of session-specific data Q_C, are treated as blocks by modelling random intercepts for each.

Subsequently Dunnett's multiple comparison procedure is used via the *multcomp* package's *glht* function, with the zero-calibration results as the control to which other systems are contrasted. From the results of this test in Table 7.3 it can be seen that there was indeed no measurable difference when compared to the model pre-trained on a Holdout subject's prior data for all approaches but the wholly within-session strategy, which was in fact significantly worse by some margin.

Linear Hypotheses	Estimate	Std Error	z value	p value
within_opt_prior – prior_pretrain	-0.06072	0.05538	-1.097	0.568
within_session – prior_pretrain	-0.20542	0.05538	-3.709	<**0.001**
xfer_gen – prior_pretrain	0.00043	0.04203	0.010	1.000
xfer_prior – prior_pretrain	0.01215	0.04203	0.251	0.995

Table 7.3: Dunnett all-vs-one contrast of Holdout Set performance of calibration systems with 8 or fewer session-specific gestures against zero-calibration baseline

The second aspect, regarding system's performance with high levels of calibration data, is more interestingly assessed with a pairwise test, in that if such systems do exceed an uncalibrated baseline it would be valuable to identify whether any one approach makes better use of this calibration data than another — or if, as the Development Set results appear to indicate, they are largely equivalent. Here the null hypothesis is thus that: *"There will be no differences in mean classification accuracy among the proposed calibration systems with access to 132 target-session gestures nor the uncalibrated system modelled on out-of-session data"*.

Again, a linear mixed-effects model was used to model effect of the system's approach on classification accuracy while accounting for variation between Holdout subjects as blocks. As here all the approaches were tested at only a single level of calibration data ($Q_C = 132$), and the zero-calibration is inherently not a function of Q_C, this was not a necessary factor to model as a random effect.

Unlike some pairwise post-hoc tests performed previously in the work, wherein the homogeneity of variances between groups being observably violated motivated the use of Dunnett's T3 pairwise test (as one of few multiple comparison procedures capable of controlling Type I errors when usual assumptions are violated [366]), here there is no reason to anticipate a lack of homoscedasticity between the approaches. Tukey's test, as the more common method, is thus used — again via *multcomp*'s *glht*.

Table 7.4 presents the results of this comparison. While it is clearly found that the model transfer from prior user data approach and both variations of the within-session learning strategy significantly outperform the uncalibrated baseline, this was not the case for the model transfer system which adapted from other-subject data ($p = 0.12$). There is nevertheless a clearly observable difference between this strategy's mean Holdout performance and the baseline. It may be that the effect size here is simply insufficient to be of

Linear Hypotheses	Estimate	Std Error	z value	p value
within_opt_prior − prior_pretrain	0.13182	0.04159	3.169	**0.0132**
within_session − prior_pretrain	0.13333	0.04159	3.206	**0.0118**
xfer_gen − prior_pretrain	0.09924	0.04159	2.386	0.1191
xfer_prior − prior_pretrain	0.13939	0.04159	3.351	**0.0072**
within_session − within_opt_prior	0.00152	0.04159	0.036	1.0000
xfer_gen − within_opt_prior	-0.03258	0.04159	-0.783	0.9356
xfer_prior − within_opt_prior	0.00758	0.04159	0.182	0.9998
xfer_gen − within_session	-0.03409	0.04159	-0.820	0.9247
xfer_prior − within_session	0.00606	0.04159	0.146	0.9999
xfer_prior − xfer_gen	0.04015	0.04159	0.965	0.8707

Table 7.4: Tukey pairwise comparison of mean Holdout Set performance of calibration systems with 132 session-specific gestures and zero-calibration baseline

statistical significance, particularly given the low sample size of the Holdout dataset. If there is a high degree of variance between so few subjects, as indeed Figures 5.10 & 5.11 and the related discussion in 5.5.2.3 would suggest, this would lead to a lower likelihood of significant effects being found than if that same variance applied over a larger dataset. Visualising the pairwise comparison as in Figure 7.14 further demonstrates the strength of evidence that this effect may in fact be present, and that further tests with larger sample sizes may find it significant, or demonstrate it more conclusively to not be genuine — again highlighting the need for greater quantities of multimodal same-limb gesture performance data by the biosignal research community.

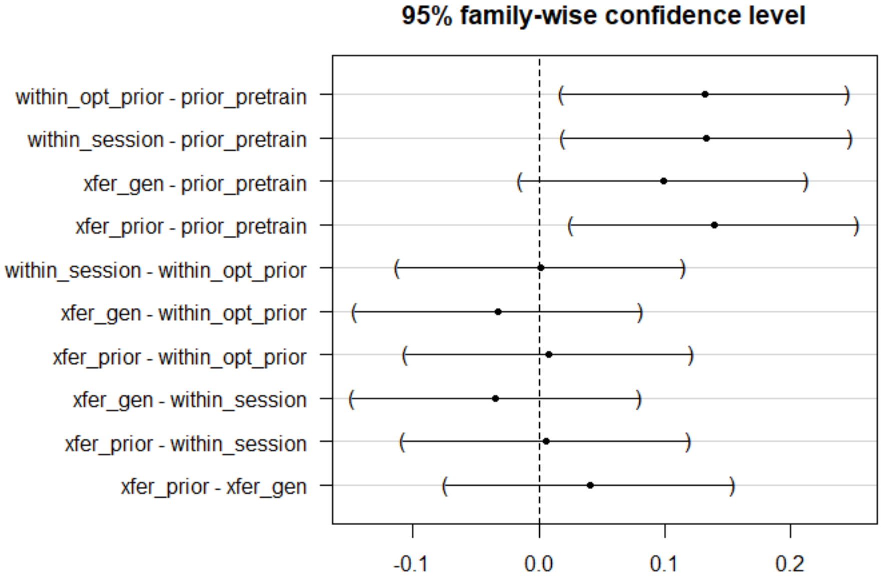


Figure 7.14: 95% Confidence Intervals of differences in means between systems at maximum level of session-specific calibration ($Q_C = 132$), estimated by Tukey pairwise contrast

Notwithstanding this, it should also be noted that the pairwise comparison did indeed find no significant differences among the various calibration systems[12], seemingly verifying this phenomenon observed from Development Set data. As mentioned, it is acknowledged that the small size of the Holdout Set here is likely to reduce the statistical power of these tests & thus make small differences difficult to distinguish, leaving open the possibility that effects were present but simply not identifiable with this test. The estimates differences in means between the systems however can be observed from Table 7.14 to be very low, below even their Standard Error; it is thus not likely that there were any meaningful differences to be found. The supposition thus holds that these approaches can, when provided with such a high level of target-session data for calibration, be considered of broadly equivalent classification ability over further target-session data.

Additionally by demonstrating these systems to have been no more accurate than the baseline when provided only 20 or fewer calibration gestures, but to have exceeded it when provided the maximum available quantity of 132 Q_C (with the single aforementioned exception), the broader finding that these systems' predictive power is greater with more session-specific data than with less is verified.

7.4.3.2 Systems without prior user data

Among these candidate systems, two approaches (and one baseline) make use of no subject-specific biosignal data other than that collected in the target session. Here by investigating these we seek to address Aim 7.2: to "*Explore strategies for session-specific gesture classification which do not require prior collection of subject-specific biosignal data*".

The first observation noted regarding such systems' performances on Development Set data was stated informally as a finding that "*With only a small amount of calibration data, they will not surpass an uncalibrated Generalist system*", though noting the significant caveats to the proxy measurement of Generalist performance on the Development Set outlined in 7.3.1.2. Figure 7.13 reveals that the level of target-session Q_C below or equal to which neither the wholly within-session strategy nor the model adaptation from other subject's data approach notably exceeded the proxy estimate of the Generalist was again 20 gestures, thus leading to the formalised null hypothesis:

$$H_0 : \mu_{subject-independent} - \mu_{no\,prior\,user\,data\;(Q_C \leq 20)} = 0. \tag{7.2}$$

As with 7.4.3.1 it should be stressed here that this specific value is not claimed to be a universal threshold, merely that which appears relevant to the particular dataset upon which these experiments are conducted.

In a similar fashion to the previous tests against the subject-specific out-of-session baseline, a Dunnett contrast can provide a suitable many-to-one test between the two "calibration" systems with no access to prior data collected from a user, and the Generalist which has access to no user data at all. As 7.3.1.2 describes, in the context of the Holdout Set a Generalist system is fully subject-independent and thus can be fairly compared here. Again the effect of a system's choice of approach on its performance is modelled with a linear mixed effects model, and variation between Holdout subjects and different quantities of calibration data Q_C accounted for as blocks in the manner previously described.

[12] Interestingly including the Model Transfer from Generalist, though there is a greater measured difference in mean accuracy between it and others.

Hypothesis	Estimate	Std Error	z value	p value
within_session – generalist_proxy	-0.21112	0.06008	-3.514	**0.0008**
xfer_gen – generalist_proxy	-0.06357	0.05259	-1.209	0.3083

Table 7.5: Dunnett contrast of mean Holdout Set performance of strategies without prior subject-specific data with access to 20 or fewer session-specific gestures against baseline Generalist system

The results of this test in Table 7.5 verify that neither strategy outperformed the Generalist at such levels of calibration data, and indeed that the wholly within-session approach was statistically weaker. It is notable that the "Transfer from Generalist" approach was competitive with the Generalist itself, given that the former draws on a much reduced total quantity of other-subject data. While this corroborates the findings presented in 6.5.2.4 (that systems with only few subject-specific data were unable to tailor a model developed on other-subject data with it to be any more accurate than an unadapted Generalist), the lack of significant performance differences here is particularly interesting in light of the systems' vastly different data requirements. It suggests that equivalent levels of accuracy can be attained either by more extensive gathering of initial biosignal data from a range of individuals before a system's deployment, or by collection of far less other-subject data and the incorporation of a per-session calibration routine. The relative costs and benefits of these two undertakings will no doubt differ according to the intended use-case of a system and the priorities of its users and developers but is evidently a factor to be considered, and certainly motivates further research into the degree to which session-specific calibration can enable a reduction in upfront data requirement and vice-versa and the extent to which this relationship generalises across datasets.

The second observation specific to these systems, that "*with sufficient calibration, performance surpassing that of an uncalibrated subject-independent system can be achieved*", identified no singular superior system and so is again most aptly investigated with an all-vs-all test.

Here the approaches' classification accuracies when applied to Holdout data at all levels of Session 3 Q_C defined in 7.2.2 greater than 20 are evaluated. The effects of variation between Holdout subjects and between different levels of Q_C are again modelled as random effects. As in 7.4.3.1 pairwise testing is again performed with Tukey's method, the results of which can be seen in Table 7.6.

Hypothesis	Estimate	Std Error	z value	p value
within_session – generalist_proxy	0.028281	0.036697	0.771	0.711
xfer_gen – generalist_proxy	0.027397	0.036697	0.747	0.726
xfer_gen – within_session	-0.000884	0.018303	-0.048	0.999

Table 7.6: Tukey pairwise comparison of mean Holdout Set performance between within-session learning and model transfer from other-subject data systems each with access to more than 20 target-session gestures, and with baseline uncalibrated Generalist system.

Somewhat surprisingly neither of the calibrated systems here significantly outperform one with access to no subject-specific data, though in both there appears to be a small measured superiority over the latter, to a similar degree for each. The aforementioned variance in the ease of classifying Holdout subjects' data may explain this in part, if it is indeed leading to effect sizes too small to be of significance at the $\alpha = 0.05$ level with so few samples to average over. It may also be pertinent to recall that, in Chapter 5, Generalist systems were found to perform slightly more accurately with Holdout subjects than they were with Development, and Bespoke systems the opposite. It is not implausible that, through random chance, there is lower within-subject similarity in the data of Holdout subjects than in that of Development. That is to say that, through either a groupwise trend or the strong influence of individual members, a random sample of a Holdout subjects' biosignal data is on average less predictive over a reserved third of their Session 3 data than is the case for Development subjects. If so, the benefit to a calibrated system of receiving increasing amounts of target-session data would be dampened for Holdout subjects by comparison to Development; there would be less gained from such specificity. Such a greater intra-subject inconsistency could also make Holdout subjects' systems on average more heavily influenced by the random downsampling of Session 3 data to a given level of Q_C, and have potential to exacerbate the variance between subjects.

Certainly further research with a wider number of subjects would be needed to investigate these possibilities in depth. The clear superiority seen among the Development Set of approaches without prior user data but which can access session-specific data over the Generalist which cannot, and the presence of a small measurable effect of a similar nature in held-out subjects' data, evidences the possibility that session-specific calibration could be of assistance in enabling gesture recognition systems which do not rely on individualised tailoring to each novel user prior to deployment. The scope for increased convenience and ease-of-access resulting from such properties of a system clearly motivate their continued exploration in future studies on the basis of the potential viability shown in this work.

7.4.3.3 Systems with access to prior subject-specific data

Having considered those approaches which forgo advance subject-specific data collection, it naturally follows to review those which do have access to such data. The two candidate approaches of this kind make use of subjects' out-of-session data in distinctly different ways — in one case solely for determining the appropriate system configuration & feature ensemble for a session-specific model, and in the other as an initial stage of training of the final model itself — it would be useful to establish if either technique is the superior option.

From the exploratory Development Set results, it was as observed above that while the two strategies were similarly performant when provided with high quantities of target-session data, the Model Transfer approach provided greater classification accuracy at lower levels of Q_C, thought to be a result of the session-specific model having insufficient training data to avoid overfitting.

Through comparing these system's performances on the Holdout Set, the two aspects of this finding can be verified. Reviewing firstly their performances where smaller amounts of Session 3 data Q_C are available, Figure 7.13 indicates the two approaches' mean classification accuracies over Development Subjects to converge at $Q_C = 60$. Of course, as previously discussed this value will not necessarily be a universal one, as datasets of different subjects whose recorded biosignals have different properties may exhibit different responses to increased levels of calibration. Nor, it should be noted, is this threshold claimed to have been found with absolute certainty even for this dataset. Modelling Q_C as a continuous variable would have inflated the computational load of these experiments to an impractical degree, and while the range of discrete values trialled for Q_C (as presented in 7.2.2) are expected to give sufficient coverage to broadly understand its impact, their precision when taken as a scale will for most cases[13] only give an indication to within ± 20 gestures; the performance of a system with one more or one fewer gesture per class, i.e. 56 or 64 gestures total, was not assessed. Nevertheless, to assess such a finding the "lower" and "higher" quantities of calibration data must be demarcated in some way, and the Development Set results provide the only suitable grounds upon which to do so while preserving the integrity of the test. To make such a determination instead on the basis of reviewing any visible trends in systems' Holdout Set performance would undermine the validity[14].

Thus, the accuracies of these systems with the five Holdout Subjects were computed for all cases where 40 or fewer session-specific gestures were made available. Similar to previous tests, the relationship between system choice and accuracy achieved was modelled as a linear mixed-effects model, accounting for both subject and calibration data quantity Q_C as blocks. The null hypothesis for this test is that "Where each has access to 40 or fewer target-session gestures of a given subject, there will be no difference in accuracy between a system using those data to adapt a model trained on data previously collected from that subject, and one which uses the data to train a model of a configuration determined by previously-collected data", or:

$$H_0 : \mu_{model\,transfer\,(Q_C \leq 40)} - \mu_{ported-config\,session-specific\,(Q_C \leq 40)} = 0. \tag{7.3}$$

Although here there is only one pair of systems to be compared, for practical convenience this was again tested with Tukey's Honestly Significant Difference pairwise method; for two groups this will be essentially equivalent to a conventional t-test of the single hypothesis.

Hypothesis	Estimate	Std Error	z value	p value
within_opt_prior – xfer_prior	-0.02386	0.01954	-1.221	0.222

Table 7.7: Tukey pairwise comparison of mean Holdout Set performance between the two candidate systems with access to prior subject-specific data at levels of $Q_C \leq 40$. (NB for a single pair this is functionally equivalent to testing as one linear hypothesis, but is performed here with Tukey's method for consistency in format of results.)

Table 7.7 shows that surprisingly there was no significant difference in these systems' performances, and while the Model Transfer approach's mean accuracy was observably greater than that of the ported-configuration within-subject strategy this was only by a very small margin. This indicates that, contrary to the apparent performance differences seen with Development Set data, here the use of Holdout Subject's Session 1 & 2 data for "cold" training of models was of no meaningful benefit; this approach was not able to compensate for the potential overfit of systems trained only on session-specific data. Again, this could potentially be explained by a difference in the within-subject informativity of Holdout subjects' data and Development subjects'; the Holdout Set may be only marginally benefitted by learning from out-of-session data. It may be that such data was sufficiently dissimilar that the "cold"-trained models were in greater need of extensive adaptation than was able to be carried out with the low quantities of target-session data

[13] Notwithstanding the nonlinear spacing of a few Q_C values

[14] See also 6.5.2.2 and other parts of 6.5.2 wherein candidate systems were similarly selected according to Development Set results, regardless of their performance on Holdout Subjects.

Q_C without high risk of overfit. It also bears mentioning that at increasingly low levels of calibration data, the Development Set results of Figure 7.13 & Table 7.2 indicate an uncalibrated pre-trained model as being an increasingly attractive option where a subjects' prior data is accessible, over either of the approaches considered here. Among Development Subjects the "model transfer from prior data" strategy drops below baseline performance where $Q_C \leq 8$ and the "within-session using ported configuration" where $Q_C \leq 20$; though these specific thresholds are unlikely universal, it may be that in some cases the two systems' accuracies are only separable at such low levels that they area both outclassed anyway.

Subsequently these systems' performances at 60 or more target-session gestures were assessed by the same process; the null hypothesis itself being not dissimilar to (7.3):

$$H_0 : \mu_{model\, transfer\, (Q_C \geq 60)} - \mu_{session-specific\, with\, ported\, config\, (Q_C \geq 60)} = 0. \tag{7.4}$$

Hypothesis	Estimate	Std Error	z value	p value
within_opt_prior – xfer_prior	-0.0001263	0.00063428	-0.02	0.984

Table 7.8: Tukey pairwise comparison of mean Holdout Set performance between the two candidate systems with access to prior subject-specific data at levels of $Q_C > 60$. (NB for a single pair this is functionally equivalent to testing as one linear hypothesis, but is performed here with Tukey's method for consistency in format of results.)

The results of this test as seen in Table 7.8 demonstrate that as had been observed with Development Set data, these systems do indeed offer equivalent classification accuracy when provided with sufficient quantities of calibration data Q_C. While in the case of Development Set results this was a convergence and here with regard to the Holdout Set it is a continuation of the system's comparable performances, it can nonetheless be concluded that they are equivalently suitable on the basis of target-session accuracy. A potential choice between the approaches could thus be made according to other considerations without adversely affecting performance. While both are similar in that they do indeed rely on some form of per-session calibration, and on obtaining subject-specific data on multiple occasions before a system can be used, the session-specific strategy wherein that prior data is used only for Feature Selection and CASH optimisation allows for such in-advance data collection sessions to be shorter in duration, and may well be viewed as preferable for this reason.

7.4.3.4 The merits of collecting subject-specific data for session-specific transfer learning

The final key observation made regarding the candidate approaches concerned the two Model Transfer approaches. These differ by the source of the data upon which they carry out the initial stages of the learning process — in one this is performed using data collected from the subject-under-test, and in the other using data from a number of unrelated individuals. From the Development Set results it appeared that subject-specific data was of greater benefit here. The merits of each approach on a practical basis in terms of their application in a deployed system, and the extent to which a given expected improvement in classification accuracy justifies the potential cost of one approach over the other, are subjective design decisions which cannot be answered here. Nonetheless it is of note that the approach using subject-specific data as the source domain for its transfer learning outperformed that which drew on other subjects' data by a clear and remarkably consistent degree. This was not unexpected considering it was in line with Chapter 6's findings that there was much less to be gained from inclusion of other-subject data by comparison to incorporation of additional subject-specific data, but warrants verifying here with the Holdout Dataset.

It should be noted that 7.4.3.1 already hypothesised, and indeed confirmed, that at very low levels of calibration data ($Q_C < 20$) these two systems could not outperform a baseline trained on out-of-session data. Similarly, 7.4.3.2 found the "Model Transfer from Generalist" approach to be no more accurate than a true subject-independent Generalist when only such small quantities of calibration data were available. In either of these systems' use-cases therefore — that where subject-specific data has been collected and that where other-subject data has been — there is no merit to performing target-session calibration if fewer than 20 gestures are available for this purpose; an uncalibrated system, which will inherently be of greater convenience, can perform just as well. Therefore these approaches were compared only on the basis of their performance at quantities of $Q_C \geq 20$, as their classification abilities below this level are irrelevant.

The previously described linear mixed-effects model, blocking for the impact of the subject and the quantity of target-session data Q_C by modelling random intercepts, was used to assess the effect of the choice of approach on system accuracy. As in 7.4.3.3, for convenience the single null hypothesis:

$$H_0 : \mu_{transfer\,from\,user\,(Q_C \geq 20)} - \mu_{transfer\,from\ others\,(Q_C \geq 20)} = 0 \tag{7.5}$$

was tested with Tukey's post-hoc method of multiple comparisons.

Hypothesis	Estimate	Std Error	z value	p value
xfer_gen – xfer_prior	-0.01705	0.01480	-1.152	0.249

Table 7.9: Tukey pairwise comparison of mean Holdout Set performance between the two candidate systems which perform model transfer from different sources at $Q_C \geq 20$. (NB for a single pair this is functionally equivalent to testing as one linear hypothesis, but is performed here with Tukey's method for consistency in format of results.)

As seen in Table 7.9, the expected consistent advantage of using subject-specific data as the basis for the model transfer was not found to be significant here & this finding could not be verified. While there was a groupwise difference in means observed this was much smaller than appeared typical of the Development Set results in Figure 7.13. It is again anticipated that this could partially be ascribed to a relatively low informativity of Holdout Subjects' out-of-session biosignal data over their Session 3 data. If this is the case for at least some of the held-out subjects, it would plausibly diminish the extent to which their own Session 1 & 2 data is of greater benefit than that of other subjects, weakening the predictive power of the "Transfer from prior data" approach. That there was at least some measurable difference between the systems' performances in both Development and Holdout sets, and that there appears a possible explanation for this effect being so diminished among Holdout subjects, evidences the merit further investigation in future work with a less limited sample size.

It is again apparent that the scarcity of public multimodal EMG/EEG datasets, particularly of same-limb gesture performances and for multiple-session data, is a significant limitation on the capacity of research in this area. It takes little imagination to conclude such challenges to be a motivating factor in the limited statistical validity of many works in the domain. While the dataset published by Jeong et al. [198] is by no means intended as a target of criticism here, there is no doubt that future research which can provide datasets of a similar nature larger in scope, or even approaching the scale and depth of the kind of gold-standard datasets which exist for unimodal biosignal research such as the NinaPro EMG datasets [195] or the "BCI competition" EEG datasets [194] will be of great benefit to the field.

7.4.3.5 Further evaluation of Holdout performance

A recurring trend among these tests was that a number observations made on the Development Set data failed to generalise to the Holdout Set, or only presented as weak trends which were not statistically significant. It appears that when systems were trialled with the Holdout data there were ultimately far fewer significant differences between the accuracies they reached than was expected. One hypothesised contributing factor for this was a greater variation, on average, of Holdouts Subjects' data on a within-subject basis. That is, that the informativity over a random reserved 33% of an individual's Session 3 data of any given other portion of their data, whether taken from the same session or a previous, was lower on average for the Holdout Subjects than it was the Development. The apparent effect of this, if it was indeed the case, being both that increased quantities of calibration data were of a diminished benefit, and that reliance upon out-of-session data was likewise less useful.

To explore this, and to further review Holdout performance on the whole, mean classification accuracies over the five Holdout subjects of all the candidate systems at all applicable quantities of target-session data Q_C are presented in Figure 7.15 & Table 7.10.

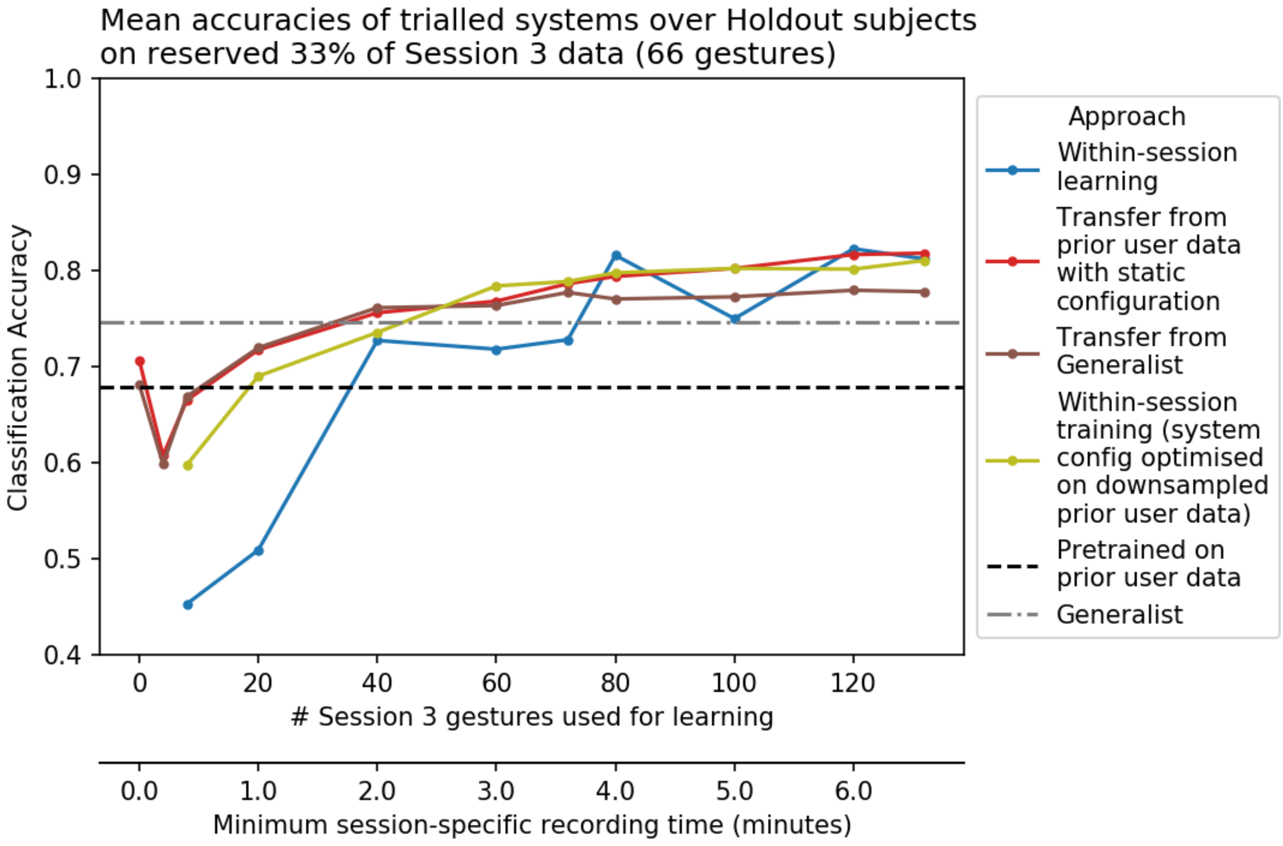


Figure 7.15: Performance of candidate systems on Holdout subjects

Qc	Within-Session		Transfer from user data		Transfer from Generalist		Within-Session (Ported config)	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
0	-	-	0.7068	0.0527	0.6818	0.0495	-	-
4	-	-	0.6076	0.1942	0.5985	0.1500	-	-
8	0.4530	0.1306	0.6652	0.1194	0.6689	0.1027	0.5977	0.0801
20	0.5091	0.1198	0.7174	0.0700	0.7197	0.0469	0.6902	0.0638
40	0.7273	0.1468	0.7561	0.0827	0.7614	0.0446	0.7356	0.0631
60	0.7182	0.0949	0.7682	0.0978	0.7636	0.0586	0.7841	0.0603
72	0.7280	0.0970	0.7864	0.0765	0.7773	0.0741	0.7886	0.0655
80	0.8159	0.0652	0.7939	0.0570	0.7705	0.0733	0.7977	0.0628
100	0.7500	0.1248	0.8023	0.0649	0.7727	0.0687	0.8023	0.0626
120	0.8227	0.0857	0.8167	0.0615	0.7795	0.0734	0.8015	0.0681
132	0.8121	0.0785	0.8182	0.0631	0.7780	0.0806	0.8106	0.0754

Table 7.10: Means and Standard Deviations in Holdout Set classification accuracies of candidate systems at differing levels of calibration data Q_C.
Generalist Baseline: Mean Accuracy = 0.7462, Std. Dev = 0.0640
Out-of-session Baseline: Mean Accuracy = 0.6788, Std. Dev = 0.0669

For transparency, it must be acknowledged here that these results were obtained before the statistical tests outlined above were performed; these means are calculated from the exact same subject-wise classification accuracies which were assessed in those tests. The decision to refrain from presenting them in this format until *after* those discussions was to ensure it was clear to the reader that design and construction of the tests, and formulation of their hypotheses, was done solely on the basis of Development Set data — not with any advance knowledge of Holdout Set performance. It also reflects the author's own experience, in that in striving to minimise the risk of inadvertently biasing any tests these data went genuinely unseen, and the mean results over holdout subjects not visualised in a comparative plot like that of Figure 7.15, until after the tests had been carried out. As has been discussed throughout the work, that many biosignal studies [30, 192] (and indeed much Machine Learning research across domains [186, 191]) are undermined by weak statistical practices and unrecognised data leakage, meant careful and transparent handling of data was considered paramount in this work and is a significant strength of it over many in the field.

By comparison of Figures 7.15 and 7.13 it can be immediately seen that systems' accuracies when trialled with Holdout subjects trended lower than those of Development Set experiments. Most strikingly, the zero-calibration baseline trained on out-of-session data was notably weaker for Holdout subjects than for Development, not even reaching a mean classification accuracy of 70%. This gives credence to the notion that Holdout subjects' data, on average, were more heterogeneous on a within-subject basis than that of Development subjects. The average predictive power of their Sessions 1 & 2 data over that of their 3rd session was evidently lower, though as discussed above further investigation would be needed to ascertain whether this is due to a group-wide trend or the presence of outlier subjects.

This, as hypothesised in 7.4.3.4 above, seems to have a knock-on effect on the accuracy of the other approach which uses out-of-session data for training. Where in Development Set experiments the "Model

Transfer from prior data" strategy consistently outperformed the "Model Transfer from Generalist", here it has no such advantage unless provided with significant quantities of calibration data. In cases where fewer than 80 target-session gestures were provided, data previously collected from a Holdout subject proved no better a basis for model transfer than data belonging to other subjects entirely. It may be noted that the mean accuracy of an unadapted Generalist was higher for Holdout subjects than for Development — though as discussed in 7.3.1.2 this is unlikely to be a true representation of a subject-independent system's performance for the latter — but this is only by a small margin. Certainly it could be that this equivalence of the two Model Transfer variants at low quantities of Q_C is influenced not only by Holdout subjects' prior data being of less benefit, but of them benefiting *more* from other-subject data than Development subjects did (perhaps as suggested in 5.5.2 due to Holdout subjects having a wider range of "other-subjects" from which to draw). However this is unlikely to account for the full extent of the performance difference of the "Transfer from prior data" system between Holdout and Development subjects, given that the out-of-session baseline's reduced accuracy for the Holdout Set already evidences a lesser within-subject consistency for Holdout subjects.

Such heterogeneity within Holdout subjects' data may also explain the non-monotonic nature of the mean within-session Holdout Set performance. Assuming a sample of data (Q_C) of a given subject's 3rd recording session is likely to share properties with other data taken from that same session (the 33% reserved for testing), one would expect this approach's accuracy to be an increasing function with respect to Q_C, or at least to not present such sharp decreases as are visible in Figure 7.15. If however the extent of this similarity is highly inconsistent among datapoints, the downsampling of the target-session data to Q_C could introduce a signficant random component — if not all the available data is of broadly equal informativity, then an increased quantity of data will not necessarily mean an increased quantity of *useful* data, as this will be in part defined by the random chance of the informative data being found among the sample. Of course a larger size for Q_C will (assuming an unbiased downsampling process) increase the likelihood of informative data being captured within it, hence there still being a clear positive correlation between Q_C and accuracy for this system, but the random noise introduced may account for the coarseness of that relationship.

Despite this the Holdout results here do overall present some broadly similar trends to those seen with the Development Set. Systems have a similar rate of improvement in accuracy in response to increasing levels of calibration data and, though at a much higher threshold of Q_C, do demonstrate an ability to outperform a pretrained model when provided sufficient target-session data (albeit not statistically significant at the $\alpha = 0.05$ level in all cases, as per 7.4.3.1 & 7.4.3.2). All the candidate approaches — except the "Transfer from Generalist" which reaches a lower peak accuracy — appear to begin saturating at a similar classification accuracy to one another, including those session-specific strategies which are notably less performant due to overfit at low quantities of Q_C.

7.4.4 Conclusions

These experimental results clearly demonstrate the potential viability of a number of strategies for classifying same-hand gestures from EMG & EEG data of a specific recording session. While session-independent, zero-calibration approaches showed strengths they were outclassed by ones which were targeted in some way on the session-under-test. Wholly session-specific systems however were undermined by reductions in the quantity of available target-session biosignal data, and thus may be less suited to use-cases where minimisation of the burden placed on a user by per-session calibration is desirable. More robust were strategies which drew on both session-specific calibration data and biosignal data collected from other sources. Exploratory experimental results showed evidence that incorporation of cross-session subject-specific data may be more successful in enabling a reduced calibration need than a similar use of other-subject biosignal data, though the signficance of this could not be confirmed with the limited amount of multimodal, multi-session, same-limb biosignal data available for use in the work.

That such use of data from other individuals was nevertheless capable of measurably benefiting minimally-calibrated systems has particularly interesting implications when considering that this approach could, in a "real-world" system, require much less (if any) user-specific setup prior to deployment than many of the competing strategies attempted here. Their viability demonstrated, the resultant potential improvements in affordability and ease-of-access of such systems over the conventional bespoke nature of many systems in the field clearly motivate their continued exploration in future research. Particularly, the extent to which such subject-independent systems can be improved through per-session calibration by a novel user could not be verifiably measured here and will naturally be of great interest to subsequent work.

This latter question was one of a number which the tests in 7.4.3 were unable to confirm with confidence; trends were measurable both in initial experimental results and when using wholly unseen data, but were not found to statistically significant when measured over the five subjects held out from the dataset during exploratory investigation. As highlighted in discussion throughout the chapter, this does not undermine the findings of this research but is in fact evidence of the work's strength. Large portions of the literature on EMG & EEG gesture classification are susceptible to the flaws of much current Machine Learning research: poor experimental design, lack or insufficiency of statistical tests, unjustified modelling choices in which bias or "cherry-picking" cannot confidently be ruled out, and more lead many studies to unreproducible claims of the superiority of certain models over others which are not genuinely supported by their results [30,31]. Whilst this work certainly does not claim to be wholly free of such faults — it is unlikely that any research ever truly could — approaching the problem with sound, unbiased experimental, statistical, and data handling practices, and transparency over them, has been a consistent and deliberate focus throughout. Thus while acknowledging that certain findings were unverifiable here — there is no intention to claim or imply significance where it was not found — these results highlight a number of areas as clearly worthy of further detailed investigation, particularly when larger multimodal datasets similar to that of Jeong et al. [198] are available.

The limitations of available biosignal datasets also restrict the extent to which the calibration strategies explored here, or in prior works, can be assessed. As discussed in 7.1, a number of both short and long term factors can lead to inter-session data drift. While multi-session datasets are limited to the order of a few weeks as in the case of that collected by Jeong et al. [198], only some of these — such as sensor positioning, environmental conditions, and minor changes in the subjects' physiology or fatigue levels — can be reasonably expected to vary between data collection sessions. Other potential contributors like the effects of ageing [365], muscle build-up or atrophy, and pertinent to prosthesis contexts the natural cortical re-mapping which can take place following amputation [99], could only be captured by longitudinal studies, and thus the capability of the proposed calibration strategies to handle such concerns would require further investigation to assess. This is a limitation of many works in the field; in many cases data were recorded on consecutive days [113,364], or a similar order of weeks to the data used in this work [175]. Seeland et al. [295], who notably found that sufficient out-of-session data could alleviate the need for calibration, and Krauledat et al. [362] who similarly found significant quantities of historical data to aid session-to-session transfer, unfortunately do not characterise the time differential between sessions. It is thus hard to ascertain whether their approaches, grounded respectively in self-supervised model retraining and computation of CSP filters usable across sessions, could be more or less suited to long-term use than this chapter's proposed strategies of using out-of-session data to augment datasets, train models for adaptation, or determine suitable modelling choices via CASH optimisation. It is clear that the creation of longitudinal biosignal datasets is vital for the advancement of cross-session transfer learning techniques to establish the confidence with which they can be employed in real-world prosthesis systems.

7.4.4.1 Implications: On the trade-offs in time, burden, and accuracy

As a final note, it would be pertinent to consider further the "real-world" implications of the calibration discussed extensively throughout this chapter.

Aim 7.4 outlined the intention to "*Identify suitable approaches for session-specific gesture classification with access to different quantities of target-session biosignal data*". This research does not claim to have found a singular universally superior strategy for using such data. While there are differences in performance among the candidate approaches, as discussed throughout 7.4.3 these differences are often marginal and were not all found to be statistically significant at the $\alpha = 0.05$ level, and thus may not be of great relevance. The comprehensive analysis of the results however paves the way for future researchers and system developers to consider the trade-offs between data requirements and achievable accuracy not only between, but also within systems — and to compare systems on the appeal of their various best-suited use-cases.

The candidate systems vary not only in the data required for their "initial" setup, but also in their response to alterations in the quantity of available calibration data. Consider the scenario presented in Figure 7.16. Here we imagine that a user or developer of a system wishes to reduce the per-session calibration burden by five minutes[15], and judges a resultant reduction in classification accuracy of no more than 2.5% to be acceptable. Figures 7.16b and 7.16c demonstrate that these criteria could not both be met, and present two possible outcomes dependent on which criterion is prioritised. The reader will note that these trade-offs are considered on the basis of the "Model Transfer from prior subject-specific data" approach, as the

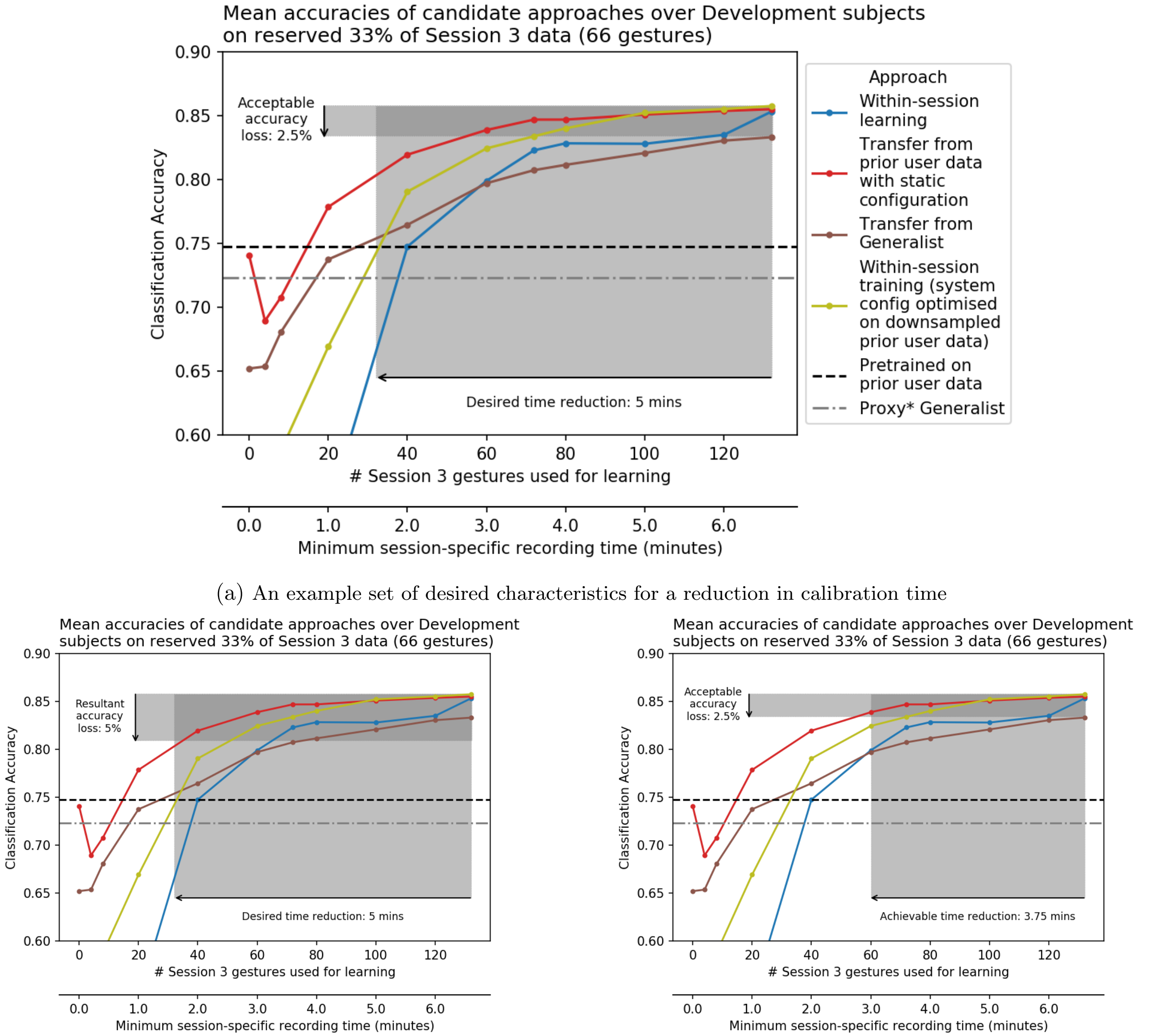
[15] For simplicity factors other than the recording of labelled calibration data which, as discussed above in 7.4.2, may also contribute to the total required calibration time are not considered for the purposes of this illustrative discussion, but should not be disregarded as unimportant.

strategy found most performant at reduced levels of calibration data Q_C. Should a system be designed for an alternative use-case such as to avoid the need for advance data collection from the subject who will use it, its response to a scarcity of calibration data will be different, and hence so will the achievable reduction in calibration time at a given level of acceptable loss; this itself is a property which a designer must take into consideration. Of course no attempt is made to be prescriptive here — the example criteria given in 7.16a are purely illustrative, and would ultimately vary according to the priorities of the individual using a deployed gesture-classification-based prosthesis system; Figure 7.17 illustrates examples of both stricter and more forgiving criteria, which would each encourage the selection of different calibration strategies. Neither are these response curves themselves claimed to be necessarily universal. Figure 7.15 demonstrates that even within the context of the dataset used in this study, while both Development and Holdout subjects clearly displayed similar trends in their systems' responses to varying quantities of calibration data, they were not exactly the same. While the Development Set results are presented here for illustrative purposes, but this is naturally a property the developer of a system would need to characterise.

The ramifications of a change in the required calibration time of a system ought also to be stressed here. It would be easy for a reduction of five minutes, as in Figure 7.16a, to instinctively seem academic. Particularly to an able-bodied researcher, for whom the real-world application of a gesture recognition system such as that in a prosthetic limb may feel distant, the impact of a system on the people who use it & their quality of life can be difficult to fully consider. Indeed some research has even advocated against tailoring of gesture recognition systems to users on the grounds that the users will themselves adapt their behaviour to the algorithm [367]; such positions are certainly not highlighted here to suggest ableist beliefs on the part of their authors in any way, but they can serve as a reminder of the risks of not duly considering the effect of a system or device on its user.

One could choose to position systems' required calibration times as relative to their expected usage. In such a light, a five minute data collection procedure for calibration purposes could be framed as a near-trivial price if it improved the accuracy of a system which went on to be used for a "session" lasting multiple hours. The able-bodied reader is invited however to consider a comparison, albeit a highly imperfect parallel, to a device they use on a regular basis for multiple hours at a time. Many, one imagines, would undoubtedly find it difficult to accept a lengthy daily calibration process — if they could countenance any at all — for their smartphone, laptop, or other devices used for work or leisure.

An accessibility tool, such as in this context a prosthesis, ought to be inobtrusive and any burden it places on its user minimised — lest it risk becoming an access barrier unto itself, rather than a means of their dismantling.

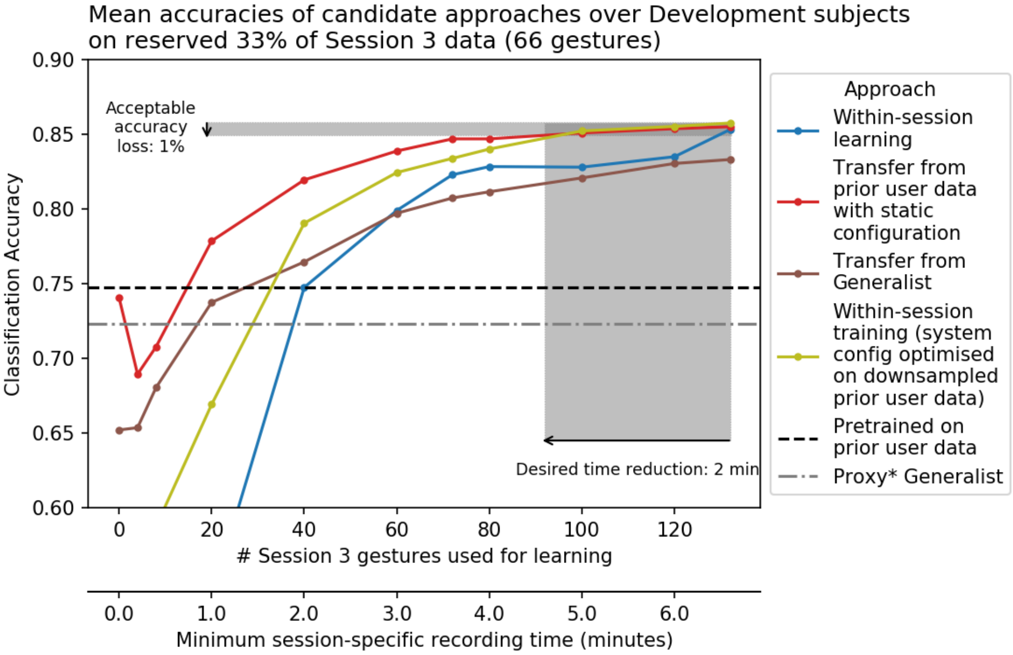


(a) An example set of desired characteristics for a reduction in calibration time

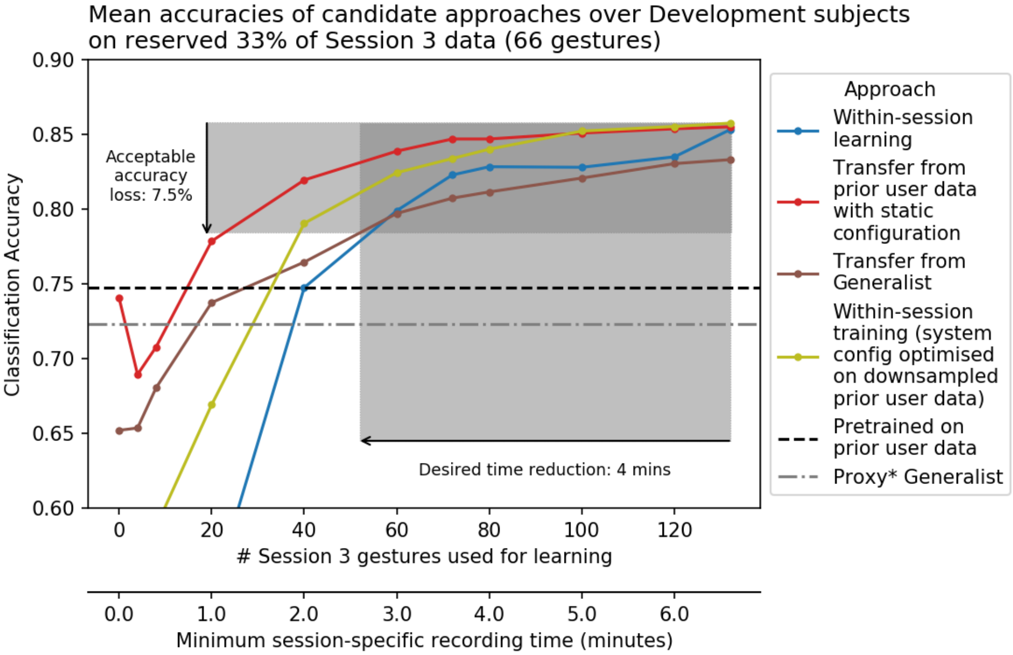
(b) The resultant loss in achievable accuracy if calibration time were so reduced

(c) The achievable reduction in calibration time while not exceeding the defined acceptable loss in accuracy

Figure 7.16: Tradeoffs in accuracy and calibration data burden



(a) A more stringent set of tradeoff criteria



(b) A more generous set of tradeoff criteria

Figure 7.17: Alternative tradeoff criteria

Conclusion

8.1 Findings & Contributions

This work has demonstrated the viability of using multimodal fusion of electromyographic and electroencephalographic data in the classification of same-hand gestures.

Chapter 5 investigated the efficacy of three distinct fusion architectures: a "Feature-Level" early fusion, a "Decision-Level" late fusion (including by stacked meta-models), and a novel "Hierarchical" approach which incorporates principles of both early and late fusion strategies. The abilities of classification systems implementing these architectures, each designed via Combined Algorithm Selection and Hyperparameter (CASH) optimisation, to accurately predict gestures from biosignal data were compared against one another and against "unimodal" classifiers modelled on only one data modality.

In a subject-specific classification scheme, wherein models were trained and tested on data belonging to single individuals, the proposed Hierarchical fusion architecture reached a mean accuracy of 88.90% across 20 subjects, surpassing that of any other approach. Performance remained high, at a mean of 86.25% accuracy, in validation of the CASH-optimised system by its training & testing on 5 unseen subjects whose data had not contribute to the selection of machine learning algorithms nor tuning of hyperparameters. The value of the CASH optimisation routine in determining the configuration of a biosignal classification system was also demonstrated. The CASH-identified Hierarchical system offered significantly higher classification accuracy ($p = 0.0079$) than the most suitable subject-specific EMG-EEG fusion design which could be gleaned solely from surveying the literature.

A Leave-One-Subject-Out training scheme was used to implement subject-independent models. In this setting all the proposed fusion architectures proved more accurate than equivalently-optimised unimodal systems. Subject-independent single-mode EEG models, while indeed outperformed by multimodal approaches, reached mean classification accuracies of up to 51.92% over held-out subjects. As 5.5.7 discusses this exceeds accuracies achieved in a number of previous works on multiclass EEG classification, despite many such works using subject-specific models [75, 128, 198]. While some works such as [64] have reported higher subject-independent accuracies (73%), this was for binary classification between movements of distinct body parts, a less complex problem than the multi-class hand gesture discrimination attempted here.

The use of CASH optimisation in designing classification systems — the methodology's first application in the field — enabled these comparisons to be fair, affording competing systems equivalent resource with which to establish suitable modelling choices, and unbiased by any cherry-picking or manual tuning of models, thereby addressing an established weakness among literature on Brain-Computer-Interfaces [30]. Further, the model configurations identified through this optimisation were validated through their application to a “Holdout” dataset consisting of subjects withheld from all stages of development. A standard of rigour rarely seen in the field [31], this ensured findings’ generalisability could be demonstrated & that reported classification accuracies were not unduly inflated by data leakage.

Chapter 5’s ancillary investigations illuminated the suitability of various specific modelling decisions in biosignal gesture classification. Detailed analysis of trends revealed by the CASH optimisation process provided a firm evidential underpinning for trends often “taken as read” among biosignal literature, such as the strength of the Linear Discriminant Analysis algorithm in classifying EEG data. Inspecting the informativity of different statistical features extracted from EEG data found predictive power in Delta-band activity in the ipsilateral motor cortex typically seen only with invasive measures such as electrocorticography.

Having demonstrated the viability of subject-independent classification, but found it routinely weaker than the subject-specific paradigm, Chapter 6 then investigated strategies for cross-subject learning. Both augmentation of a given subjects’ dataset with data of other individuals, and the use of transfer learning to directly adapt a model trained on others’ data to the subject, demonstrated an ability to classify an unseen portion of the subject’s data, with increasing accuracies as the quantity of subject-specific data was increased. Neither of these strategies however offered peak accuracy significantly different from that of a subject-specific system modelled on an equivalent level of the subject’s data. Though they did demonstrably achieve greater accuracies where subject-specific data was scarce, these were not significantly greater than those reached by a wholly subject-independent system. This implied future gesture classification systems to be better served by subject-specific modelling where significant quantities of subject data are available, and by subject-independent modelling where it is not; that calibration by the techniques explored here was not of benefit.

Finally Chapter 7 approached the problem from a cross-session perspective, somewhat rare among research but highly relevant to the context of prosthesis control. Given the importance of an accessibility device being itself accessible, there is a strong motivation to minimise the burden placed upon users by session-specific calibration. Here a number of potential approaches for such calibration were proposed and their ability to accurately classify a subject’s gestures when provided varying levels of session-specific data explored. Zero-calibration approaches were outperformed by those which utilised session-specific data in all but the most extreme cases where sufficiently little calibration data were available that such adaptation caused models to overfit. The use of either cross-subject data, or data collected from the same subject in previous sessions, as the source domain for a transfer learning process were much more robust to reductions in the session-specific data than models trained solely on said data. While performing such a domain transfer from a subject’s own data proved capable of more accurate classification than doing so from others’ data, both showed promising results. The latter’s viability is notable given its distinct use-case, and the savings in time, cost, and convenience which could be made for a user by avoiding the need to collect biosignal data from them in advance of a system’s deployment.

8.2 Recommendations for future work

A natural extension of Chapter 7's demonstration various per-session calibration strategies' viability is their trialling in experiments designed with an increased verisimilitude. Given the motivation of such approaches to positively impact users' experience of a system, there would be value in better relating offline experiments to real-world BCI operation [193]. In particular, the calibration data available to a system would logically be that collected earliest in its operation within a given session. In this work, calibration data were selected by a stratified sample — on the basis of whole gestures so as to avoid temporal data leakage [191,192] — in varying quantities. The properties of calibration data sampled at the beginning of a session ought not in principle to differ notably from that taken from random points throughout it as in the offline tests conducted here, but the randomness of this sampling may add some variability which would not be reflected in a real-world system.

The potential effect of fatigue may also warrant consideration. While controlled for by inclusion of rest periods in the dataset collected by Jeong et al. [198] used in this study, some works have induced fatigue in participants [139] or simulated in processing their muscular data [146] and found it to impact the quality of gesture classification. Drift between calibration data collected at the start of a session & fatigued data collected after a system's extensive use may impact the longevity of the calibration process's impact; future research intending to extend this work to better resemble real-world applications must not ignore this factor.

Assessing system's real-time performance would also be of benefit in enabling the translation of this research to prosthesis applications. Online classification accuracy is naturally the ideal test of this but work such as [223] has been able to emulate real-time classification in offline testing. The sliding time-window approach to feature extraction discussed in 4.3 essentially places a limit on the classification speed of a system; the 1-second windows with 50% overlap used here would result in a maximum prediction rate of two per second. While not a focus of this study, increasing the overlap between successive windows could plausibly lead to more frequent predictions, provided the underlying feature extraction & classification were sufficiently fast or parallelised [234]. This could be used to enable a "smoothing" of a prosthetic limb's actuation decisions: determining gesture intent as, for instance, a vote of 10 predictions each made 100 milliseconds apart (i.e. a 90% overlap between neighbouring windows) could plausibly mitigate the impact of individual misclassifications.

Real-time classification could further enable the extension of the per-session calibration strategies presented in Chapter 7 to a continuous adaptation paradigm. As discussed in 7.3.3.3, some algorithms such as the Gaussian Naïve Bayes classifier are suitable for incremental training; a mechanism could be implemented for on-the-fly "correction" of a system's predictions. In its simplest form this would involve a model updating when manually prompted by a user, though more advanced methods could include leveraging the EEG data for automatic detection of unexpected actuation as in [115], or retraining algorithms continually throughout a system's usage as in [175]. Such adaptation may even prove useful as a means of mitigating the aforementioned potential effect of temporal drift between calibration & test data.

The three types of grasp which defined this work's gesture classes are among the movements noted by [28] as most valuable to prosthesis users due to their frequent use in Activities of Daily Living. Though both biosignal literature and commercial prostheses frequently define only generic "hand close" and "hand open" gestures, the spherical, lateral, and cylindrical grasps are subtly but importantly distinct in their purposes and not easily interchangeable. It would be difficult to justify, for example, offering only a whole-hand cylindrical grasp and expecting it to meet a prosthesis user's needs in more delicate tasks. By virtue of being all right-hand movements they are not only more task-relevant but also fundamentally less distinct than the gestures used in many studies. The same muscle groups generally act in each, controlled by the same region of the motor cortex, thus making their classification a more difficult task than those of works wherein classes correspond to different body parts such as the right and left hands, or more dissimilar same-limb gestures such as the separation of wrist from elbow movements. The work could be extended however by also considering such types of movement. One avenue for doing so would be simply integrating them into the suite of available gestures, and attempting a wider multiclass problem. Perhaps more valuable, and better representative of a "real-world" use case, would be to embed a recognition of the fact that everyday human gestures are not often strictly delineated between individual movements of isolated body parts, but rather more combined. Consider for example the motions involved in opening a locked door. One could break this down as: grasp the key, rotate it, release the key, grasp the handle (with a different hand-shape than the key was held with), rotate it while pushing the door, release the handle, perform the same in reverse — but these would typically be done in a more fluid way. There might hence be merit in expanding a system not into a more complex multiclass problem but into one of multi-label classification, identifying in parallel whether for example a change in gesture state was required in the fingers and thumb, wrist, and even elbow[1], at a given point in time. From these movement components a prosthesis' resultant robotic response could be either selected from among pre-programmed multi-degree-of-freedom actuation routines, or even synthesised combinatorially. Having demonstrated a number of systems' viability for classifying similar same-limb gestures this work motivates further research in such a direction, to lead towards more naturalistic control for gesture-recognition based prostheses.

The acquisition and dissemination of more extensive multimodal biosignal datasets is paramount for furthering research in the domain. The limited availability of multiple-session data constrains the continued exploration of systems' robustness to changes in sensor fit, environmental conditions, and any longer term time-related data drift. Datasets of larger samples of participants are also necessary to ensure the quality and generalisability of future research. In this work to ensure sufficient data were available for modelling, only five subjects were reserved for the validation necessary to avoid the pitfalls of over-optimisation and data leakage common to Brain-Computer-Interface research [31]. Such a small population of this held-out dataset naturally limits the sensitivity of tests performed with it; 5.5.2.3 discussed the subject-wise variation in systems' classification accuracies, and as noted in 7.4.3 this may have affected the extent to which significant effects could be identified.

[1] In the case of transhumeral amputees — though the absence of residual forearm muscles may affect the fidelity of systems for this group.

Wider population samples may also open lines of investigation into cross-subject learning which were not viable here. As discussed in 6.3.1.1, some prior works have "screened" the data used to augment a subject-specific system by identified individuals whose biosignal data were similarly distributed [182,184,337]. Requiring sight of novel users' data to assess said similarity however precludes such a strategy from use in a context which minimises pre-deployment setup such as a wholly subject-independent system or one applying per-session calibration to a baseline "generalist" model as proposed in 7.3.3.4. While steps towards subject-independence could be of clear benefit, with the potential to reduce access barriers to users in terms of time, cost, and availability (given the lesser need for bespoke customisation of classifiers), it is no surprise that their reduced specificity routinely leads to lower accuracies. They may well have the most to gain from pre-selecting suitable data for augmentation but doing so with data-driven techniques evidently undermines their practical benefits. However, with sufficient quantity and diversity of data from which to draw in cross-subject learning, there may be alternative means of selection, such as by some measure of demographic similarity. Aging, for instance, is known not only to affect individuals' musculature physically, but also to modify the nature of the motor unit recruitment (see 2.1.1) involved in their movement [365]. Some work has similarly found differences in bioelectric signal properties between different genders [368]. Trends of this nature are of course relevant largely at the population-level, and are limited in nuance — for example, as is unfortunately the case in many fields the biosignal literature on gender-related differences rarely considers those whose gender does not fall under the binary categories of "male" and "female", nor the experience of binary transgender individuals who may or may not have received gender-affirming care[2]. Nevertheless, broad demographic trends could be used to provide some degree of tailoring — it is plausible for example that a child may see greater success in using a gesture classification system trained wholly or partially on data of other young people than of those advanced in age, and vice versa. Under a wholly subject-independent mode of deployment, one could then envisage a user being able to select an appropriate system from a range pre-trained on different age groups. Given the high variance between individuals however, proper investigation of such effects, let alone the development of "tailored" pre-trained systems, would require datasets of vastly higher participants numbers. Indeed this applies also to unimodal datasets — while some established public datasets such as various NinaPro EMG sets have up to 77 participants (including one dataset of over 10 upper-limb amputees) [226], many of the BCI Competition datasets considered gold-standard in the EEG domain contain fewer than 10, or in some cases fewer even than 5, subjects's data [194].

[2] Some work should be recognised here as having considered such matters, such as Künzel et al. in investigating the impact of Hormore Replacement Therapy on the sleep EEG of transgender women [369], & Hazin et al.'s exploration of EMG measured from pelvic floor muscles before & after gender-affirming surgery [370].

8.3 Concluding thoughts

Of course, the most important population to consider in translating the findings of this research into real-world advances in robotic prosthesis control applications is amputees themselves. As exploratory research this work was deliberate in its choice to use data collected from individuals without upper-limb differences; 4.1.1.1 discussed the motivation for this decision and indeed the evidence from Scheme & Englehart for trends found with non-amputee data to be reflected in that of amputees.

Nevertheless before systems evidenced here as viable could be put into use they would need further exploration no only using amputees' data but in conjunction with them as meaningful contributors to development. Individuals with lived experience of prosthesis use are naturally best-informed as to their desirable characteristics. As has been made clear throughout this thesis no attempt is made here to speak for amputees, nor suggest its contributions to the field of gesture classification as having necessarily guaranteed some sweeping transformation to their lives. The true potential impact of findings beyond academia could only ever be fully judged by amputees who use gesture-recognition prostheses; "Nothing About Us Without Us" is a well-established philosophy in the disability rights movement [371] and while the author is themselves disabled, as a non-amputee their understanding of amputees' needs is necessarily incomplete.

Perhaps the most significant aspect of the systems explored in this work on which better research of amputees' opinions is needed is actually one of the most fundamental: the merits of using neural data measured by Electroencephalography to supplement traditionally Electromyographic gesture recognition. Certainly the motivation for investigating the approach is clear — as has been discussed earlier in the work the capabilities of EMG-based prostheses are often limited, offering low dexterity due to the challenges of reliable gesture recognition. In particular, that amputees have varying levels of precision in the voluntary control of residual limb muscles, yet can perform Kinaesthetic Motor Imagery of the amputated hand, makes EEG as a non-invasive means of measuring the motor cortex an appealing option. Further, Chapter 5's evidence that inclusion of EEG data can lead to higher subject-independent accuracy than the use of EMG alone demonstrates the potential of the technique to lead to more accessible systems of lower burden to users in terms of time and mass-availability. EEG itself however is not without inconvenience. Changes in environmental conditions can affect the quality of data, and the technology involved in its measurement (though rapidly advancing with recent developments in high-density configurations [372], sensors which avoid the cost and inconvenience of conductive electrolyte gel [373], and electrodes better-suited for individuals with natural, coarse, and curly hair [374, 375]) is not inexpensive. Further analysis of the costs and benefits of multimodality in biosignal classification, while outside the scope of this work, will certainly be of great merit — and must as discussed be conducted in conjunction with prosthesis users and indeed with amputees who have not sought or been able to acquire prostheses due to systemic ableism & the resultant structural or financial barriers placed upon them.

Bibliography

[1] J. J. Bird, M. Pritchard, A. Fratini, A. Ekárt, and D. R. Faria, "Synthetic biological signals machine-generated by gpt-2 improve the classification of eeg and emg through data augmentation," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3498–3504, Feb. 2021, © 2021 IEEE. Reprinted, with permission, from J. J. Bird, M. Pritchard (co-first authors), A. Fratini, A. Ekárt, and D. R. Faria, Synthetic Biological Signals Machine-Generated by GPT-2 Improve the Classification of EEG and EMG Through Data Augmentation, IEEE Robotics and Automation Letters, February 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9345373

[2] K. Ziegler-Graham, E. J. MacKenzie, P. L. Ephraim, T. G. Travison, and R. Brookmeyer, "Estimating the prevalence of limb loss in the united states: 2005 to 2050," *Arch. Phys. Med. Rehabil.*, vol. 89, no. 3, pp. 422–429, Mar. 2008. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/18295618/

[3] S. Katz, "Assessing self-maintenance: activities of daily living, mobility, and instrumental activities of daily living," *Journal of the American Geriatrics Society*, vol. 31, no. 12, pp. 721–727, Dec. 1983. [Online]. Available: https://agsjournals.onlinelibrary.wiley.com/doi/10.1111/j.1532-5415.1983.tb03391.x

[4] P. F. Edemekong, D. L. Bomgaars, S. Sukumaran, and C. Schoo, *Activities of Daily Living.* StatPearls Publishing, Jun. 2023.

[5] P. Montoya, K. Ritter, E. Huse, W. Larbig, C. Braun, S. Töpfner, W. Lutzenberger, W. Grodd, H. Flor, and N. Birbaumer, "The cortical somatotopic map and phantom phenomena in subjects with congenital limb atrophy and traumatic amputees with phantom limb pain," *European Journal of Neuroscience*, vol. 10, no. 3, pp. 1095–1102, 1998. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1046/j.1460-9568.1998.00122.x

[6] P. Hernigou, "Ambroise paré iv: The early history of artificial limbs (from robotic to prostheses)," *International Orthopaedics*, vol. 37, no. 6, pp. 1195–1197, Jun. 2013. [Online]. Available: https://doi.org/10.1007/s00264-013-1884-7

[7] W. H. Weihe, "In memoriam reinhold reiter 17 november 1920–24 september 1998," *International Journal of Biometeorology*, vol. 43, no. 2, pp. 96–98, Oct. 1999. [Online]. Available: https://doi.org/10.1007/s004840050122

[8] Otto Bock Health Care LP, "Fascinated. with michelangelo – perfect use of precision technology," 2014. [Online]. Available: https://www.ottobockus.com/media/local-media/prosthetics/upper-limb/michelangelo/files/michelangelo-brochure.pdf

[9] M. Jagannathan, "This duo wants to do for prosthetics what fashion designers did for eyeglasses," Jul. 2018. [Online]. Available: https://www.marketwatch.com/story/this-duo-wants-to-do-for-prosthetics-what-fashion-designers-did-for-eyeglasses-2018-07-06-12883815

[10] NHS England. (2022, Nov.) Nhs offers life-changing bionic arms to all amputees. [Online]. Available: https://www.england.nhs.uk/2022/11/nhs-offers-life-changing-bionic-arms-to-all-amputees/

[11] Open Bionics, *Hero Arm User Guide - Version 100583_01_0*, Feb. 2019. [Online]. Available: https://openbionics.com/hero-arm-user-guide/

[12] S. M. Wurth and L. J. Hargrove, "A real-time comparison between direct control, sequential pattern recognition control and simultaneous pattern recognition control using a fitts' law style assessment procedure," *Journal of NeuroEngineering and Rehabilitation*, vol. 11, no. 1, p. 91, May 2014. [Online]. Available: https://jneuroengrehab.biomedcentral.com/articles/10.1186/1743-0003-11-91

[13] E. Campbell, A. Phinyomark, A. H. Al-Timemy, R. N. Khushaba, G. Petri, and E. Scheme, "Differences in emg feature space between able-bodied and amputee subjects for myoelectric control," in *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, 2019, pp. 33–36.

[14] E. Scheme and K. Englehart, "Electromyogram pattern recognition for control of powered upper-limb prostheses: state of the art and challenges for clinical use," *J. Rehabil. Res. Dev.*, vol. 48, no. 6, pp. 643–659, 2011. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/21938652/

[15] T. A. Kuiken, L. A. Miller, K. Turner, and L. J. Hargrove, "A comparison of pattern recognition control and direct control of a multiple degree-of-freedom transradial prosthesis," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 4, pp. 1–8, 2016. [Online]. Available: https://ieeexplore.ieee.org/document/7752831

[16] L. Resnik, H. H. Huang, A. Winslow, D. L. Crouch, F. Zhang, and N. Wolk, "Evaluation of emg pattern recognition for upper limb prosthesis control: a case study in comparison with direct myoelectric control," *Journal of NeuroEngineering and Rehabilitation*, vol. 15, no. 1, p. 23, Mar. 2018. [Online]. Available: https://jneuroengrehab.biomedcentral.com/articles/10.1186/s12984-018-0361-3

[17] Esper Bionics, "Esper bionics," 2023, accessed: December 2023. [Online]. Available: https://esperbionics.com/

[18] L. Dickstein, "Esper hand: The 200 best inventions of 2022," *TIME Magazine*, Nov. 2022. [Online]. Available: https://time.com/collection/best-inventions-2022/6228818/esper-hand/

[19] J. Dubernard, P. Henry, H. Parmentier, B. Vallet, D. Vial, L. Badet, P. Petruzzo, N. Lefrançois, M. Lanzetta, E. Owen, and N. Hakim, “Première transplantation des deux mains : résultats à 18 mois [first double hand transplantation: results after 18 months],” *Annales de Chirurgie*, vol. 127, no. 1, pp. 19 – 25, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S000339440100668X

[20] G. Brandacher, M. Ninkovic, H. Piza-Katzer, M. Gabl, H. Hussl, M. Rieger, M. Schocke, K. Egger, W. Loescher, B. Zelger, M. Ninkovic, H. Bonatti, C. Boesmueller, W. Mark, R. Margreiter, and S. Schneeberger, “The innsbruck hand transplant program: Update at 8 years after the first transplant,” *Transplantation Proceedings*, vol. 41, no. 2, pp. 491 – 494, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0041134509000141

[21] M. Bumbaširević, A. Lesic, T. Palibrk, D. Milovanovic, M. Zoka, T. Kravić-Stevović, and S. Raspopovic, “The current state of bionic limbs from the surgeon’s viewpoint,” *EFORT Open Reviews*, vol. 5, no. 2, pp. 65–72, 2020. [Online]. Available: https://doi.org/10.1302/2058-5241.5.180038

[22] E&T Editorial Staff, “Mind-controlled bionic arm with sense of touch ‘could be available in two years’,” *E&T Magazine*, Apr. 2020. [Online]. Available: https://eandt.theiet.org/content/articles/2020/04/mind-controlled-bionic-arm-with-sense-of-touch-could-be-available-in-two-years/

[23] J. Loughran, “Prosthetic limb brain interface could enable finger movement for amputees,” *E&T Magazine*, Mar. 2020. [Online]. Available: https://eandt.theiet.org/content/articles/2020/03/prosthetic-limb-with-individual-finger-movement-made-possible-with-brain-interface/

[24] E. A. Biddiss and T. T. Chau, “Upper limb prosthesis use and abandonment: a survey of the last 25 years,” *Prosthetics and Orthotics International*, vol. 31, no. 3, pp. 236–257, Sep. 2007. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/17979010/

[25] E. Biddiss, D. Beaton, and T. Chau, “Consumer design priorities for upper limb prosthetics,” *Disability and Rehabilitation: Assistive Technology*, vol. 2, no. 6, pp. 346–357, Nov. 2007. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/19263565/

[26] P. J. Kyberd and W. Hill, “Survey of upper limb prosthesis users in sweden, the united kingdom and canada,” *Prosthetics and Orthotics International*, vol. 35, no. 2, pp. 234–241, 2011. [Online]. Available: https://journals.sagepub.com/doi/10.1177/0309364611409099

[27] S. M. Engdahl, B. P. Christie, B. Kelly, A. Davis, C. A. Chestek, and D. H. Gates, “Surveying the interest of individuals with upper limb loss in novel prosthetic control techniques,” *Journal of NeuroEngineering and Rehabilitation*, vol. 12, no. 1, p. 53, Jun. 2015. [Online]. Available: https://jneuroengrehab.biomedcentral.com/track/pdf/10.1186/s12984-015-0044-2.pdf

[28] F. Cordella, A. L. Ciancio, R. Sacchetti, A. Davalli, A. G. Cutti, E. Guglielmelli, and L. Zollo, “Literature review on needs of upper limb prosthesis users,” *Frontiers in Neuroscience*, vol. 10, 2016. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2016.00209

[29] I. S. Engagement and P. Editorial, “How can engineers and technologists involve patients in the design process of new solutions for healthcare?” Feb. 2021. [Online]. Available: https://communities.theiet.org/blogs/822/7162

[30] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, “A review of classification algorithms for eeg-based brain–computer interfaces,” *Journal of Neural Engineering*, vol. 4, no. 2, p. R1, Jan. 2007. [Online]. Available: https://dx.doi.org/10.1088/1741-2560/4/2/R01

[31] M. Hosseini, M. Powell, J. Collins, C. Callahan-Flintoft, W. Jones, H. Bowman, and B. Wyble, “I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data,” *Neuroscience & Biobehavioral Reviews*, vol. 119, pp. 456–467, Dec. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0149763420305868

[32] H. Gray, *Anatomy of the human body*, 20th ed. Philadelphia, Pennsylvania, USA: Lea & Febiger, 1918, illus. H. V. Carter, rev. W. H. Lewis. [Online]. Available: http://www.bartleby.com/107/

[33] Blausen.com staff. (2014, Mar.) Medical gallery of Blausen Medical 2014. CC 3.0 Unported. [Online]. Available: https://commons.wikimedia.org/wiki/File:Blausen_0103_Brain_Sensory%26Motor.png

[34] J. G. Betts, K. A. Young, J. A. Wise, E. Johnson, B. Poe, D. H. Kruse, O. Korol, J. E. Johnson, M. Womble, and P. DeSaix, *Anatomy and Physiology.* OpenStax, Apr. 2013, ch. 14.2, cC 4.0 International. [Online]. Available: https://openstax.org/books/anatomy-and-physiology/pages/14-2-central-processing

[35] L. M. McPherson, F. Negro, C. K. Thompson, L. Sanchez, C. J. Heckman, J. Dewald, and D. Farina, “Properties of the motor unit action potential shape in proximal and distal muscles of the upper limb in healthy and post-stroke individuals,” in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016, pp. 335–339. [Online]. Available: https://ieeexplore.ieee.org/document/7590708

[36] I. Rodríguez-Carreño, L. Gila-Useros, and A. Malanda-Trigueros, “Motor unit action potential duration: Measurement and significance,” in *Advances in Clinical Neurophysiology*, I. M. Ajeena, Ed. Rijeka: IntechOpen, Oct. 2012, ch. 7. [Online]. Available: https://www.intechopen.com/chapters/40104

[37] C. Katsis, Y. Goletsis, A. Likas, D. Fotiadis, and I. Sarmas, “A novel method for automated emg decomposition and muap classification,” *Artificial Intelligence in Medicine*, vol. 37, no. 1, pp. 55–64, 2006, intelligent Data Analysis in Medicine. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365705001065

[38] I. Kuzborskij, A. Gijsberts, and B. Caputo, “On the challenge of classifying 52 hand movements from surface electromyography,” in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2012, pp. 4931–4937. [Online]. Available: https://ieeexplore.ieee.org/document/6347099

[39] S. H. Jaffer and N. H. Ghaeb, "Important features of emg signal under simple load conditions," *Journal of Polytechnic*, vol. 7, pp. 1–01, 2017. [Online]. Available: https://www.researchgate.net/publication/317605481_Important_features_of_EMG_signal_under_simple_load_conditions

[40] T. R. Farrell and R. F. Weir, "Pilot comparison of surface vs. implanted emg for multifunctional prosthesis control," in *9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005.*, 2005, pp. 277–280. [Online]. Available: https://dukespace.lib.duke.edu/dspace/handle/10161/2706

[41] P. R. Troyk, G. A. DeMichele, D. A. Kerns, and R. F. Weir, "Imes: An implantable myoelectric sensor," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2007, pp. 1730–1733. [Online]. Available: https://ieeexplore.ieee.org/document/4352644

[42] R. F. Weir, P. R. Troyk, G. A. DeMichele, D. A. Kerns, J. F. Schorsch, and H. Maas, "Implantable myoelectric sensors (imess) for intramuscular electromyogram recording," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 1, pp. 159–171, 2009. [Online]. Available: https://ieeexplore.ieee.org/document/4633666

[43] A. Waris, I. K. Niazi, M. Jamil, O. Gilani, K. Englehart, W. Jensen, M. Shafique, and E. N. Kamavuako, "The effect of time on emg classification of hand motions in able-bodied and transradial amputees," *Journal of Electromyography and Kinesiology*, vol. 40, pp. 72–80, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1050641117304790

[44] C. Pylatiuk, M. Muller-Riederer, A. Kargov, S. Schulz, O. Schill, M. Reischl, and G. Bretthauer, "Comparison of surface emg monitoring electrodes for long-term use in rehabilitation device control," in *2009 IEEE International Conference on Rehabilitation Robotics*, 2009, pp. 300–304. [Online]. Available: https://ieeexplore.ieee.org/document/5209576

[45] Thalmic Labs, Inc, "Tech specs: Myo battery life, dimensions, compatibility, and more," Web Archive, 2016. [Online]. Available: https://web.archive.org/web/20170701185826/https://www.myo.com/\techspecs

[46] P. Visconti, F. Gaetani, G. Zappatore, and P. P. and, "Technical features and functionalities of myo armband: An overview on related literature and advanced applications of myoelectric armbands mainly focused on arm prostheses," *International Journal on Smart Sensing and Intelligent Systems*, vol. 11, no. 1178-5608, pp. 1–25, 2018. [Online]. Available: https://www.exeley.com/in_jour_smart_sensing_and_intelligent_systems/doi/10.21307/ijssis-2018-005

[47] J. G. Abreu, J. M. Teixeira, L. S. Figueiredo, and V. Teichrieb, "Evaluating sign language recognition using the myo armband," in *2016 XVIII Symposium on Virtual and Augmented Reality (SVR)*, 2016, pp. 64–70.

[48] A. B. H. Amor, O. Ghoul, and M. Jemni, "Toward sign language handshapes recognition using myo armband," in *2017 6th International Conference on Information and Communication Technology and Accessibility (ICTA)*, 2017, pp. 1–6.

[49] S. He, C. Yang, M. Wang, L. Cheng, and Z. Hu, “Hand gesture recognition using myo armband,” in *2017 Chinese Automation Congress (CAC)*, 2017, pp. 4850–4855.

[50] E. Kaya and T. Kumbasar, “Hand gesture recognition systems with the wearable myo armband,” in *2018 6th International Conference on Control Engineering Information Technology (CEIT)*, 10 2018.

[51] *Thalmic Labs Myo armband allows amputee to control prosthetic limb*, ser. Cantech Letter, 2016. [Online]. Available: https://www.cantechletter.com/2016/01/thalmic-labs-myo-armband-allows-amputee-to-control-prosthetic-limb/

[52] S. Pitou, F. Wu, A. Shafti, B. Michael, R. Stopforth, and M. Howard, “Embroidered electrodes for control of affordable myoelectric prostheses,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 1812–1817. [Online]. Available: https://ieeexplore.ieee.org/document/8461066

[53] P. Zipp, “Recommendations for the standardization of lead positions in surface electromyography,” *European Journal of Applied Physiology and Occupational Physiology*, vol. 50, no. 1, pp. 41–54, Feb. 1982. [Online]. Available: https://doi.org/10.1007/BF00952243

[54] E.-P. Takala and R. Toivonen, “Placement of forearm surface EMG electrodes in the assessment of hand loading in manual tasks,” *Ergonomics*, vol. 56, no. 7, pp. 1159–1166, 2013. [Online]. Available: https://doi.org/10.1080/00140139.2013.799235

[55] A. L. Hof, “The relationship between electromyogram and muscle force,” *Sportverletzung Sportschaden : Organ der Gesellschaft für Orthopädisch-Traumatologische Sportmedizin*, vol. 11, pp. 79–86, 10 1997. [Online]. Available: https://www.researchgate.net/publication/13878512_The_relationship_between_electromyogram_and_muscle_force

[56] E. Scheme and K. Englehart, “Training strategies for mitigating the effect of proportional control on classification in pattern recognition–based myoelectric control,” *JPO: Journal of Prosthetics and Orthotics*, vol. 25, no. 2, 2013. [Online]. Available: https://journals.lww.com/jpojournal/Fulltext/2013/04000/Training_Strategies_for_Mitigating_the_Effect_of.4.aspx

[57] M. Seyedali, J. M. Czerniecki, D. C. Morgenroth, and M. E. Hahn, “Co-contraction patterns of trans-tibial amputee ankle and knee musculature during gait,” *Journal of NeuroEngineering and Rehabilitation*, vol. 9, no. 1, p. 29, May 2012. [Online]. Available: https://jneuroengrehab.biomedcentral.com/articles/10.1186/1743-0003-9-29

[58] H. Yuan and B. He, “Brain–computer interfaces using sensorimotor rhythms: Current state and future perspectives,” *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 5, pp. 1425–1435, 2014. [Online]. Available: https://ieeexplore.ieee.org/document/6775293

[59] K. Jerbi, J.-P. Lachaux, K. N’Diaye, D. Pantazis, R. M. Leahy, L. Garnero, and S. Baillet, “Coherent neural representation of hand speed in humans revealed by meg imaging,” *Proceedings*

of the National Academy of Sciences, vol. 104, no. 18, pp. 7676–7681, 2007. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.0609632104

[60] J. Gross, L. Timmermann, J. Kujala, M. Dirks, F. Schmitz, R. Salmelin, and A. Schnitzler, “The neural basis of intermittent motor control in humans,” *Proceedings of the National Academy of Sciences*, vol. 99, no. 4, pp. 2299–2302, 2002. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.032682099

[61] D. J. Lee, E. Kulubya, P. Goldin, A. Goodarzi, and F. Girgis, “Review of the Neural Oscillations Underlying Meditation,” *Frontiers in Neuroscience*, vol. 12, 2018.

[62] R. Ferri, F. I. I. Cosentino, M. Elia, S. A. Musumeci, R. Marinig, and P. Bergonzi, “Relationship between delta, sigma, beta, and gamma EEG bands at REM sleep onset and REM sleep end,” *Clinical Neurophysiology*, vol. 112, no. 11, pp. 2046 – 2052, 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1388245701006563

[63] A. E. Symons, W. El-Deredy, M. Schwartze, and S. A. Kotz, “The functional role of neural oscillations in non-verbal emotional communication,” *Frontiers in human neuroscience*, vol. 10, pp. 239–239, May 2016. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/27252638

[64] S. Fazli, F. Popescu, M. Danóczy, B. Blankertz, K.-R. Müller, and C. Grozea, “Subject-independent mental state classification in single trials,” *Neural Networks*, vol. 22, no. 9, pp. 1305–1312, 2009, brain-Machine Interface. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608009001063

[65] E. Başar, M. Schürmann, C. Başar-Eroglu, and S. Karakaş, “Alpha oscillations in brain functioning: an integrative theory,” *International Journal of Psychophysiology*, vol. 26, no. 1, pp. 5–29, Jun. 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167876097007538

[66] G. G. Knyazev, “Motivation, emotion, and their inhibitory control mirrored in brain oscillations,” *Neuroscience & Biobehavioral Reviews*, vol. 31, no. 3, pp. 377 – 395, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0149763406001163

[67] M. Steriade, “Corticothalamic resonance, states of vigilance and mentation,” *Neuroscience*, vol. 101, no. 2, pp. 243–276, 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306452200003535

[68] R. Huber, M. Felice Ghilardi, M. Massimini, and G. Tononi, “Local sleep and learning,” *Nature*, vol. 430, no. 6995, pp. 78–81, Jul. 2004. [Online]. Available: https://doi.org/10.1038/nature02663

[69] V. Nácher, A. Ledberg, G. Deco, and R. Romo, “Coherent delta-band oscillations between cortical areas correlate with decision making,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 37, pp. 15 085–15 090, Aug. 2013. [Online]. Available: https://www.pnas.org/doi/10.1073/pnas.1314681110

[70] A. K. Engel and P. Fries, "Beta-band oscillations—signalling the status quo?" *Current Opinion in Neurobiology*, vol. 20, no. 2, pp. 156 – 165, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0959438810000395

[71] D. L. Pritchett, J. H. Siegle, C. A. Deister, and C. I. Moore, "For things needing your attention: the role of neocortical gamma in sensory perception," *Current Opinion in Neurobiology*, vol. 31, pp. 254 – 263, Apr. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0959438815000343

[72] R. Salmelin, M. Hámáaláinen, M. Kajola, and R. Hari, "Functional segregation of movement-related rhythmic activity in the human brain," *NeuroImage*, vol. 2, no. 4, pp. 237–243, 1995. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811985710312

[73] G. Pfurtscheller and F. Lopes da Silva, "Event-related eeg/meg synchronization and desynchronization: basic principles," *Clinical Neurophysiology*, vol. 110, no. 11, pp. 1842–1857, 1999. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1388245799001418

[74] H. Toriyama, J. Ushiba, and J. Ushiyama, "Subjective vividness of kinesthetic motor imagery is associated with the similarity in magnitude of sensorimotor event-related desynchronization between motor execution and motor imagery," *Frontiers in Human Neuroscience*, vol. 12, p. 295, 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnhum.2018.00295

[75] P. Ofner, A. Schwarz, J. Pereira, and G. R. Müller-Putz, "Upper limb movements can be decoded from the time-domain of low-frequency eeg," *PLOS ONE*, vol. 12, no. 8, pp. 1–24, 08 2017. [Online]. Available: https://doi.org/10.1371/journal.pone.0182578

[76] A. Guillot, F. Lebon, D. Rouffet, S. Champely, J. Doyon, and C. Collet, "Muscular responses during motor imagery as a function of muscle contraction types," *International Journal of Psychophysiology*, vol. 66, no. 1, pp. 18–27, Oct. 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167876007001201

[77] G. Schalk, J. Kubánek, K. J. Miller, N. R. Anderson, E. C. Leuthardt, J. G. Ojemann, D. Limbrick, D. Moran, L. A. Gerhardt, and J. R. Wolpaw, "Decoding two-dimensional movement trajectories using electrocorticographic signals in humans," *Journal of Neural Engineering*, vol. 4, no. 3, p. 264, Jun. 2007. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2560/4/3/012

[78] W. Shain, L. Spataro, J. Dilgen, K. Haverstick, S. Retterer, M. Isaacson, M. Saltzman, and J. Turner, "Controlling cellular reactive responses around neural prosthetic devices using peripheral and local intervention strategies," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, no. 2, pp. 186–188, Jun. 2003. [Online]. Available: https://ieeexplore.ieee.org/document/1214717

[79] D. Szarowski, M. Andersen, S. Retterer, A. Spence, M. Isaacson, H. Craighead, J. Turner, and W. Shain, "Brain responses to micro-machined silicon devices," *Brain Research*, vol. 983, no. 1, pp. 23–35, Sep. 2003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0006899303030233

[80] J. Kubánek, K. J. Miller, J. G. Ojemann, J. R. Wolpaw, and G. Schalk, "Decoding flexion of individual fingers using electrocorticographic signals in humans," *Journal of Neural Engineering*, vol. 6, no. 6, p. 066001, Oct. 2009. [Online]. Available: https://doi.org/10.1088%2F1741-2560%2F6%2F6%2F066001

[81] T. Pistohl, T. Ball, A. Schulze-Bonhage, A. Aertsen, and C. Mehring, "Prediction of arm movement trajectories from ecog-recordings in humans," *Journal of Neuroscience Methods*, vol. 167, no. 1, pp. 105 – 114, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165027007004840

[82] J. L. O'Leary, "Discoverer of the brain wave: <i>hans berger on the electroencephalogram of man</i>. the fourteen original reports on the human electroencephalogram. translated from the german and edited by pierre gloor. elsevier, new york, 1969. xii + 350 pp., illus. $31.50. electroencephalography and clinical neurophysiology, supplement 28." *Science*, vol. 168, no. 3931, pp. 562–563, May 1970. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.168.3931.562

[83] N. Jamil, A. N. Belkacem, S. Ouhbi, and A. Lakas, "Noninvasive electroencephalography equipment for assistive, adaptive, and rehabilitative brain–computer interfaces: A systematic literature review," *Sensors*, vol. 21, no. 14, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/14/4754

[84] H. H. Jasper, "The ten-twenty electrode system of the international federation," *Electroencephalogr. Clin. Neurophysiol.*, vol. 10, pp. 370–375, 1958.

[85] E. L. G. E. Chatrian and P. L. Nelson, "Ten percent electrode system for topographic studies of spontaneous and evoked eeg activities," *American Journal of EEG Technology*, vol. 25, no. 2, pp. 83–92, 1985. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/00029238.1985.11080163

[86] R. Oostenveld, "High-density eeg electrode placement," https://robertoostenveld.nl/electrode/#:~:text=The%205%25%20or%20the%2010,of%20the%2010%2D10%20system., Jan. 2006.

[87] G. H. Klem, H. O. Lüders, H. H. Jasper, and C. Elger, "The ten-twenty electrode system of the international federation," *Recommendations for the Practice of Clinical Neurophysiology: Guidelines of the International Federation of Clinical Physiology (EEG Supplement)*, vol. 52, pp. 3–6, 1999. [Online]. Available: https://pdfs.semanticscholar.org/53a7/\cf6bf8568c660240c080125e55836d507098.pdf

[88] R. Oostenveld and P. Praamstra, "The five percent electrode system for high-resolution EEG and ERP measurements," *Clinical Neurophysiology*, vol. 112, no. 4, pp. 713–719, Apr. 2001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1388245700005277

[89] V. Jurcak, D. Tsuzuki, and I. Dan, "10/20, 10/10, and 10/5 systems revisited: Their validity as relative head-surface-based positioning systems," *NeuroImage*, vol. 34, no. 4, pp. 1600–1611, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811906009724

[90] T. Warbrick, "Simultaneous EEG-fMRI: What have we learned and what does the future hold?" *Sensors (Basel)*, vol. 22, no. 6, Mar. 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8952790/

[91] B. Burle, L. Spieser, C. Roger, L. Casini, T. Hasbroucq, and F. Vidal, "Spatial and temporal resolutions of EEG: Is it really black and white? a scalp current density view," *Int. J. Psychophysiol.*, vol. 97, no. 3, pp. 210–220, Sep. 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4548479/

[92] C. M. Michel, M. M. Murray, G. Lantz, S. Gonzalez, L. Spinelli, and R. Grave de Peralta, "Eeg source imaging," *Clinical Neurophysiology*, vol. 115, no. 10, pp. 2195–2222, Oct. 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1388245704002135

[93] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-r. Muller, "Optimizing spatial filters for robust eeg single-trial analysis," *IEEE Signal Processing Magazine*, vol. 25, no. 1, pp. 41–56, 2008. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/4408441

[94] P. L. Nunez, R. Srinivasan, A. F. Westdorp, R. S. Wijesinghe, D. M. Tucker, R. B. Silberstein, and P. J. Cadusch, "Eeg coherency: I: statistics, reference electrode, volume conduction, laplacians, cortical imaging, and interpretation at multiple scales," *Electroencephalography and Clinical Neurophysiology*, vol. 103, no. 5, pp. 499–515, Nov. 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0013469497000667

[95] R. Grech, T. Cassar, J. Muscat, K. P. Camilleri, S. G. Fabri, M. Zervakis, P. Xanthopoulos, V. Sakkalis, and B. Vanrumste, "Review on solving the inverse problem in eeg source analysis," *Journal of NeuroEngineering and Rehabilitation*, vol. 5, no. 1, p. 25, Nov. 2008. [Online]. Available: https://jneuroengrehab.biomedcentral.com/articles/10.1186/1743-0003-5-25

[96] S. Dalal, S. Rampp, F. Willomitzer, and S. Ettl, "Consequences of eeg electrode position error on ultimate beamformer source reconstruction performance," *Frontiers in Neuroscience*, vol. 8, p. 42, 2014. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2014.00042

[97] B. U. Westner, S. S. Dalal, A. Gramfort, V. Litvak, J. C. Mosher, R. Oostenveld, and J.-M. Schoffelen, "A unified view on beamformers for m/eeg source reconstruction," *NeuroImage*, vol. 246, p. 118789, Feb. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811921010612

[98] C. M. Michel and D. Brunet, "Eeg source imaging: A practical review of the analysis steps," *Frontiers in Neurology*, vol. 10, Apr. 2019. [Online]. Available: https://www.frontiersin.org/journals/neurology/articles/10.3389/fneur.2019.00325

[99] M. E. Gunduz, C. B. Pinto, F. G. Saleh Velez, D. Duarte, K. Pacheco-Barrios, F. Lopes, and F. Fregni, "Motor cortex reorganization in limb amputation: A systematic review of TMS motor mapping studies," *Frontiers in Neuroscience*, vol. 14, p. 314, Apr. 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7187753/

[100] A. Pascual-Leone, M. Peris, J. M. Tormos, A. P. Pascual, and M. D. Catalá, "Reorganization of human cortical motor output maps following traumatic forearm amputation," *NeuroReport*, vol. 7, no. 13, pp. 2068–2070, Sep. 1996. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/8930960/

[101] M. Lotze, H. Flor, W. Grodd, W. Larbig, and N. Birbaumer, "Phantom movements and pain. an fMRI study in upper limb amputees," *Brain*, vol. 124, no. Pt 11, pp. 2268–2277, Nov. 2001. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/11673327/

[102] T. R. Makin, J. Scholz, N. Filippini, D. Henderson Slater, I. Tracey, and H. Johansen-Berg, "Phantom pain is associated with preserved structure and function in the former hand area," *Nature Communications*, vol. 4, p. 1570, Mar. 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3615341/

[103] T. R. Makin and H. Flor, "Brain (re)organisation following amputation: Implications for phantom limb pain," *Neuroimage*, vol. 218, no. 116943, p. 116943, Sep. 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7422832/

[104] M. Gagné, S. Hétu, K. T. Reilly, and C. Mercier, "The map is not the territory: motor system reorganization in upper limb amputees," *Hum. Brain Mapp.*, vol. 32, no. 4, pp. 509–519, Apr. 2011. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/21391244/

[105] G. Pfurtscheller, B. Allison, G. Bauernfeind, C. Brunner, T. Solis Escalante, R. Scherer, T. Zander, G. Mueller-Putz, C. Neuper, and N. Birbaumer, "The hybrid bci," *Frontiers in Neuroscience*, vol. 4, 2010. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnpro.2010.00003

[106] A. Gijsberts, M. Atzori, C. Castellini, H. Müller, and B. Caputo, "Movement error rate for evaluation of machine learning methods for semg-based hand movement classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 4, pp. 735–744, 2014.

[107] E. Rocon, J. Gallego, L. Barrios, A. Victoria, J. Ibánez, D. Farina, F. Negro, J. Dideriksen, S. Conforto, T. D'Alessio, G. Severini, J. Belda-Lois, L. Popovic, G. Grimaldi, M. Manto, and J. Pons, "Multimodal bci-mediated fes suppression of pathological tremor," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, 2010, pp. 3337–3340. [Online]. Available: https://ieeexplore.ieee.org/document/5627914

[108] T. D. Lalitharatne, K. Teramoto, Y. Hayashi, and K. Kiguchi, "Towards hybrid eeg-emg-based control approaches to be used in bio-robotics applications: Current status, challenges and future directions," *Paladyn, Journal of Behavioral Robotics*, vol. 4, no. 2, pp. 147–154, 2013. [Online]. Available: https://www.degruyter.com/document/doi/10.2478/pjbr-2013-0009/html

[109] A. Sarasola-Sanz, N. Irastorza-Landa, E. López-Larraz, C. Bibián, F. Helmhold, D. Broetz, N. Birbaumer, and A. Ramos-Murguialday, "A hybrid brain-machine interface based on eeg and emg activity for the motor rehabilitation of stroke patients," in *2017 International Conference on Rehabilitation Robotics (ICORR)*, 2017, pp. 895–900. [Online]. Available: https://ieeexplore.ieee.org/document/8009362

[110] A. H. Khan, I. N. Khan, and M. A. R. Sarkar, “Development of a prosthetic hand operated by eeg brain signals and emg muscle signals,” *International Journal of Control Theory and Applications*, vol. 8, no. 3, pp. 941–948, Dec. 2015. [Online]. Available: https://www.researchgate.net/publication/323918367_Development_of_a_Prosthetic_Hand_Operated_by_EEG_Brain_Signals_and_EMG_Muscle_Signals

[111] Y. Du, X. Zhang, Y. Wang, and T. Mu, “Design on exoskeleton robot intellisense system based on multi-dimensional information fusion,” in *2012 IEEE International Conference on Mechatronics and Automation*, 2012, pp. 2435–2439. [Online]. Available: https://ieeexplore.ieee.org/document/6285727

[112] N. Hooda, R. Das, and N. Kumar, “Fusion of eeg and emg signals for classification of unilateral foot movements,” *Biomedical Signal Processing and Control*, vol. 60, p. 101990, Jul. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809420301464

[113] O. Ozdenizci, S. Y. Gunay, F. Quivira, and D. Erdogmug, “Hierarchical graphical models for context-aware hybrid brain-machine interfaces,” *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 2018, pp. 1964–1967, Jul. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6525618/

[114] K. Kiguchi and Y. Hayashi, “A study of emg and eeg during perception-assist with an upper-limb power-assist robot,” in *2012 IEEE International Conference on Robotics and Automation*, 2012, pp. 2711–2716. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6225027

[115] K. Förster, A. Biasiucci, R. Chavarriaga, J. del R. Millán, D. Roggen, and G. Tröster, “On the use of brain decoded signals for online user adaptive gesture recognition systems,” in *International Conference on Pervasive Computing*, Berlin, Heidelberg, 2010, pp. 427–444.

[116] A. Riccio, E. Holz, P. Aricò, F. Leotta, F. Aloise, L. Desideri, M. Rimondini, A. Kubler, D. Mattia, and F. Cincotti, “Towards a hybrid control of a p300-based bci for communication in severely disabled end-users,” in *TOBI Workshop IV*. Sion, Switzerland: Tools for Brain Computer Interaction Project, Jan. 2013. [Online]. Available: https://www.semanticscholar.org/paper/Towards-a-Hybrid-Control-of-a-P300-based-BCI-for-in-Riccio-Holz/3587a3b803e9315df6fd7a3063efac3116506135

[117] T. W. Picton, “The p300 wave of the human event-related potential,” *Journal of Clinical Neurophysiology*, vol. 9, no. 4, Oct. 1992. [Online]. Available: https://journals.lww.com/clinicalneurophys/fulltext/1992/10000/the_p300_wave_of_the_human_event_related_potential.2.aspx

[118] L. Farwell and E. Donchin, “Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials,” *Electroencephalography and Clinical Neurophysiology*, vol. 70, no. 6, pp. 510–523, 1988. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0013469488901496

[119] J. Polich, “Updating p300: An integrative theory of p3a and p3b,” *Clinical Neurophysiology*, vol. 118, no. 10, pp. 2128–2148, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1388245707001897

[120] S. G. Hart and L. E. Staveland, “Development of nasa-tlx (task load index): Results of empirical and theoretical research,” in *Human Mental Workload*, ser. Advances in Psychology, P. A. Hancock and N. Meshkati, Eds. North-Holland, 1988, vol. 52, pp. 139–183. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0166411508623869

[121] K. KIGUCHI, T. D. LALITHARATNE, and Y. HAYASHI, “Estimation of forearm supination/pronation motion based on eeg signals to control an artificial arm,” *Journal of Advanced Mechanical Design, Systems, and Manufacturing*, vol. 7, no. 1, pp. 74–81, 2013. [Online]. Available: https://www.jstage.jst.go.jp/article/jamdsm/7/1/7_74/_pdf/-char/en

[122] I. Ruhunage, C. J. Perera, K. Nisal, J. Subodha, and T. D. Lalitharatne, “Emg signal controlled transhumerai prosthetic with eeg-ssvep based approch for hand open/close,” in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct. 2017, pp. 3169–3174. [Online]. Available: https://ieeexplore.ieee.org/document/8123115

[123] R. Srinivasan, F. A. Bibi, and P. L. Nunez, “Steady-state visual evoked potentials: Distributed local sources and wave-like dynamics are sensitive to flicker frequency,” *Brain Topography*, vol. 18, no. 3, pp. 167–187, Mar. 2006. [Online]. Available: https://link.springer.com/article/10.1007/s10548-006-0267-4

[124] O. Friman, I. Volosyak, and A. Graser, “Multiple channel detection of steady-state visual evoked potentials for brain-computer interfaces,” *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 4, pp. 742–750, Mar. 2007. [Online]. Available: https://ieeexplore.ieee.org/document/4132932

[125] J. Zhang, B. Wang, C. Zhang, Y. Xiao, and M. Y. Wang, “An eeg/emg/eog-based multimodal human-machine interface to real-time control of a soft robot hand,” *Frontiers in Neurorobotics*, vol. 13, p. 7, 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnbot.2019.00007

[126] P. Sawers, “Thalmic labs rebrands as north, launches $999 alexa-powered holographic glasses,” Oct. 2018. [Online]. Available: https://venturebeat.com/2018/10/23/thalmic-labs-rebrands-as-north-launches-999-alexa-powered-holographic-glasses/

[127] M. Pawłowski, A. Wróblewska, and S. Sysko-Romańczuk, “Effective techniques for multimodal data fusion: A comparative analysis,” *Sensors (Basel)*, vol. 23, no. 5, p. 2381, Feb. 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10007548/

[128] S. Y. Gordleeva, S. A. Lobov, N. A. Grigorev, A. O. Savosenkov, M. O. Shamshin, M. V. Lukoyanov, M. A. Khoruzhko, and V. B. Kazantsev, “Real-time eeg–emg human–machine interface-based control system for a lower-limb exoskeleton,” *IEEE Access*, vol. 8, pp. 84 070–84 081, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9084126

[129] H. Aly and S. M. Youssef, “Bio-signal based motion control system using deep learning models: a deep learning approach for motion classification using eeg and emg signal fusion,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 2, pp. 991–1002, Jul. 2021. [Online]. Available: https://link.springer.com/article/10.1007/s12652-021-03351-1

[130] X. Li, O. W. Samuel, X. Zhang, H. Wang, P. Fang, and G. Li, “A motion-classification strategy based on semg-eeg signal combination for upper-limb amputees,” *Journal of NeuroEngineering and Rehabilitation*, vol. 14, no. 1, p. 2, Jan. 2017. [Online]. Available: https://jneuroengrehab.biomedcentral.com/articles/10.1186/s12984-016-0212-z

[131] B. A. Conway, D. M. Halliday, S. F. Farmer, U. Shahani, P. Maas, A. I. Weir, and J. R. Rosenberg, “Synchronization between motor cortex and spinal motoneuronal pool during the performance of a maintained motor task in man,” *J. Physiol.*, vol. 489 (Pt 3), pp. 917–924, Dec. 1995. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/8788955/

[132] D. M. Halliday, B. A. Conway, S. F. Farmer, and J. R. Rosenberg, “Using electroencephalography to study functional coupling between cortical activity and electromyograms during voluntary contractions in humans,” *Neuroscience Letters*, vol. 241, no. 1, pp. 5–8, Jan. 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304394097009646

[133] G. Severini, S. Conforto, M. Schmid, and T. D’Alessio, “A multivariate auto-regressive method to estimate cortico-muscular coherence for the detection of movement intent,” *Applied Bionics and Biomechanics*, vol. 9, no. 353272, pp. 135–143, 2012. [Online]. Available: https://www.hindawi.com/journals/abb/2012/353272/

[134] A. Chowdhury, H. Raza, Y. K. Meena, A. Dutta, and G. Prasad, “An eeg-emg correlation-based brain-computer interface for hand orthosis supported neuro-rehabilitation,” *Journal of Neuroscience Methods*, vol. 312, pp. 1–11, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165027018303790

[135] X. Lou, S. Xiao, Y. Qi, X. Hu, Y. Wang, and X. Zheng, “Corticomuscular coherence analysis on hand movement distinction for active rehabilitation,” *Computational and Mathematical Methods in Medicine*, vol. 2013, no. 908591, Apr. 2013. [Online]. Available: https://www.hindawi.com/journals/cmmm/2013/908591/

[136] A. Chowdhury, A. Dutta, and G. Prasad, “Corticomuscular co-activation based hybrid brain-computer interface for motor recovery monitoring,” *IEEE Access*, vol. 8, pp. 174 542–174 557, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9204608/

[137] H. I. Aly, S. Youssef, and C. Fathy, “Hybrid brain computer interface for movement control of upper limb prostheses,” in *2018 International Conference on Biomedical Engineering and Applications (ICBEA)*, 2018, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/8471729

[138] Y. Zhang, B. Liu, X. Ji, and D. Huang, “Classification of eeg signals based on autoregressive model and wavelet packet decomposition,” *Neural Processing Letters*, vol. 45, no. 2, pp. 365–378, Apr. 2017. [Online]. Available: https://link.springer.com/article/10.1007/s11063-016-9530-1

[139] M. S. Al-Quraishi, I. Elamvazuthi, T. B. Tang, M. Al-Qurishi, S. Parasuraman, and A. Borboni, “Multimodal fusion approach based on eeg and emg signals for lower limb movement recognition,” *IEEE Sensors Journal*, vol. 21, no. 24, pp. 27 640–27 650, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9565878

[140] M. Haghighat, M. Abdel-Mottaleb, and W. Alhalabi, “Discriminant correlation analysis for feature level fusion with application to multimodal biometrics,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Mar. 2016, pp. 1866–1870. [Online]. Available: https://ieeexplore.ieee.org/document/7472000

[141] J. Tryon, E. Friedman, and A. L. Trejos, “Performance evaluation of eeg/emg fusion methods for motion classification,” in *16th IEEE International Conference on Rehabilitation Robotics (ICORR)*. Toronto, Canada: IEEE, Jun. 2019, pp. 971–976. [Online]. Available: https://ieeexplore.ieee.org/document/8779465

[142] J. Tryon and A. L. Trejos, “Classification of task weight during dynamic motion using eeg–emg fusion,” *IEEE Sensors Journal*, vol. 21, no. 4, pp. 5012–5021, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9237976

[143] ——, “Evaluating convolutional neural networks as a method of EEG-EMG fusion,” *Front. Neurorobot.*, vol. 15, p. 692183, Nov. 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8649783/

[144] Q. Tao and R. Veldhuis, “Optimal decision fusion for a face verification system,” in *Advances in Biometrics*, S.-W. Lee and S. Z. Li, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 958–967. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-540-74549-5_100

[145] C. Cui, G. Bian, Z. Hou, J. Zhao, and H. Zhou, “A multimodal framework based on integration of cortical and muscular activities for decoding human intentions about lower limb motions,” *IEEE Transactions on Biomedical Circuits and Systems*, vol. 11, no. 4, pp. 889–899, 2017. [Online]. Available: https://ieeexplore.ieee.org/document/7983375

[146] R. Leeb, H. Sagha, R. Chavarriaga, and J. del R. Millán, “Multimodal fusion of muscle and brain signals for a hybrid-bci,” in *32nd Annual International Conference of the IEEE EMBS*. Buenos Aires, Argentina: IEEE, Aug. 2010, pp. 4343–4346. [Online]. Available: https://ieeexplore.ieee.org/document/5626233

[147] M. Pritchard, A. I. Weinberg, J. A. R. Williams, F. Campelo, H. Goldingay, and D. R. Faria, “Dynamic fusion of electromyographic and electroencephalographic data towards use in robotic prosthesis control,” *Journal of Physics: Conference Series*, vol. 1828, no. 1, p. 012056, Feb. 2021. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/1828/1/012056

[148] Z. Wang and R. Suppiah, "Upper limb movement recognition utilising eeg and emg signals for rehabilitative robotics," in *Advances in Information and Communication*, K. Arai, Ed. Cham: Springer Nature Switzerland, 2023, pp. 676–695. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-28076-4_49

[149] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in Large Margin Classifiers*, vol. 10, pp. 61–64, Mar. 1999. [Online]. Available: https://home.cs.colorado.edu/~mozer/Teaching/syllabi/6622/papers/Platt1999.pdf

[150] M. Kurzynski, "On a two-level multiclassifier system with error correction applied to the control of bioprosthetic hand," *Studies in Health Technology and Informatics*, vol. 192, p. 1093, 2013. [Online]. Available: https://ebooks.iospress.nl/publication/34309

[151] M. Kim, J. Lee, and K. Kim, "Enhancement of semg-based gesture classification using mahanobis distance metric," in *2016 6th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, Jun. 2016, pp. 1117–1122. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7523781

[152] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update," *Journal of Neural Engineering*, vol. 15, no. 3, Apr. 2018. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2552/aab2f2

[153] A. S. Royer, A. J. Doud, M. L. Rose, and B. He, "Eeg control of a virtual helicopter in 3-dimensional space using intelligent control strategies," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 18, no. 6, pp. 581–589, Sep. 2010. [Online]. Available: https://ieeexplore.ieee.org/document/5585778

[154] K. LaFleur, K. Cassady, A. Doud, K. Shades, E. Rogin, and B. He, "Quadcopter control in three-dimensional space using a noninvasive motor imagery-based brain–computer interface," *Journal of Neural Engineering*, vol. 10, no. 4, p. 046003, Jun. 2013. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2560/10/4/046003

[155] G. Pfurtscheller, C. Guger, G. Müller, G. Krausz, and C. Neuper, "Brain oscillations control hand orthosis in a tetraplegic," *Neuroscience Letters*, vol. 292, no. 3, pp. 211–214, Oct. 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304394000014713

[156] G. Pfurtscheller, G. R. Müller, J. Pfurtscheller, H. J. Gerner, and R. Rupp, "'thought' – control of functional electrical stimulation to restore hand grasp in a patient with tetraplegia," *Neuroscience Letters*, vol. 351, no. 1, pp. 33–36, Nov. 2003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304394003009479

[157] J. R. Wolpaw and D. J. McFarland, "Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 51, pp. 17 849–17 854, Dec. 2004. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC535103/

[158] M. Jochumsen, I. K. Niazi, K. Dremstrup, and E. N. Kamavuako, "Detecting and classifying three different hand movement types through electroencephalography recordings for neurorehabilitation," *Medical & Biological Engineering & Computing*, vol. 54, no. 10, pp. 1491–1501, Oct. 2016. [Online]. Available: https://link.springer.com/article/10.1007/s11517-015-1421-5

[159] R. Xiao and L. Ding, "Evaluation of eeg features in decoding individual finger movements from one hand," *Computational and Mathematical Methods in Medicine*, vol. 2013, p. 243257, Apr. 2013. [Online]. Available: https://doi.org/10.1155/2013/243257

[160] K. Liao, R. Xiao, J. Gonzalez, and L. Ding, "Decoding individual finger movements from one hand using human eeg signals," *PLOS ONE*, vol. 9, no. 1, pp. 1–12, 01 2014. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0085192

[161] R. Alazrai, H. Alwanni, and M. I. Daoud, "Eeg-based bci system for decoding finger movements within the same hand," *Neuroscience Letters*, vol. 698, pp. 113–120, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304394018309029

[162] I. n. Iturrate, R. Leeb, R. Chavarriaga, and J. d. R. Millán, "Decoding of two hand grasping types from eeg," in *6th International Brain-Computer Interface Meeting*. Verlag der Technischen Universität Graz, May 2016. [Online]. Available: https://www.semanticscholar.org/paper/Decoding-of-two-hand-grasping-types-from-EEG-Iturrate-Leeb/dfe35c308ccb758ac1fbbcbbfbc1c992b0faf88d

[163] A. Schwarz, P. Ofner, J. Pereira, A. I. Sburlea, and G. R. Müller-Putz, "Decoding natural reach-and-grasp actions from human EEG," *Journal of Neural Engineering*, vol. 15, no. 1, p. 016005, Dec. 2017. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2552/aa8911

[164] S. S. Mohseni Salehi, M. Moghadamfalahi, F. Quivira, A. Piers, H. Nezamfar, and D. Erdogmus, "Decoding complex imagery hand gestures," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2017, Jul. 2017, pp. 2968–2971. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6525619/

[165] K. C. Ames and M. M. Churchland, "Motor cortex signals for each arm are mixed across hemispheres and neurons yet partitioned within the population response," *eLife*, vol. 8, p. e46159, Oct. 2019. [Online]. Available: https://doi.org/10.7554/eLife.46159

[166] K. J. Miller, S. Zanos, E. E. Fetz, M. den Nijs, and J. G. Ojemann, "Decoupling the cortical power spectrum reveals real-time representation of individual finger movements in humans," *Journal of Neuroscience*, vol. 29, no. 10, pp. 3132–3137, 2009. [Online]. Available: https://www.jneurosci.org/content/29/10/3132

[167] T. Pistohl, A. Schulze-Bonhage, A. Aertsen, C. Mehring, and T. Ball, "Decoding natural grasp types from human ecog," *NeuroImage*, vol. 59, no. 1, pp. 248 – 260, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S105381191100749X

[168] T. Pistohl, T. S. B. Schmidt, T. Ball, A. Schulze-Bonhage, A. Aertsen, and C. Mehring, "Grasp detection from human ecog during natural reach-to-grasp movements," *PLOS ONE*, vol. 8, pp. 1–11, 01 2013. [Online]. Available: https://doi.org/10.1371/journal.pone.0054658

[169] R. Leeb, H. Sagha, R. Chavarriaga, and J. del R Millán, "A hybrid brain–computer interface based on the fusion of electroencephalographic and electromyographic activities," *Journal of Neural Engineering*, vol. 8, no. 2, p. 025011, Mar. 2011. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2560/8/2/025011

[170] C. R. Pernet, P. Sajda, and G. A. Rousselet, "Single-trial analyses: why bother?" *Front. Psychol.*, vol. 2, p. 322, Nov. 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3210509/

[171] I. Ketykó, F. Kovács, and K. Z. Varga, "Domain adaptation for semg-based gesture recognition with recurrent neural networks," in *2019 International Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1–7. [Online]. Available: https://ieeexplore.ieee.org/document/8852018

[172] Z. Lu, X. Chen, Q. Li, X. Zhang, and P. Zhou, "A hand gesture recognition framework and wearable gesture-based interaction prototype for mobile devices," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 2, pp. 293–299, 2014. [Online]. Available: https://ieeexplore.ieee.org/document/6748952

[173] M. E. Benalcázar, C. Motoche, J. A. Zea, A. G. Jaramillo, C. E. Anchundia, P. Zambrano, M. Segura, F. Benalcázar Palacios, and M. Pérez, "Real-time hand gesture recognition using the myo armband and muscle activity detection," in *2017 IEEE Second Ecuador Technical Chapters Meeting (ETCM)*, 2017, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/8247458

[174] C. Castellini, A. E. Fiorilla, and G. Sandini, "Multi-subject/daily-life activity emg-based control of mechanical hands," *Journal of NeuroEngineering and Rehabilitation*, vol. 6, no. 1, p. 41, Nov. 2009. [Online]. Available: https://jneuroengrehab.biomedcentral.com/articles/10.1186/1743-0003-6-41

[175] Y. Du, W. Jin, W. Wei, Y. Hu, and W. Geng, "Surface EMG-based inter-session gesture recognition enhanced by deep domain adaptation," *Sensors (Basel)*, vol. 17, no. 3, Feb. 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5375744/

[176] H. Kang, Y. Nam, and S. Choi, "Composite common spatial pattern for subject-to-subject transfer," *IEEE Signal Processing Letters*, vol. 16, no. 8, pp. 683–686, 2009. [Online]. Available: https://ieeexplore.ieee.org/document/4912345

[177] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularized common spatial patterns with generic learning for eeg signal classification," in *2009 Annual International Conference of the*

IEEE Engineering in Medicine and Biology Society, Sep. 2009, pp. 6599–6602. [Online]. Available: https://ieeexplore.ieee.org/document/5332554

[178] P. Gaur, K. McCreadie, R. B. Pachori, H. Wang, and G. Prasad, "Tangent space features-based transfer learning classification model for two-class motor imagery brain-computer interface," *Int. J. Neural Syst.*, vol. 29, no. 10, p. 1950025, Dec. 2019, https://pure.ulster.ac.uk/ws/portalfiles/portal/77556642/ws_ijns_revision7.pdf. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31711330/

[179] P. Gaur, A. Chowdhury, K. McCreadie, R. B. Pachori, and H. Wang, "Logistic regression with tangent space-based cross-subject learning for enhancing motor imagery classification," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 3, pp. 1188–1197, Jul. 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9495927

[180] H. Kang and S. Choi, "Bayesian common spatial patterns for multi-subject eeg classification," *Neural Networks*, vol. 57, pp. 39–50, Sep. 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608014001117

[181] ——, "Bayesian common spatial patterns with dirichlet process priors for multi-subject eeg classification," in *The 2012 International Joint Conference on Neural Networks (IJCNN)*. IEEE, Jun. 2012, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6252554

[182] F. Lotte and C. Guan, "Learning from other subjects helps reducing brain-computer interface calibration time," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, Mar. 2010, pp. 614–617. [Online]. Available: https://ieeexplore.ieee.org/document/5495183

[183] M. A. M. Joadder, J. J. Myszewski, M. H. Rahman, and I. Wang, "A performance based feature selection technique for subject independent mi based bci," *Health Information Science and Systems*, vol. 7, no. 1, p. 15, Aug. 2019. [Online]. Available: https://link.springer.com/article/10.1007/s13755-019-0076-2

[184] A. M. Azab, L. Mihaylova, K. K. Ang, and M. Arvaneh, "Weighted transfer learning for improving motor imagery-based brain–computer interface," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 7, pp. 1352–1359, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8737742

[185] O. A. Gonzales-Huisa, G. Oshiro, V. E. Abarca, J. G. Chavez-Echajaya, and D. A. Elias, "Emg and imu data fusion for locomotion mode classification in transtibial amputees," *Prosthesis*, vol. 5, no. 4, pp. 1232–1256, 2023. [Online]. Available: https://www.mdpi.com/2673-1592/5/4/85

[186] E. Gibney, "Could machine learning fuel a reproducibility crisis in science?" *Nature*, vol. 608, pp. 250–251, Jul. 2022. [Online]. Available: https://www.nature.com/articles/d41586-022-02035-w

[187] E. Sohn, "The reproducibility issues that haunt health-care ai," *Nature*, vol. 613, pp. 402–403, Jan. 2023. [Online]. Available: https://www.nature.com/articles/d41586-023-00023-2

[188] E. Dubnansky and M. B. Omary, "Acknowledging joint first authors of published work: The time has come," *Gastroenterology*, vol. 143, no. 4, pp. 879–880, Oct. 2012. [Online]. Available: https://www.gastrojournal.org/article/S0016-5085(12)01193-6/fulltext

[189] K. S. Kim, H. H. Choi, C. S. Moon, and C. W. Mun, "Comparison of k-nearest neighbor, quadratic discriminant and linear discriminant analysis in classification of electromyogram signals based on the wrist-motion directions," *Current Applied Physics*, vol. 11, no. 3, pp. 740–745, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1567173910004153

[190] R. Whelan and H. Garavan, "When optimism hurts: Inflated predictions in psychiatric neuroimaging," *Biological Psychiatry*, vol. 75, no. 9, pp. 746–748, 2014, mechanisms of Aging and Cognition. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0006322313004575

[191] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, Aug. 2023. [Online]. Available: https://www.cell.com/patterns/abstract/S2666-3899(23)00159-9

[192] R. Li, J. S. Johansen, H. Ahmed, T. V. Ilyevsky, R. B. Wilbur, H. M. Bharadwaj, and J. M. Siskind, "The perils and pitfalls of block design for eeg classification experiments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 316–333, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9264220

[193] M. Billinger, I. Daly, V. Kaiser, J. Jin, B. Z. Allison, G. R. Müller-Putz, and C. Brunner, *Is It Significant? Guidelines for Reporting BCI Performance.* Berlin, Heidelberg: Springer Berlin Heidelberg, Jul. 2012, pp. 333–354. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-29746-5_17

[194] M. Tangermann, K.-R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. Mueller-Putz, G. Nolte, G. Pfurtscheller, H. Preissl, G. Schalk, A. Schlögl, C. Vidaurre, S. Waldert, and B. Blankertz, "Review of the bci competition iv," *Frontiers in Neuroscience*, vol. 6, Jul. 2012. [Online]. Available: https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2012.00055

[195] M. Atzori, A. Gijsberts, C. Castellini, B. Caputo, A.-G. M. Hager, S. Elsig, G. Giatsidis, F. Bassetto, and H. Müller, "Electromyography data for non-invasive naturally-controlled robotic hand prostheses," *Scientific Data*, vol. 1, no. 1, p. 140053, Dec. 2014. [Online]. Available: https://doi.org/10.1038/sdata.2014.53

[196] D. Gwon, K. Won, M. Song, C. S. Nam, S. C. Jun, and M. Ahn, "Review of public motor imagery and execution datasets in brain-computer interfaces," *Front. Hum. Neurosci.*, vol. 17, p. 1134869, Mar. 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10101208

[197] M. D. Luciw, E. Jarocka, and B. B. Edin, “Multi-channel eeg recordings during 3,936 grasp and lift trials with varying weight and friction,” *Scientific Data*, vol. 1, no. 1, p. 140047, Nov. 2014. [Online]. Available: https://www.nature.com/articles/sdata201447

[198] J.-H. Jeong, J.-H. Cho, K.-H. Shim, B.-H. Kwon, B.-H. Lee, D.-Y. Lee, D.-H. Lee, and S.-W. Lee, “Multimodal signal dataset for 11 intuitive movement tasks from single upper extremity during multiple recording sessions,” *Gigascience*, vol. 9, no. 10, Oct. 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7539536/

[199] ——, “Supporting data for "multimodal signal dataset for 11 intuitive movement tasks from single upper extremity during multiple recording sessions",” 2020. [Online]. Available: http://dx.doi.org/10.5524/100788

[200] T. Feix, J. Romero, H.-B. Schmiedmayer, A. M. Dollar, and D. Kragic, “The grasp taxonomy of human grasp types,” *IEEE Transactions on human-machine systems*, vol. 46, no. 1, pp. 66–77, 2015, https://is.mpg.de/uploads_file/attachment/attachment/256/grasp_taxonomy.pdf. [Online]. Available: https://ieeexplore.ieee.org/document/7243327

[201] J. T. Panachakel, N. N. Vinayak, M. Nunna, A. G. Ramakrishnan, and K. Sharma, “An improved eeg acquisition protocol facilitates localized neural activation,” Mar. 2020. [Online]. Available: https://arxiv.org/abs/2003.10212

[202] R. Joshi, P. Saraswat, and R. Gajendran, “A novel mu rhythm-based brain computer interface design that uses a programmable system on chip,” *Journal of medical signals and sensors*, vol. 2, no. 1, pp. 11–16, Jan. 2012. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/23493871

[203] B. Amanpour and A. Erfanian, “Classification of brain signals associated with imagination of hand grasping, opening and reaching by means of wavelet-based common spatial pattern and mutual information,” in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2013, pp. 2224–2227. [Online]. Available: https://ieeexplore.ieee.org/document/6609978

[204] L. R. Krol. (2020, Nov.) Eeg 10-10 system with additional information. CC0 1.0. [Online]. Available: https://commons.wikimedia.org/wiki/File:EEG_10-10_system_with_additional_information.svg

[205] *MATLAB version 9.8 (R2020a)*, The Mathworks, Inc., Natick, Massachusetts, USA, 2020.

[206] D. Robertson, J. Barden, and J. Dowling, “Response characteristics of different butterworth low-pass digital filters,” *Journal of Biomechanics*, vol. 26, no. 3, p. 299, Mar. 1993. [Online]. Available: https://www.sciencedirect.com/science/article/pii/002192909390410G

[207] T.-P. JUNG, S. MAKEIG, C. HUMPHRIES, T.-W. LEE, M. J. McKEOWN, V. IRAGUI, and T. J. SEJNOWSKI, “Removing electroencephalographic artifacts by blind source separation,” *Psychophysiology*,

vol. 37, no. 2, p. 163–178, Mar. 2000. [Online]. Available: https://www.cambridge.org/core/journals/psychophysiology/article/abs/removing-electroencephalographic-artifacts-by-blind-source-separation/2548D35629CAE17E6956C2FFF1B6C8AB

[208] M. Val-Calvo, J. R. Álvarez Sánchez, J. M. Ferrández-Vicente, and E. Fernández, “Optimization of real-time eeg artifact removal and emotion estimation for human-robot interaction applications,” *Frontiers in Computational Neuroscience*, vol. 13, Nov. 2019. [Online]. Available: https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2019.00080

[209] N. Mammone, F. La Foresta, and F. C. Morabito, “Automatic artifact rejection from multichannel scalp eeg by wavelet ica,” *IEEE Sensors Journal*, vol. 12, no. 3, pp. 533–542, Mar. 2012. [Online]. Available: https://ieeexplore.ieee.org/document/5713804

[210] K.-J. Huang, J.-C. Liao, W.-Y. Shih, C.-W. Feng, J.-C. Chang, C.-C. Chou, and W.-C. Fang, “A real-time processing flow for ica based eeg acquisition system with eye blink artifact elimination,” in *SiPS 2013 Proceedings*, 2013, pp. 237–240. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6674511

[211] D. Steyrl, R. Scherer, J. Faller, and G. R. Müller-Putz, “Random forests in non-invasive sensorimotor rhythm brain-computer interfaces: a practical and convenient non-linear classifier,” *Biomedical Engineering / Biomedizinische Technik*, vol. 61, no. 1, pp. 77–86, 2016. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/bmt-2014-0117/html

[212] L. van der Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, Nov. 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

[213] H. Raza, D. Rathee, S.-M. Zhou, H. Cecotti, and G. Prasad, “Covariate shift estimation based adaptive ensemble learning for handling non-stationarity in motor imagery related eeg-based brain-computer interface,” *Neurocomputing*, vol. 343, pp. 154–166, May 2019, learning in the Presence of Class Imbalance and Concept Drift. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231219301560

[214] B. Marcheix, B. Gardiner, and S. Coleman, “Adaptive gesture recognition system for robotic control using surface emg sensors,” in *2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Dec. 2019, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/9001765

[215] B. Hudgins, P. Parker, and R. Scott, “A new strategy for multifunction myoelectric control,” *IEEE Transactions on Biomedical Engineering*, vol. 40, no. 1, pp. 82–94, 1993.

[216] W. Li, P. Shi, and H. Yu, “Gesture recognition using surface electromyography and deep learning for prostheses hand: State-of-the-art, challenges, and future,” *Frontiers in Neuroscience*, vol. 15, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.621885

[217] K. Englehart and B. Hudgins, “A robust, real-time control scheme for multifunction myoelectric control,” *IEEE transactions on bio-medical engineering*, vol. 50, pp. 848–54, 08 2003.

[218] J. Bird, L. Manso, E. Ribeiro, A. Ekart, and D. Faria, “A study on mental state classification using eeg-based brain-machine interface,” in *2018 International Conference on Intelligent Systems (IS)*, Funchal, Madeira Island, Portugal, Sep. 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8710576

[219] S. Rimbert, C. Lindig-León, M. Fedotenkova, and L. Bougrain, “Modulation of beta power in eeg during discrete and continuous motor imageries,” in *2017 8th International IEEE/EMBS Conference on Neural Engineering (NER)*, May 2017, pp. 333–336. [Online]. Available: https://ieeexplore.ieee.org/document/8008358

[220] Y. J. Yang, E. J. Jeon, J. S. Kim, and C. K. Chung, “Characterization of kinesthetic motor imagery compared with visual motor imageries,” *Scientific Reports*, vol. 11, no. 1, p. 3751, Feb. 2021. [Online]. Available: https://www.nature.com/articles/s41598-021-82241-0

[221] H. Candra, M. Yuwono, R. Chai, A. Handojoseno, I. Elamvazuthi, H. T. Nguyen, and S. Su, “Investigation of window size in classification of eeg-emotion signal with wavelet entropy and support vector machine,” in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug. 2015, pp. 7250–7253. [Online]. Available: https://ieeexplore.ieee.org/document/7320065

[222] T. R. Farrell and R. F. Weir, “The optimal controller delay for myoelectric prostheses,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 15, no. 1, pp. 111–118, 2007.

[223] C. Dolopikos, M. Pritchard, J. J. Bird, and D. R. Faria, “Electromyography signal-based gesture recognition for human-machine interaction in real-time through model calibration,” in *Advances in Information and Communication. Future of Information and Communication Conference (FICC) 2021*, ser. Advances in Intelligent Systems and Computing, K. Arai, Ed., vol. 1364. Cham: Springer International Publishing, Apr. 2021, pp. 898–914. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-73103-8_65

[224] M. Shahzaib and S. Shakil, “Hand electromyography circuit and signals classification using artificial neural network,” in *2018 14th International Conference on Emerging Technologies (ICET)*, 2018, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/8603587

[225] R. N. Khushaba, A. Al-Ani, A. Al-Timemy, and A. Al-Jumaily, “A fusion of time-domain descriptors for improved myoelectric hand control,” in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2016, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/7850064

[226] M. Atzori, A. Gijsberts, S. Heynen, A.-G. M. Hager, O. Deriaz, P. van der Smagt, C. Castellini, B. Caputo, and H. Müller, “Building the ninapro database: A resource for the biorobotics community,” in *2012 4th IEEE RAS & EMBS International Conference on*

Biomedical Robotics and Biomechatronics (BioRob), 2012, pp. 1258–1265. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6290287

[227] M. Zardoshti-Kermani, B. Wheeler, K. Badie, and R. Hashemi, "Emg feature evaluation for movement control of upper extremity prostheses," *IEEE Transactions on Rehabilitation Engineering*, vol. 3, no. 4, pp. 324–333, Dec. 1995. [Online]. Available: https://ieeexplore.ieee.org/document/481972

[228] R. Menon, G. Di Caterina, H. Lakany, L. Petropoulakis, B. A. Conway, and J. J. Soraghan, "Study on interaction between temporal and spatial information in classification of emg signals for myoelectric prostheses," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 10, pp. 1832–1842, 2017. [Online]. Available: https://ieeexplore.ieee.org/document/7904664

[229] L. H. Smith, L. J. Hargrove, B. A. Lock, and T. A. Kuiken, "Determining the optimal window length for pattern recognition-based myoelectric control: balancing the competing effects of classification error and controller delay," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 19, no. 2, pp. 186–192, Apr. 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4241762/

[230] J. Kobylarz, J. J. Bird, D. R. Faria, E. P. Ribeiro, and A. Ekárt, "Thumbs up, thumbs down: non-verbal human-robot interaction through real-time emg classification via inductive and supervised transductive transfer learning," *Journal of Ambient Intelligence and Humanized Computing*, Mar. 2020. [Online]. Available: https://doi.org/10.1007/s12652-020-01852-z

[231] A. Phinyomark, F. Quaine, Y. Laurillau, S. Thongpanja, C. Limsakul, and P. Phukpattaranont, "Emg amplitude estimators based on probability distribution for muscle-computer interface," *Fluctuation and Noise Letters*, vol. 12, 09 2013. [Online]. Available: https://www.researchgate.net/publication/262011230_EMG_Amplitude_Estimators_Based_on_Probability_Distribution_for_Muscle-Computer_Interface

[232] S. Lobov, V. Mironov, I. Kastalskiy, and V. Kazantsev, "Combined use of command-proportional control of external robotic devices based on electromyography signals," *Sovremennye tehnologii v medicine*, vol. 7, pp. 30–38, 12 2015. [Online]. Available: http://stm-journal.ru/en/numbers/2015/4/1189

[233] X. Chen, A. Ke, X. Ma, and J. He, "Soc-based architecture for robotic prosthetics control using surface electromyography," in *2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, vol. 01, 2016, pp. 134–137.

[234] M. Atzori and H. Müller, "Pawfe: Fast signal feature extraction using parallel time windows," *Frontiers in Neurorobotics*, vol. 13, p. 74, 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnbot.2019.00074

[235] W. Rose, "Electromyogram analysis," Oct. 2019. [Online]. Available: https://www1.udel.edu/biology/rosewc/kaap686/notes/EMG%20analysis.pdf

[236] B. Wan, R. Wu, K. Zhang, and L. Liu, "A new subtle hand gestures recognition algorithm based on emg and fsr," in *2017 IEEE 21st International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 2017, pp. 127–132.

[237] I. Mohd Khairuddin, S. N. Sidek, A. P P Abdul Majeed, M. A. Mohd Razman, A. Ahmad Puzi, and H. Md Yusof, "The classification of movement intention through machine learning models: the identification of significant time-domain EMG features," *PeerJ Comput. Sci.*, vol. 7, no. e379, p. e379, Feb. 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7959624/

[238] J. Kim, S. Mastnik, and E. André, "Emg-based hand gesture recognition for realtime biosignal interfacing," in *Proceedings of the 13th International Conference on Intelligent User Interfaces*, ser. IUI '08. New York, NY, USA: Association for Computing Machinery, Jan. 2008, p. 30–39. [Online]. Available: https://doi.org/10.1145/1378773.1378778

[239] G. N. Saridis and T. P. Gootee, "Emg pattern analysis and classification for a prosthetic arm," *IEEE Transactions on Biomedical Engineering*, vol. BME-29, no. 6, pp. 403–412, 1982.

[240] X. Chen and Z. J. Wang, "Pattern recognition of number gestures based on a wireless surface emg system," *Biomedical Signal Processing and Control*, vol. 8, no. 2, pp. 184–192, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809412000870

[241] A. Phinyomark, P. Phukpattaranont, and C. Limsakul, "Feature reduction and selection for emg signal classification," *Expert Systems with Applications*, vol. 39, no. 8, pp. 7420–7431, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417412001200

[242] K. Englehart, B. Hudgin, and P. Parker, "A wavelet-based continuous classification scheme for multifunction myoelectric control," *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 3, pp. 302–311, 2001.

[243] Y. Zhang, G. Wang, C. Teng, Z. Sun, and J. Wang, "The analysis of hand movement distinction based on relative frequency band energy method," *BioMed Research International*, vol. 2014, p. 781769, Nov. 2014. [Online]. Available: https://doi.org/10.1155/2014/781769

[244] D. Farina and R. Merletti, "Comparison of algorithms for estimation of emg variables during voluntary isometric contractions," *Journal of Electromyography and Kinesiology*, vol. 10, no. 5, pp. 337–349, 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1050641100000250

[245] D. Farina, R. Merletti, and R. M. Enoka, "The extraction of neural strategies from the surface emg," *Journal of Applied Physiology*, vol. 96, no. 4, pp. 1486–1495, 2004, pMID: 15016793. [Online]. Available: https://journals.physiology.org/doi/full/10.1152/japplphysiol.01070.2003

[246] S. Orcioni, F. D. Nardo, S. Fioretti, M. Conti, R. Seepold, M. Gaiduk, and N. M. Madrid, "Preliminary results of homomorphic deconvolution application to surface emg signals during

walking," *Procedia Comput. Sci.*, vol. 192, no. C, p. 3272–3280, Jan. 2021. [Online]. Available: https://doi.org/10.1016/j.procs.2021.09.100

[247] G. Pfurtscheller and C. Neuper, "Motor imagery activates primary sensorimotor area in humans," *Neuroscience Letters*, vol. 239, no. 2, pp. 65–68, 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304394097008896

[248] J.-H. Park, H.-S. Cynn, K. S. Cha, K. H. Kim, and H.-S. Jeon, "Event-related desynchronization of mu rhythms during concentric and eccentric contractions," *Journal of Motor Behavior*, vol. 50, no. 4, pp. 457–466, 2018. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/00222895.2017.1367638

[249] Siuly, Y. Li, J. Wu, and J. Yang, "Developing a logistic regression model with cross-correlation for motor imagery signal recognition," in *The 2011 IEEE/ICME International Conference on Complex Medical Engineering*, May 2011, pp. 502–507. [Online]. Available: https://ieeexplore.ieee.org/document/5876793

[250] C. Thornton, F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Auto-weka: Combined selection and hyperparameter optimization of classification algorithms," 2013.

[251] N. S. Young, J. P. A. Ioannidis, and O. Al-Ubaydli, "Why current publication practices may distort science," *PLoS Med.*, vol. 5, no. 10, p. e201, Oct. 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2561077/

[252] M. Garouani, A. Ahmad, M. Bouneffa, M. Hamlich, G. Bourguin, and A. Lewandowski, "Using meta-learning for automated algorithms selection and configuration: an experimental framework for industrial big data," *Journal of Big Data*, vol. 9, no. 1, p. 57, Apr. 2022. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00612-4

[253] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds., vol. 24. Curran Associates, Inc., 2011. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf

[254] J. Bergstra, D. Yamins, and D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in *Proceedings of the 30th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, S. Dasgupta and D. McAllester, Eds., vol. 28, no. 1. Atlanta, Georgia, USA: PMLR, Jun. 2013, pp. 115–123. [Online]. Available: https://proceedings.mlr.press/v28/bergstra13.html

[255] A. Jaffe, E. Fetaya, B. Nadler, T. Jiang, and Y. Kluger, "Unsupervised ensemble learning with dependent classifiers," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Gretton and C. C. Robert, Eds., vol. 51. Cadiz, Spain: PMLR, May 2016, pp. 351–360. [Online]. Available: https://proceedings.mlr.press/v51/jaffe16.html

[256] E. A. Kirchner, M. Tabie, and A. Seeland, “Multimodal movement prediction - towards an individual assistance of patients,” *PLOS ONE*, vol. 9, no. 1, pp. 1–9, 01 2014. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0085060

[257] A. Manolova, G. Tsenov, V. Lazarova, and N. Neshov, “Combined eeg and emg fatigue measurement framework with application to hybrid brain-computer interface,” in *2016 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom)*, Jun. 2016, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/7901569

[258] E. Ciaccio, S. Dunn, and Y. Akay, “Biosignal pattern recognition and interpretation systems. 2. methods for feature extraction and selection,” *Engineering in Medicine and Biology Magazine, IEEE*, vol. 12, pp. 106 – 113, 01 1994, https://www.researchgate.net/publication/3244876_Biosignal_pattern_recognition_and_interpretation_systems_2_Methods_for_feature_extraction_and_selection.

[259] I. Kononenko, E. Šimec, and M. Robnik-Šikonja, “Overcoming the myopia of inductive learning algorithms with relieff,” *Applied Intelligence*, vol. 7, no. 1, pp. 39–55, Jan. 1997. [Online]. Available: https://link.springer.com/article/10.1023/A:1008280620621

[260] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

[261] B. Komer, J. Bergstra, and C. Eliasmith, *Hyperopt-Sklearn.* Cham: Springer International Publishing, 2019, pp. 97–111. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-05318-5_5

[262] A. Phinyomark and E. Scheme, “Emg pattern recognition in the era of big data and deep learning,” *Big Data and Cognitive Computing*, vol. 2, no. 3, Aug. 2018. [Online]. Available: https://www.mdpi.com/2504-2289/2/3/21

[263] D. Garrett, D. Peterson, C. Anderson, and M. Thaut, “Comparison of linear, nonlinear, and feature selection methods for eeg signal classification,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, no. 2, pp. 141–144, 2003.

[264] K.-R. Muller, C. Anderson, and G. Birch, “Linear and nonlinear methods for brain-computer interfaces,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, no. 2, pp. 165–169, Jun. 2003.

[265] L. J. Hargrove, K. Englehart, and B. Hudgins, “A comparison of surface and intramuscular myoelectric signal classification,” *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 5, pp. 847–853, 2007.

[266] M. Gandolla, S. Ferrante, G. Ferrigno, D. Baldassini, F. Molteni, E. Guanziroli, M. C. Cottini, C. Seneci, and A. Pedrocchi, “Artificial neural network emg classifier for functional hand grasp

movements prediction," *Journal of International Medical Research*, vol. 45, no. 6, pp. 1831–1847, Sep. 2017. [Online]. Available: https://journals.sagepub.com/doi/10.1177/0300060516656689

[267] S. Qi, X. Wu, W.-H. Chen, J. Liu, J. Zhang, and J. Wang, "semg-based recognition of composite motion with convolutional neural network," *Sensors and Actuators A: Physical*, vol. 311, p. 112046, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924424719313305

[268] J. Ashford, J. Bird, F. Campelo, and D. Faria, "Classification of eeg signals based on image representation of statistical features," in *UK Workshop on Computational Intelligence.* Springer, 2019, pp. 449–460.

[269] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[270] G. Biau and E. Scornet, "A random forest guided tour," *TEST*, vol. 25, no. 2, pp. 197–227, Jun. 2016. [Online]. Available: https://link.springer.com/article/10.1007/s11749-016-0481-7

[271] R. Su, X. Chen, S. Cao, and X. Zhang, "Random forest-based recognition of isolated sign language subwords using data from accelerometers and surface electromyographic sensors," *Sensors*, vol. 16, no. 1, 2016. [Online]. Available: https://www.mdpi.com/1424-8220/16/1/100

[272] C. Tang, D. Garreau, and U. von Luxburg, "When do random forests fail?" in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/204da255aea2cd4a75ace6018fad6b4d-Paper.pdf

[273] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?" in *Machine Learning and Data Mining in Pattern Recognition*, P. Perner, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 154–168.

[274] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[275] E. Frank, M. A. Hall, and I. H. Witten, *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, 4th ed., 2016. [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf

[276] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[277] H. Nisar, K. Wee Boon, Y. Kim Ho, and T. Shen Khang, "Brain-computer interface: Feature extraction and classification of motor imagery-based cognitive tasks," in *2022 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, 2022, pp. 42–47. [Online]. Available: https://ieeexplore.ieee.org/document/9815460

[278] A. Schlögl, F. Lee, H. Bischof, and G. Pfurtscheller, "Characterization of four-class motor imagery eeg data for the bci-competition 2005," *Journal of Neural Engineering*, vol. 2, no. 4, p. L14, Aug. 2005. [Online]. Available: https://dx.doi.org/10.1088/1741-2560/2/4/L02

[279] J. Große Sundrup and K. Mombaur, "On the distribution of muscle signals: A method for distance-based classification of human gestures," *Sensors*, vol. 23, no. 17, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/17/7441

[280] A. Singh, A. Yadav, and A. Rana, "Article: K-means with three different distance metrics," *International Journal of Computer Applications*, vol. 67, no. 10, pp. 13–17, Apr. 2013. [Online]. Available: https://www.ijcaonline.org/archives/volume67/number10/11430-6785

[281] A. Tharwat, "Linear vs. quadratic discriminant analysis classifier: a tutorial," *International Journal of Applied Pattern Recognition*, vol. 3, no. 2, pp. 145–180, 2016, pMID: 79050. [Online]. Available: https://www.inderscienceonline.com/doi/abs/10.1504/IJAPR.2016.079050

[282] K. D. Katyal, M. S. Johannes, T. G. McGee, A. J. Harris, R. S. Armiger, A. H. Firpi, D. McMullen, G. Hotson, M. S. Fifer, N. E. Crone, R. J. Vogelstein, and B. A. Wester, "Harmonie: A multimodal control framework for human assistive robotics," in *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, 2013, pp. 1274–1278. [Online]. Available: https://ieeexplore.ieee.org/document/6696173

[283] M. Linderman, M. A. Lebedev, and J. S. Erlichman, "Recognition of handwriting from electromyography," *PLoS One*, vol. 4, no. 8, p. e6791, Aug. 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2727961/

[284] J. Schäfer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, p. Article32, Nov. 2005. [Online]. Available: https://www.degruyter.com/document/doi/10.2202/1544-6115.1175/html

[285] A. Mkhadri, "Shrinkage parameter for the modified linear discriminant analysis," *Pattern Recognition Letters*, vol. 16, no. 3, pp. 267–275, 1995. [Online]. Available: https://www.sciencedirect.com/science/article/pii/016786559400100H

[286] H. Zhang, "The optimality of naive bayes," in *The Florida AI Research Society*, 2004. [Online]. Available: https://api.semanticscholar.org/CorpusID:8891634

[287] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, ser. COLT '92. New York, NY, USA: Association for Computing Machinery, Jul. 1992, p. 144–152. [Online]. Available: https://dl.acm.org/doi/10.1145/130385.130401

[288] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, May 2011. [Online]. Available: https://doi.org/10.1145/1961189.1961199

[289] J.-P. Vert, K. Tsuda, and B. Schölkopf, "A Primer on Kernel Methods," in *Kernel Methods in Computational Biology.* The MIT Press, 07 2004. [Online]. Available: https://doi.org/10.7551/mitpress/4057.003.0004

[290] Y. Fujiwara, R. Matsumoto, T. Nakae, K. Usami, M. Matsuhashi, T. Kikuchi, K. Yoshida, T. Kunieda, S. Miyamoto, T. Mima, A. Ikeda, and R. Osu, "Neural pattern similarity between contra- and ipsilateral movements in high-frequency band of human electrocorticograms," *NeuroImage*, vol. 147, pp. 302–313, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811916306760

[291] M. Tavakolan, Z. Frehlick, X. Yong, and C. Menon, "Classifying three imaginary states of the same upper extremity using time-domain features," *PLOS ONE*, vol. 12, no. 3, pp. 1–18, 03 2017. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0174161

[292] A. Ameri, E. N. Kamavuako, E. J. Scheme, K. B. Englehart, and P. A. Parker, "Support vector regression for improved real-time, simultaneous myoelectric control," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 6, pp. 1198–1209, 2014. [Online]. Available: https://ieeexplore.ieee.org/document/6817581

[293] L. R. Quitadamo, F. Cavrini, L. Sbernini, F. Riillo, L. Bianchi, S. Seri, and G. Saggio, "Support vector machines to detect physiological patterns for EEG and EMG-based human–computer interaction: a review," *Journal of Neural Engineering*, vol. 14, no. 1, p. 011001, Jan. 2017. [Online]. Available: https://doi.org/10.1088%2F1741-2552%2F14%2F1%2F011001

[294] X. Yong and C. Menon, "Eeg classification of different imaginary movements within the same limb," *PLOS ONE*, vol. 10, no. 4, pp. 1–24, 04 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0121896

[295] A. Seeland, M. Tabie, S. K. Kim, F. Kirchner, and E. A. Kirchner, "Adaptive multimodal biosignal control for exoskeleton supported stroke rehabilitation," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct. 2017, pp. 2431–2436. [Online]. Available: https://ieeexplore.ieee.org/document/8122987

[296] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, no. 61, pp. 1871–1874, 2008. [Online]. Available: http://jmlr.org/papers/v9/fan08a.html

[297] The MathWorks, Inc., *Support Vector Machine Classification — MATLAB & Simulink*, 2023, accessed: October 2023. [Online]. Available: https://uk.mathworks.com/help/stats/support-vector-machine-classification.html

[298] N. M. Razali and Y. B. Wah, "Power comparisons of shapiro-wilk , kolmogorov-smirnov , lilliefors and anderson-darling tests," *Journal of Statistical Modeling and Analytics*, vol. 2, pp. 21–33, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:18639594

[299] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/

[300] G. Müller-Putz, R. Scherer, C. Brunner, R. Leeb, and G. Pfurtscheller, "Better than random? a closer look on bci results," *International Journal of Bioelectromagnetism*, vol. 10, pp. 52–55, 01 2008.

[301] J. Ashford, J. Reis-Cunha, I. Lobo, F. Lobo, and F. Campelo, "Organism-specific training improves performance of linear B-cell epitope prediction," *Bioinformatics*, vol. 37, no. 24, pp. 4826–4834, 07 2021, Supplementary File 6: Feature Relevance. [Online]. Available: https://academic.oup.com/bioinformatics/article/37/24/4826/6325084

[302] V. M.K and K. K, "A survey on similarity measures in text mining," in *Machine Learning and Applications: An International Journal*, Mar. 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:62118842

[303] T. M. Hall, F. de Carvalho, and A. Jackson, "A common structure underlies low-frequency cortical dynamics in movement, sleep, and sedation," *Neuron*, vol. 83, no. 5, pp. 1185–1199, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0896627314006333

[304] B. Morillon, L. H. Arnal, C. E. Schroeder, and A. Keitel, "Prominence of delta oscillatory rhythms in the motor cortex and their relevance for auditory and speech perception," *Neuroscience & Biobehavioral Reviews*, vol. 107, pp. 136–142, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0149763419300922

[305] S. Waldert, H. Preissl, E. Demandt, C. Braun, N. Birbaumer, A. Aertsen, and C. Mehring, "Hand movement direction decoded from meg and eeg," *Journal of Neuroscience*, vol. 28, no. 4, pp. 1000–1008, 2008. [Online]. Available: https://www.jneurosci.org/content/28/4/1000

[306] S. Waldert, T. Pistohl, C. Braun, T. Ball, A. Aertsen, and C. Mehring, "A review on directional information in neural signals for brain-machine interfaces," *Journal of Physiology-Paris*, vol. 103, no. 3, pp. 244–254, 2009, neurorobotics. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0928425709000485

[307] A. Hamel-Thibault, F. Thénault, K. Whittingstall, and P.-M. Bernier, "Delta-Band Oscillations in Motor Regions Predict Hand Selection for Reaching," *Cerebral Cortex*, vol. 28, no. 2, pp. 574–584, 12 2016. [Online]. Available: https://academic.oup.com/cercor/article/28/2/574/2725380

[308] Y. Kim, J. Ryu, K. K. Kim, C. C. Took, D. P. Mandic, and C. Park, "Motor imagery classification using mu and beta rhythms of eeg with strong uncorrelating transform based complex common spatial

patterns," *Computational Intelligence and Neuroscience*, vol. 2016, p. 1489692, Oct. 2016. [Online]. Available: https://doi.org/10.1155/2016/1489692

[309] D. T. Bundy and E. C. Leuthardt, "The cortical physiology of ipsilateral limb movements," *Trends Neurosci.*, vol. 42, no. 11, pp. 825–839, Nov. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0166223619301584

[310] K. J. Wisneski, N. Anderson, G. Schalk, M. Smyth, D. Moran, and E. C. Leuthardt, "Unique cortical physiology associated with ipsilateral hand movements and neuroprosthetic implications," *Stroke*, vol. 39, no. 12, pp. 3351–3359, Dec. 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18927456

[311] E. A. Heming, K. P. Cross, T. Takei, D. J. Cook, and S. H. Scott, "Independent representations of ipsilateral and contralateral limbs in primary motor cortex," *eLife*, vol. 8, p. e48190, Oct. 2019. [Online]. Available: https://doi.org/10.7554/eLife.48190

[312] J. Diedrichsen, T. Wiestler, and J. W. Krakauer, "Two distinct ipsilateral cortical representations for individuated finger movements," *Cereb. Cortex*, vol. 23, no. 6, pp. 1362–1377, Jun. 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3643717/

[313] E. Berlot, G. Prichard, J. O'Reilly, N. Ejaz, and J. Diedrichsen, "Ipsilateral finger representations in the sensorimotor cortex are driven by active movement processes, not passive sensory input," *J. Neurophysiol.*, vol. 121, no. 2, pp. 418–426, Feb. 2019. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30517048/

[314] L. Brinkman, A. Stolk, H. C. Dijkerman, F. P. de Lange, and I. Toni, "Distinct roles for alpha- and beta-band oscillations during mental simulation of goal-directed actions," *Journal of Neuroscience*, vol. 34, no. 44, pp. 14 783–14 792, 2014. [Online]. Available: https://www.jneurosci.org/content/34/44/14783

[315] F. Hamzei, C. Dettmers, R. Rzanny, J. Liepert, C. Büchel, and C. Weiller, "Reduction of excitability ("inhibition") in the ipsilateral primary motor cortex is mirrored by fmri signal decreases," *NeuroImage*, vol. 17, no. 1, pp. 490–496, 2002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811902910773

[316] J. M. Newton, A. Sunderland, and P. A. Gowland, "fmri signal decreases in ipsilateral primary motor cortex during unilateral hand movements are related to duration and side of movement," *NeuroImage*, vol. 24, no. 4, pp. 1080–1087, 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811904005944

[317] J. Liepert, C. Dettmers, C. Terborg, and C. Weiller, "Inhibition of ipsilateral motor cortex during phasic generation of low force," *Clinical Neurophysiology*, vol. 112, no. 1, pp. 114–121, 2001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1388245700005034

[318] M. Tinazzi and G. Zanette, “Modulation of ipsilateral motor cortex in man during unimanual finger movements of different complexities,” *Neuroscience Letters*, vol. 244, no. 3, pp. 121–124, 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304394098001505

[319] P. C. Bucy and J. F. Fulton, “IPSILATERAL REPRESENTATION IN THE MOTOR AND PREMOTOR CORTEX OF MONKEYS,” *Brain*, vol. 56, no. 3, pp. 318–342, 09 1933. [Online]. Available: https://academic.oup.com/brain/article/56/3/318/285381

[320] M. Diers, C. Christmann, C. Koeppe, M. Ruf, and H. Flor, “Mirrored, imagined and executed movements differentially activate sensorimotor cortex in amputees with and without phantom limb pain,” *PAIN*, vol. 149, no. 2, May 2010. [Online]. Available: https://journals.lww.com/pain/fulltext/2010/05000/mirrored,_imagined_and_executed_movements.21.aspx

[321] Pancrat and Iamozy. (2014, Jul.) Human motor cortex. CC Attribution-Share Alike 3.0 Unported. [Online]. Available: https://commons.wikimedia.org/wiki/File:Human_motor_cortex.jpg

[322] L. Fadiga and L. Craighero, “Hand actions and speech representation in broca’s area,” *Cortex*, vol. 42, no. 4, pp. 486–490, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010945208703836

[323] F. Geranmayeh, S. L. E. Brownsett, and R. J. S. Wise, “Task-induced brain activity in aphasic stroke patients: what is driving recovery?” *Brain*, vol. 137, no. 10, pp. 2632–2648, Jun. 2014. [Online]. Available: https://academic.oup.com/brain/article/137/10/2632/2846855

[324] G. M. Rojas, C. Alvarez, C. E. Montoya, M. de la Iglesia-Vayá, J. E. Cisternas, and M. Gálvez, “Study of resting-state functional connectivity networks using eeg electrodes position as seed,” *Frontiers in Neuroscience*, vol. 12, Apr. 2018. [Online]. Available: https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2018.00235

[325] M. Xia, J. Wang, and Y. He, “Brainnet viewer: A network visualization tool for human brain connectomics,” *PLOS ONE*, vol. 8, no. 7, pp. 1–15, Jul. 2013. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0068910

[326] D. Muret, V. Root, P. Kieliba, D. Clode, and T. R. Makin, “Beyond body maps: Information content of specific body parts is distributed across the somatosensory homunculus,” *Cell Reports*, vol. 38, no. 11, p. 110523, Mar. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2211124722002595

[327] M. H. Schieber and L. S. Hibbard, “How somatotopic is the motor cortex hand area?” *Science*, vol. 261, no. 5120, pp. 489–492, 1993. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.8332915

[328] C. W. Dunnett, "Pairwise multiple comparisons in the unequal variance case," *Journal of the American Statistical Association*, vol. 75, no. 372, pp. 796–800, 1980. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1980.10477552

[329] M. C. Shingala and A. Rajyaguru, "Comparison of post hoc tests for unequal variance," in *International Journal of New Technologies in Science and Engineering*, vol. 2, Nov. 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:212480596

[330] T. Pohlert, *PMCMRplus: Calculate Pairwise Multiple Comparisons of Mean Rank Sums Extended*, 2020, r package version 1.7.0. [Online]. Available: https://CRAN.R-project.org/package=PMCMRplus

[331] B. S. Holland and M. D. Copenhaver, "An improved sequentially rejective bonferroni test procedure," *Biometrics*, vol. 43, no. 2, pp. 417–423, Jun. 1987. [Online]. Available: http://www.jstor.org/stable/2531823

[332] G. Hommel and G. Bernhard, "Bonferroni procedures for logically related hypotheses," *Journal of Statistical Planning and Inference*, vol. 82, no. 1, pp. 119–128, Dec. 1999. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S037837589900035X

[333] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. Mullers, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (Cat. No.98TH8468)*, 1999, pp. 41–48. [Online]. Available: https://ieeexplore.ieee.org/document/788121

[334] S. Tortora, L. Tonin, C. Chisari, S. Micera, E. Menegatti, and F. Artoni, "Hybrid human-machine interface for gait decoding through bayesian fusion of eeg and emg classifiers," *Frontiers in Neurorobotics*, vol. 14, 2020. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85097068162&doi=10.3389%2ffnbot.2020.582728&partnerID=40&md5=65c1b696e5a60ca2be928a10e96af8a3

[335] R. O. R. Tasé, D. D. Rodríguez, O. W. Samuel, and A. L. Delis, "A hybrid brain-computer interface using extreme learning machines for motor intention detection," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13055 LNCS, p. 115 – 123, 2021. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85119827605&doi=10.1007%2f978-3-030-89691-1_12&partnerID=40&md5=9cc1a6b5a47a324356bb5023ec246cd9

[336] J.-H. Cho, J.-H. Jeong, and S.-W. Lee, "Neurograsp: Real-time eeg classification of high-level motor imagery tasks using a dual-stage deep learning framework," *IEEE Transactions on Cybernetics*, vol. 52, no. 12, pp. 13 279–13 292, 2022. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9606552

[337] E. Dagois, A. Khalaf, E. Sejdic, and M. Akcakaya, “Transfer learning for a multimodal hybrid eeg-ftcd brain–computer interface,” *IEEE Sensors Letters*, vol. 3, no. 1, pp. 1–4, Jan. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8520792

[338] A. Khalaf, M. Sybeldon, E. Sejdic, and M. Akcakaya, “A brain-computer interface based on functional transcranial doppler ultrasound using wavelet transform and support vector machines,” *Journal of Neuroscience Methods*, vol. 293, pp. 174–182, Jan. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165027017303515

[339] S. Kullback and R. A. Leibler, “On information and sufficiency,” *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, Mar. 195. [Online]. Available: https://doi.org/10.1214/aoms/1177729694

[340] H. Min, Z. Chen, B. Fang, Z. Xia, Y. Song, Z. Wang, Q. Zhou, F. Sun, and C. Liu, “Cross-individual gesture recognition based on long short-term memory networks,” *Scientific Programming*, vol. 2021, p. 6680417, Jul. 2021. [Online]. Available: https://www.hindawi.com/journals/sp/2021/6680417/

[341] Z. Wan, R. Yang, M. Huang, N. Zeng, and X. Liu, “A review on transfer learning in eeg signal analysis,” *Neurocomputing*, vol. 421, pp. 1–14, Jan. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231220314223

[342] T. F. Chan, G. H. Golub, and R. J. LeVeque, “Updating formulae and a pairwise algorithm for computing sample variances,” Stanford University Computer Science, Stanford, CA, USA, Tech. Rep., Nov. 1979, sTAN-CS-79-773. [Online]. Available: http://i.stanford.edu/pub/cstr/reports/cs/tr/79/773/CS-TR-79-773.pdf

[343] N. Bacaër, *Verhulst and the logistic equation (1838).* London: Springer London, 2011, pp. 35–39. [Online]. Available: https://link.springer.com/chapter/10.1007/978-0-85729-115-8_6

[344] A. Sultana, F. Ahmed, and M. S. Alam, “A systematic review on surface electromyography-based classification system for identifying hand and finger movements,” *Healthcare Analytics*, vol. 3, p. 100126, Nov. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772442522000661

[345] M. Yokoyama and M. Yanagisawa, “Logistic regression analysis of multiple interosseous hand-muscle activities using surface electromyography during finger-oriented tasks,” *Journal of Electromyography and Kinesiology*, vol. 44, pp. 117–123, Feb. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1050641118303791

[346] V. H. Cene and A. Balbinot, “Upper-limb movement classification through logistic regression semg signal processing,” in *2015 Latin America Congress on Computational Intelligence (LA-CCI)*, 2015, pp. 1–5.

[347] Y. Diao, Q. Chen, Y. Liu, L. He, Y. Sun, X. Li, Y. Chen, G. Li, and G. Zhao, “A fuzzy granular logistic regression algorithm for semg-based cross-individual prosthetic hand gesture

classification," *Journal of Neural Engineering*, vol. 20, no. 2, p. 026029, Apr. 2023. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2552/acc42a

[348] V. Barigala, S. Peddapalli, P. Govarthan, S. Pj, M. Aasaithambi, N. Ganapathy, K. Pa, D. Kumar, and J. Fredo, "Identifying the optimal location of facial emg for emotion detection using logistic regression," *Studies in health technology and informatics*, vol. 305, pp. 81–84, Jun. 2023, volume 305: Healthcare Transformation with Informatics and Artificial Intelligence. [Online]. Available: https://ebooks.iospress.nl/doi/10.3233/SHTI230429

[349] R. A. Khan, N. Rashid, M. Shahzaib, U. F. Malik, A. Arif, J. Iqbal, M. Saleem, U. S. Khan, and M. Tiwana, "A novel framework for classification of two-class motor imagery EEG signals using logistic regression classification algorithm," *PLoS One*, vol. 18, no. 9, p. e0276133, Sep. 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10490872/

[350] R. Tomioka, K. Aihara, and K.-R. Müller, "Logistic regression for single trial eeg classification," in *Proceedings of the 19th International Conference on Neural Information Processing Systems*, ser. NIPS'06. Cambridge, MA, USA: MIT Press, Dec. 2006, p. 1377–1384. [Online]. Available: https://dl.acm.org/doi/abs/10.5555/2976456.2976629

[351] A. Subasi and E. Erçelebi, "Classification of eeg signals using neural network and logistic regression," *Computer Methods and Programs in Biomedicine*, vol. 78, no. 2, pp. 87–99, May 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260705000246

[352] D.-W. Chen, R. Miao, Z.-Y. Deng, Y.-Y. Lu, Y. Liang, and L. Huang, "Sparse logistic regression with l1/2 penalty for emotion recognition in electroencephalography classification," *Frontiers in Neuroinformatics*, vol. 14, 2020. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fninf.2020.00029/full

[353] M. Schmidt, N. Le Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," *Mathematical Programming*, vol. 162, no. 1, pp. 83–112, Mar. 2017. [Online]. Available: https://link.springer.com/article/10.1007/s10107-016-1030-6

[354] H.-L. Halme and L. Parkkonen, "Across-subject offline decoding of motor imagery from MEG and EEG," *Sci. Rep.*, vol. 8, no. 1, p. 10087, Jul. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6031658/

[355] K. H. Lee, J. Y. Min, and S. Byun, "Electromyogram-based classification of hand and finger gestures using artificial neural networks," *Sensors*, vol. 22, no. 1, Dec. 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/1/225

[356] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:160025533

[357] R. Niu, Y. Wang, H. Xi, Y. Hao, and M. Zhang, “Epileptic seizure prediction by synthesizing eeg signals through gpt,” in *Proceedings of the 2021 4th International Conference on Artificial Intelligence and Pattern Recognition*, ser. AIPR ’21. New York, NY, USA: Association for Computing Machinery, Sep. 2021, p. 419–423. [Online]. Available: https://dl.acm.org/doi/10.1145/3488933.3489016

[358] F. P. Carrle, Y. Hollenbenders, and A. Reichenbach, “Generation of synthetic eeg data for training algorithms supporting the diagnosis of major depressive disorder,” *Frontiers in Neuroscience*, vol. 17, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2023.1219133

[359] S. J. Lehmler, M. S. ur Rehman, G. Tobias, and I. Iossifidis, “Deep transfer learning compared to subject-specific models for semg decoders,” *Journal of Neural Engineering*, vol. 19, no. 5, p. 056039, Oct. 2022. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2552/ac9860/

[360] M. Ison and P. Artemiadis, “The role of muscle synergies in myoelectric control: trends and challenges for simultaneous multifunction control,” *Journal of Neural Engineering*, vol. 11, no. 5, p. 051001, Sep. 2014. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2560/11/5/051001

[361] L. Hargrove, K. Englehart, and B. Hudgins, “The effect of electrode displacements on pattern recognition based myoelectric control,” in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, 2006, pp. 2203–2206. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/4462227

[362] M. Krauledat, M. Tangermann, B. Blankertz, and K.-R. Müller, “Towards zero training for brain-computer interfacing,” *PLOS ONE*, vol. 3, no. 8, pp. 1–12, Aug. 2008. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0002967

[363] Y. Li, C. Guan, H. Li, and Z. Chin, “A self-training semi-supervised svm algorithm and its application in an eeg-based brain computer interface speller system,” *Pattern Recognition Letters*, vol. 29, no. 9, pp. 1285–1294, Jul. 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016786550800055X

[364] A. Abu-Rmileh, E. Zakkay, L. Shmuelof, and O. Shriki, “Co-adaptive training improves efficacy of a multi-day EEG-based motor imagery BCI training,” *Front. Hum. Neurosci.*, vol. 13, p. 362, Oct. 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6802491/

[365] S. M. Ling, R. A. Conwit, L. Ferrucci, and E. J. Metter, “Age-associated changes in motor unit physiology: observations from the baltimore longitudinal study of aging,” *Archives of physical medicine and rehabilitation*, vol. 90, no. 7, pp. 1237–1240, Jul. 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5496096/

[366] D. C. Sauder and C. E. DeMars, “An updated recommendation for multiple comparisons,” *Advances in Methods and Practices in Psychological Science*, vol. 2, no. 1, pp. 26–44, Jan. 2019. [Online]. Available: https://journals.sagepub.com/doi/10.1177/2515245918808784

[367] N. Jiang, I. Vujaklija, H. Rehbaum, B. Graimann, and D. Farina, "Is accurate mapping of emg signals on kinematics needed for precise online myoelectric control?" *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 3, pp. 549–558, 2014. [Online]. Available: https://ieeexplore.ieee.org/document/6648468

[368] S. Arjunan, D. Kumar, C. Kalra, J. Burne, and T. Bastos, "Effect of age and gender on the surface electromyogram during various levels of isometric contraction," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2011, pp. 3853–3856. [Online]. Available: https://ieeexplore.ieee.org/document/6090957/

[369] H. E. Künzel, H. Murck, G. K. Stalla, and A. Steiger, "Changes in the sleep electroencephalogram (eeg) during male to female transgender therapy," *Psychoneuroendocrinology*, vol. 36, no. 7, pp. 1005–1009, Aug. 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306453011000023

[370] M. Hazin, C. W. S. Ferreira, R. Andrade, E. Moretti, D. R. da Silva, J. H. Policarpo, L. Barbosa, and A. Lemos, "Assessment of the strength and electrical activity of the pelvic floor muscles of male-to-female transgender patients submitted to gender-affirming surgery: A case series," *Neurourology and Urodynamics*, vol. 40, no. 6, pp. 1625–1633, Jun. 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/nau.24728

[371] P. Harpur and M. A. Stein, "The convention on the rights of persons with disabilities as a global tipping point for the participation of persons with disabilities," Oct. 2022. [Online]. Available: https://oxfordre.com/politics/view/10.1093/acrefore/9780190228637.001.0001/acrefore-9780190228637-e-245

[372] g.tec medical engineering GmbH, "g.PANGOLIN ULTRA HIGH-DENSITY EEG/EMG/ECG." [Online]. Available: https://www.gtec.at/product/g-pangolin-electrodes/

[373] ——, "g.SAHARA HYBRID ACTIVE HYBRID EEG ELECTRODES." [Online]. Available: https://www.gtec.at/product/g-sahara-hybrid-eeg-electrodes/

[374] A. Etienne, T. Laroia, H. Weigle, A. Afelin, S. K. Kelly, A. Krishnan, and P. Grover, "Novel electrodes for reliable eeg recordings on coarse and curly hair," *bioRxiv*, Feb. 2020. [Online]. Available: https://www.biorxiv.org/content/early/2020/02/27/2020.02.26.965202

[375] L. Sanders, "New electrodes can better capture brain waves of people with natural hair," in *Science News*. Society for Science, Mar. 2020. [Online]. Available: https://www.sciencenews.org/article/electrodes-brain-waves-eeg-black-african-american-natural-hair

[376] J. J. Bird, "A socially interactive multimodal human-robot interaction framework through studies on machine and deep learning," PhD thesis, Aston University, United Kingdom, 2021, available at https://publications.aston.ac.uk/id/eprint/43472/.

Supplementary Data

A.1 Supplementary Data for Chapter 5

A.1.1 Post-hoc analysis of modelling hyperparameters

A.1.1.1 Linear Discriminant Analysis Solvers

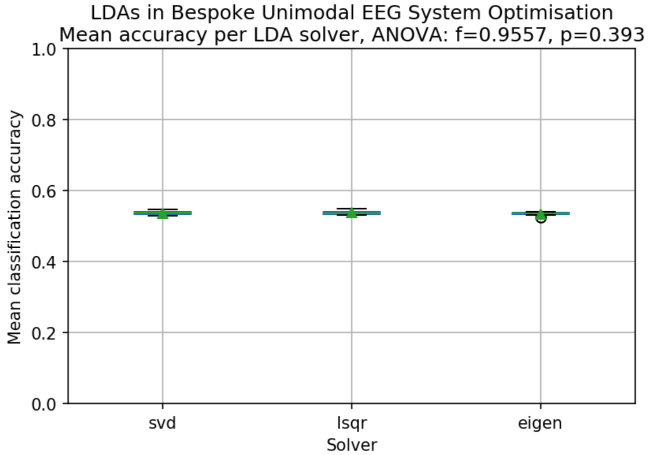


Figure A.1: Development Set accuracies achieved by different LDA solvers in CASH optimisation of Bespoke Unimodal EEG system

Hypothesis	t value	p value
LSQR – SVD	0.475	0.9508
Eigen – SVD	-0.991	0.6966
Eigen – LSQR	-1.306	0.5028

Table A.1: Dunnett's T3 comparisons of LDA solvers in Bespoke Unimodal EEG system

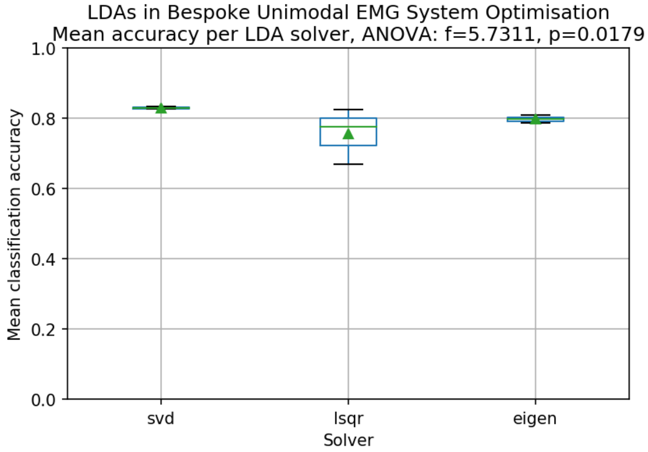
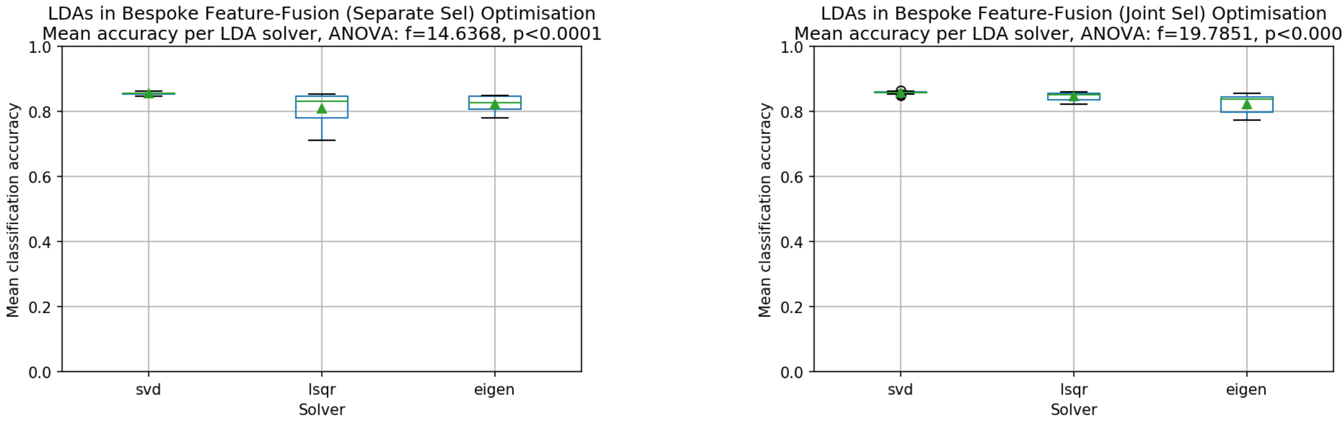


Figure A.2: Development Set accuracies achieved by different LDA solvers in CASH optimisation of Bespoke Unimodal EMG system

Hypothesis	t value	p value
LSQR – SVD	-1.571	0.4837
Eigen – SVD	-2.712	0.3669
Eigen – LSQR	0.858	0.7884

Table A.2: Dunnett's T3 comparisons of LDA solvers in Bespoke Unimodal EMG system



(a) Separate selection

(b) Joint selection

Figure A.3: Development Set accuracies achieved by different LDA solvers in CASH optimisation of Bespoke Feature-Level Fusion system with separate feature selection (left) & joint feature selection (right)

Hypothesis	t value	p value
LSQR – SVD	-2.560	0.0994
Eigen – SVD	-2.793	0.0973
Eigen – LSQR	0.664	0.8781

(a) Separate selection

Hypothesis	t value	p value
LSQR – SVD	-2.639	0.0803
Eigen – SVD	-2.936	0.0685
Eigen – LSQR	-1.883	0.2436

(b) Joint selection

Table A.3: Pairwise comparisons of LDA solvers' Development Set accuracies in CASH optimisation of Bespoke Feature-Level Fusion systems with separate feature selection (left) & joint feature selection (right), using Dunnett's T3 test.

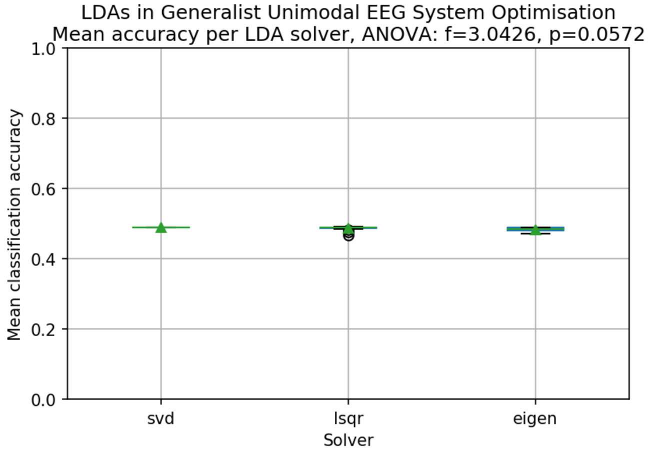


Figure A.4: Development Set accuracies achieved by different LDA solvers in CASH optimisation of Generalist Unimodal EEG system

Hypothesis	t value	p value
LSQR – SVD	-0.883	0.7593
Eigen – SVD	-1.964	0.2268
Eigen – LSQR	-1.605	0.3543

Table A.4: Dunnett's T3 comparisons of LDA solvers in Generalist Unimodal EEG system

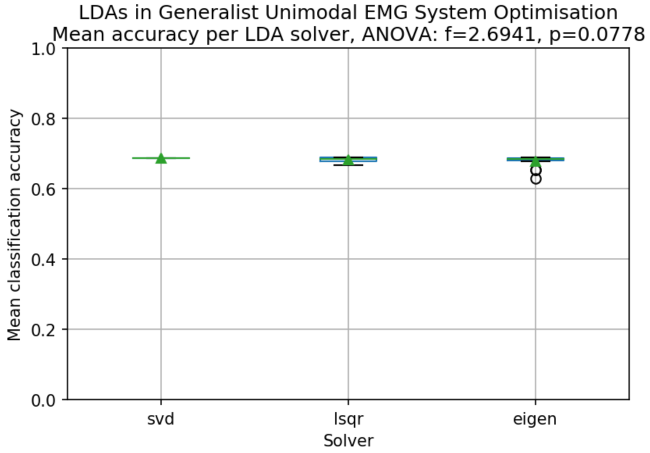


Figure A.5: Development Set accuracies achieved by different LDA solvers in CASH optimisation of Generalist Unimodal EMG system

Hypothesis	t value	p value
LSQR – SVD	-2.925	**0.0242**
Eigen – SVD	-1.985	0.1802
Eigen – LSQR	-1.018	0.6759

Table A.5: Dunnett's T3 comparisons of LDA solvers in Generalist Unimodal EMG system

A.1.1.2 Shrinkage in LDAs

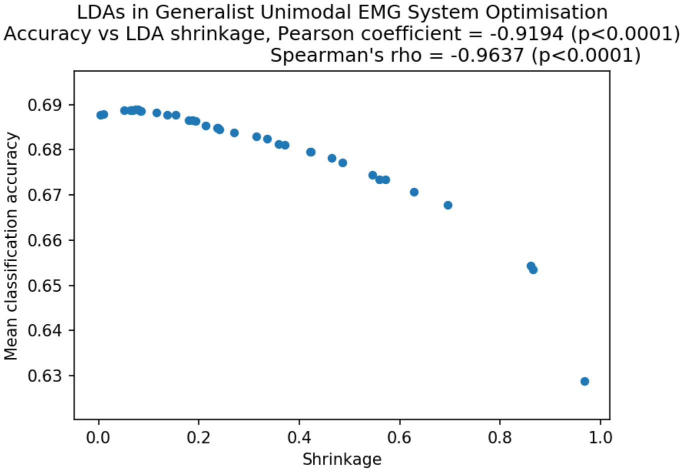
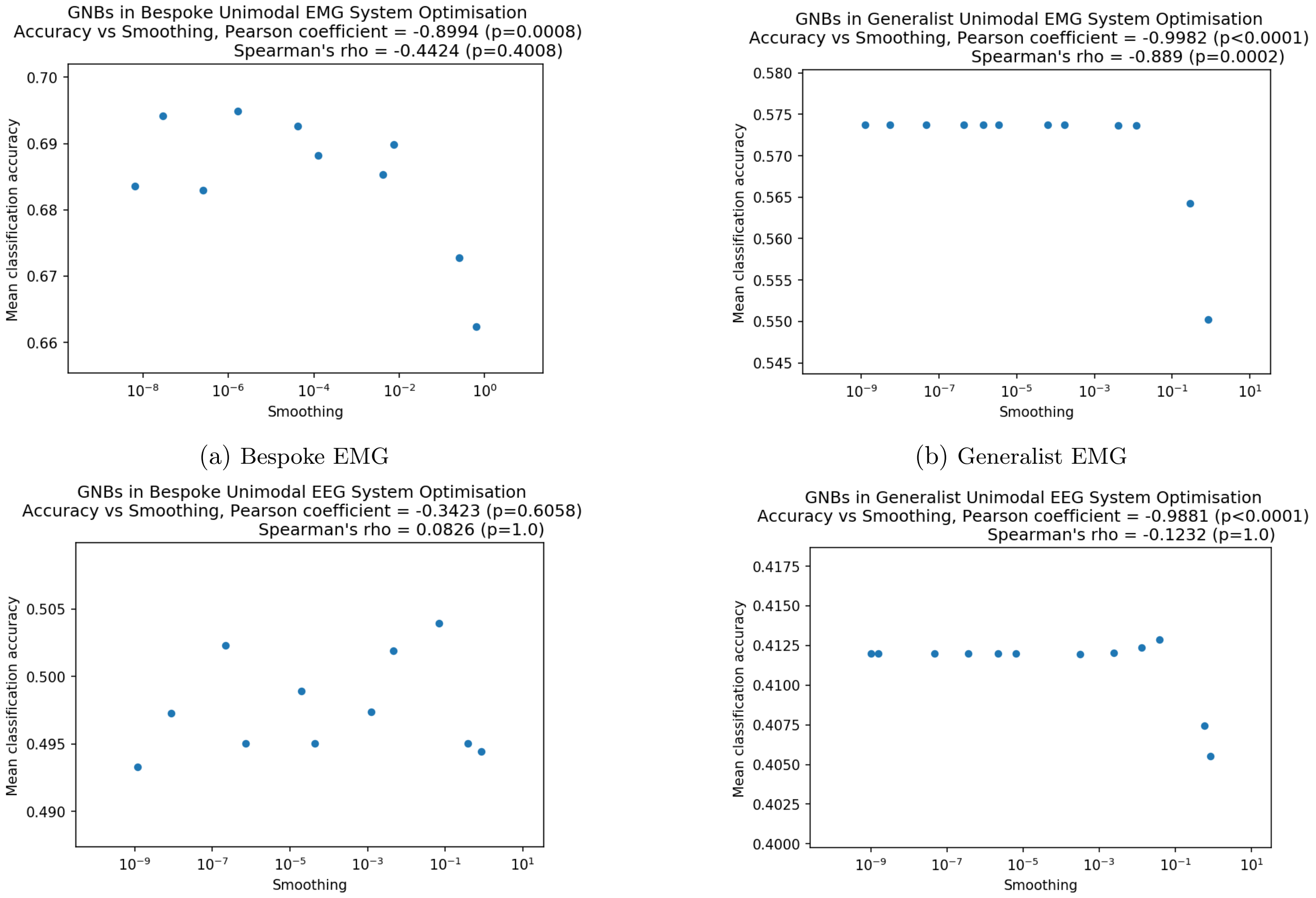


Figure A.6: Development Set accuracy against LDA Shrinkage in CASH optimisation of Generalist Unimodal EMG system. Reported p-values adjusted by Bonferroni correction.

System	Solver	Pearson's R		Spearman's Rho	
		r	p	ρ	p
Bespoke Unimodal EEG	Least Squares Solution	-0.2088	0.8114	-0.0877	1.0
	Eigenvalue Decomposition	-0.5835	0.3380	-0.3214	0.9641
Bespoke Unimodal EMG	Least Squares Solution	-0.9177	0.5202	-1.0	<0.0001
	Eigenvalue Decomposition	N/A*		N/A*	
Bespoke Feature Fusion (Separate Selection)	Least Squares Solution	-0.9276	0.0018	-1.0	<0.0001
	Eigenvalue Decomposition	-0.9735	0.0021	-0.9429	0.0096
Bespoke Feature Fusion (Joint Selection)	Least Squares Solution	-0.9783	<0.0001	-0.9500	0.0002
	Eigenvalue Decomposition	-0.9612	0.0011	-0.9643	0.0009
Generalist Unimodal EEG	Least Squares Solution	-0.8872	<0.0001	-0.8594	<0.0001
	Eigenvalue Decomposition	-0.9640	0.0002	-0.9524	0.0005
Generalist Unimodal EMG	Least Squares Solution	-0.9856	<0.0001	-0.9658	<0.0001
	Eigenvalue Decomposition	-0.9445	<0.0001	-0.9607	<0.0001
Generalist Feature Fusion (Separate Selection)	Least Squares Solution	-0.9113	0.0002	-0.9294	0.0001
	Eigenvalue Decomposition	-0.9595	<0.0001	-0.9257	<0.0001
Generalist Feature Fusion (Joint Selection)	Least Squares Solution	-0.8585	<0.0001	-0.5155	0.0043
	Eigenvalue Decomposition	-0.8398	0.0361	-0.9643	0.0009

Table A.6: Correlation coefficients between LDA shrinkage & Development Set accuracy, broken down by LDA solver. Reported p-values for each system type adjusted by Bonferroni correction.
NB: As shrinkage was optimised across both of these solvers together, we cannot assume any distinctions are genuinely related to the solvers having different interactions with shrinkage.
* Only two LDAs using the Eigenvalue Decomposition solver were trialled in Bespoke Unimodal EMG optimisation; correlational analysis is not suitable here.

A.1.1.3 Smoothing in Gaussian Naïve Bayes Models



(a) Bespoke EMG

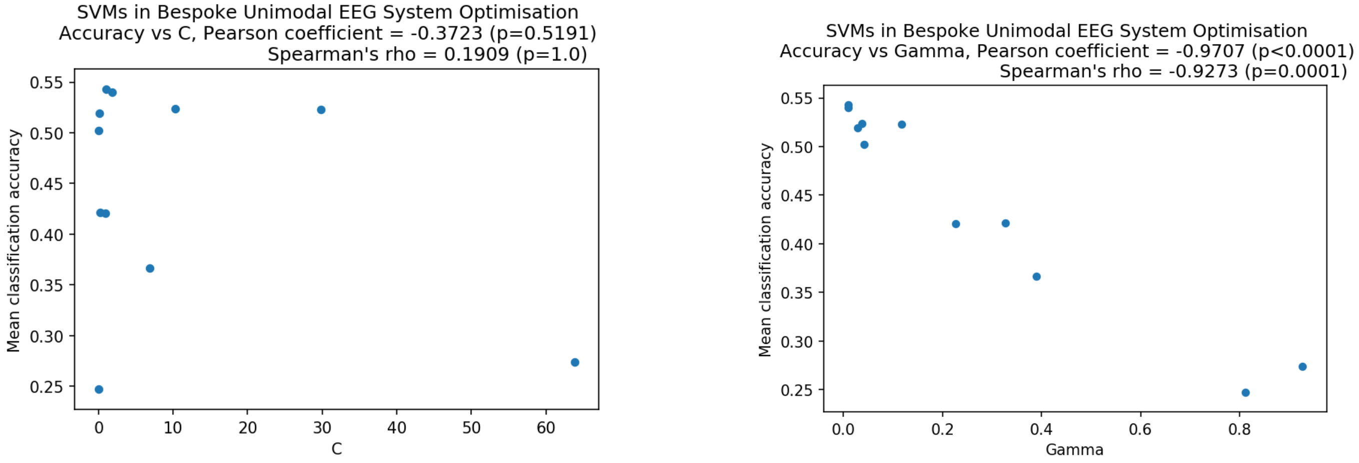
(b) Generalist EMG

(c) Bespoke EEG

(d) Generalist EEG

Figure A.7: Development Set accuracy against Gaussian Naïve Bayes Smoothing in CASH optimisation of Unimodal EMG (above) and EEG (below) systems in both Bespoke (left) and Generalist (right) cases.
Reported p-values adjusted by Bonferroni correction.

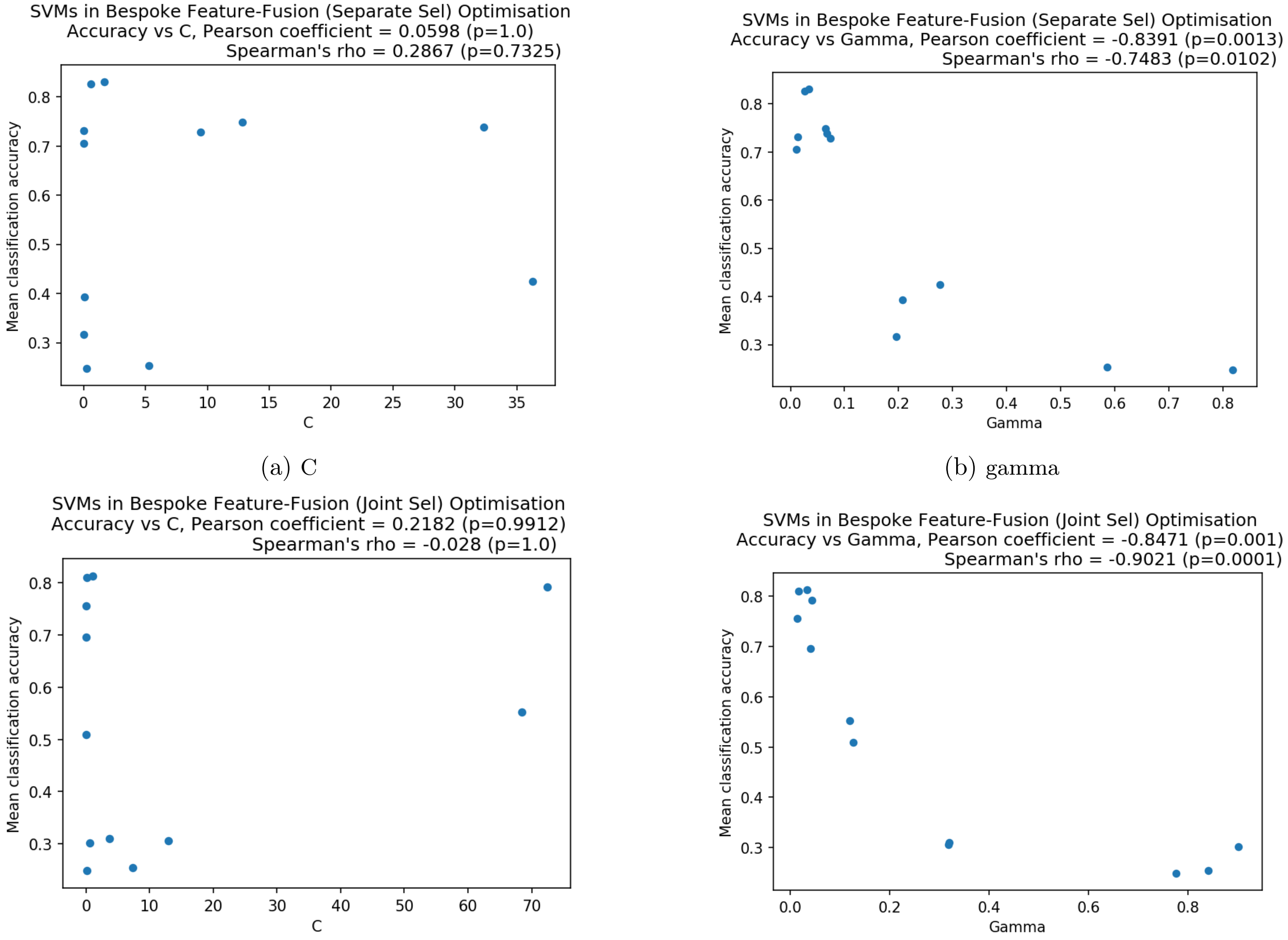
A.1.1.4 C and γ in non-EMG Support Vector Machines



(a) C

(b) gamma

Figure A.8: Influence of SVM hyperparameters C (left) and gamma (right) on Development Set accuracy in CASH optimisation of Bespoke Unimodal EEG system. Reported p-values adjusted by Bonferroni correction.



(a) C

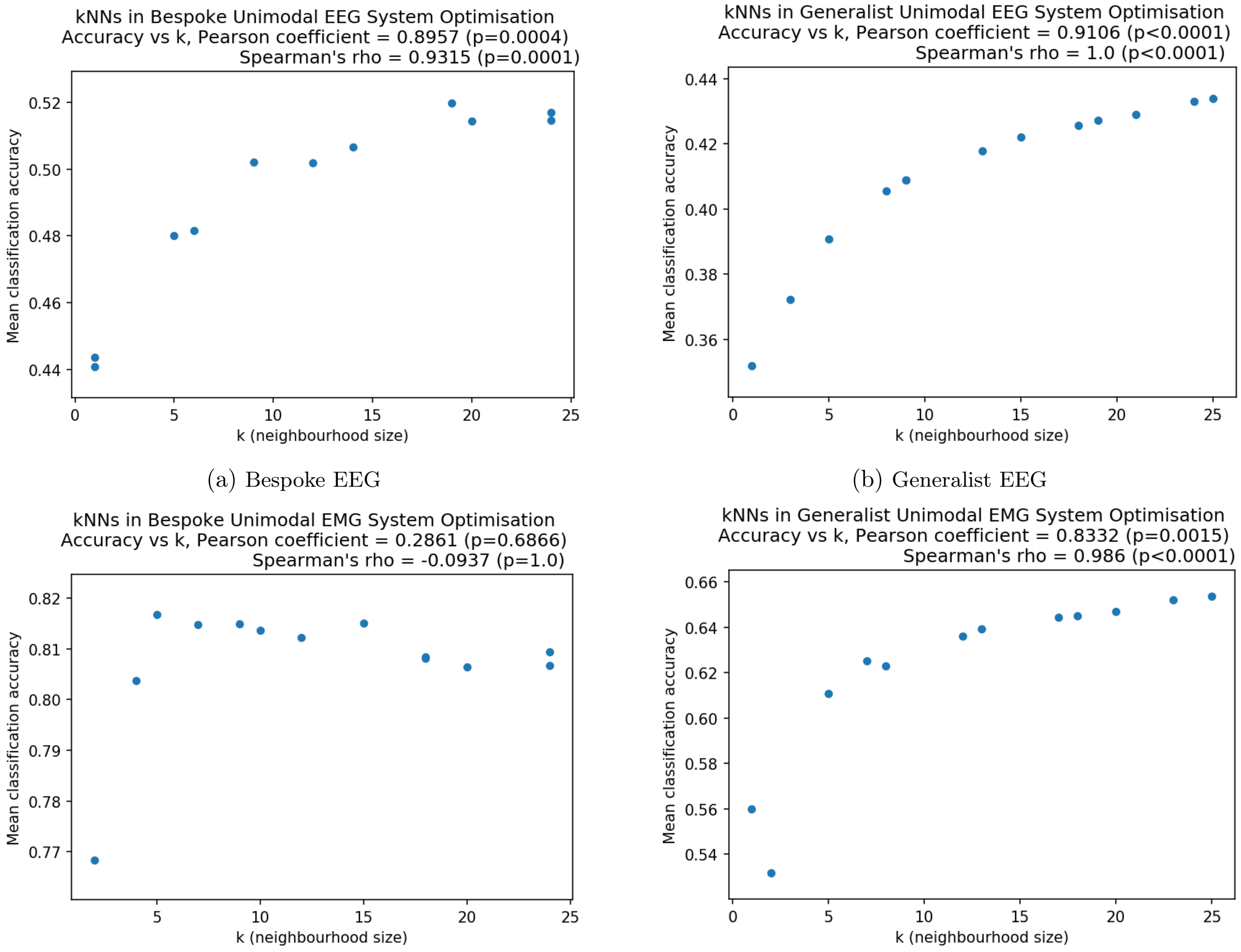
(b) gamma

(c) C

(d) gamma

Figure A.9: Influence of SVM hyperparameters C (left) and gamma (right) on Development Set accuracy in CASH optimisation of Bespoke Feature-level Fusion with Separate (above) and Joint (below) feature selection.
Reported p-values adjusted by Bonferroni correction.

A.1.1.5 k in k- Nearest Neighbour models



(a) Bespoke EEG

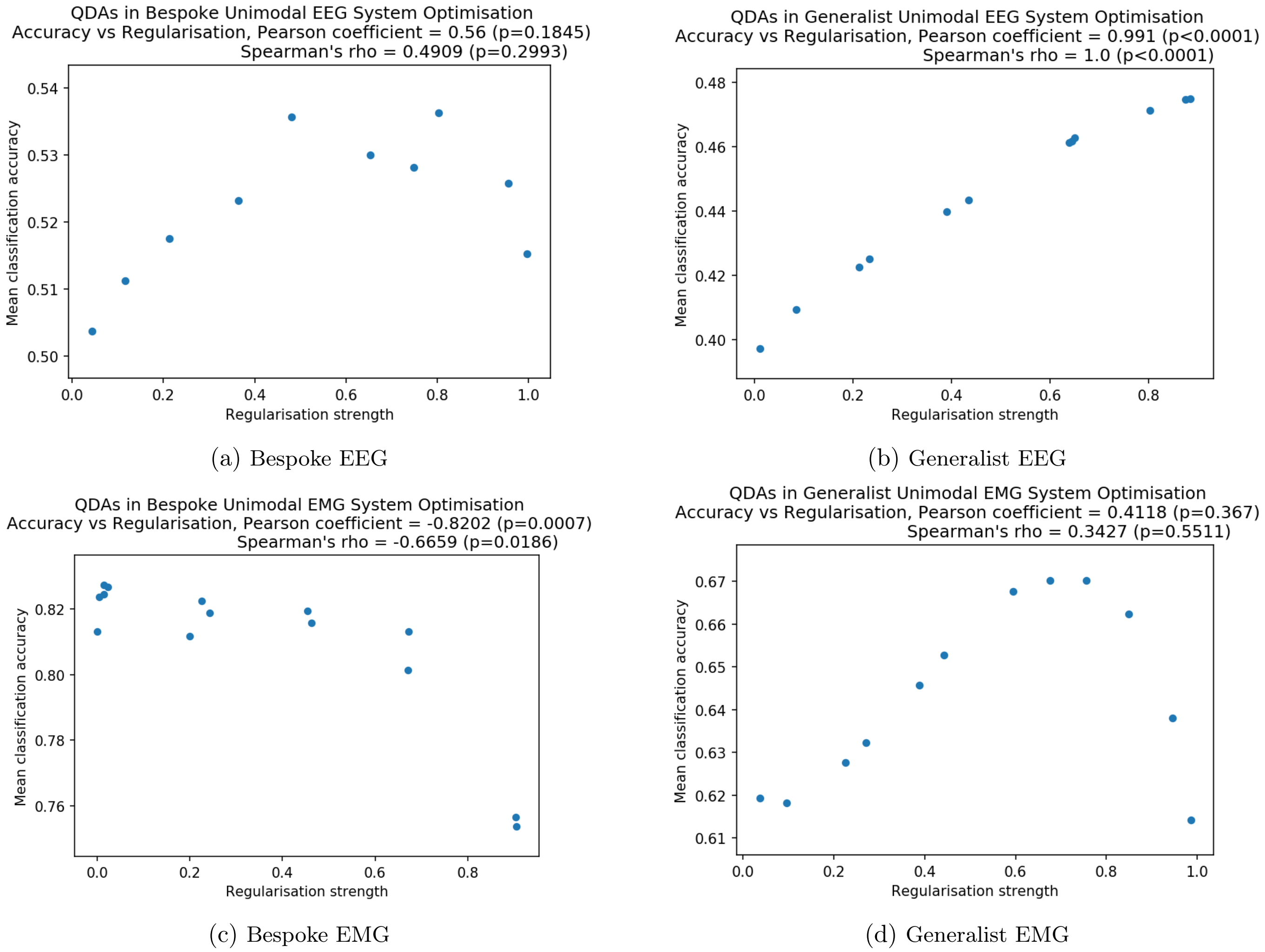
(b) Generalist EEG

(c) Bespoke EMG

(d) Generalist EMG

Figure A.10: Influence of neighbourhood size hyperparameter k on Unimodal EEG (above) and EMG (below) kNNs' Development Set accuracy in CASH optimisation of Bespoke (left) and Generalist (right) cases.
Reported p-values adjusted by Bonferroni correction.

A.1.1.6 Regularisation in Quadratic Discriminant Analysis models



QDAs in Bespoke Unimodal EEG System Optimisation
Accuracy vs Regularisation, Pearson coefficient = 0.56 (p=0.1845)
Spearman's rho = 0.4909 (p=0.2993)

(a) Bespoke EEG

QDAs in Generalist Unimodal EEG System Optimisation
Accuracy vs Regularisation, Pearson coefficient = 0.991 (p<0.0001)
Spearman's rho = 1.0 (p<0.0001)

(b) Generalist EEG

QDAs in Bespoke Unimodal EMG System Optimisation
Accuracy vs Regularisation, Pearson coefficient = -0.8202 (p=0.0007)
Spearman's rho = -0.6659 (p=0.0186)

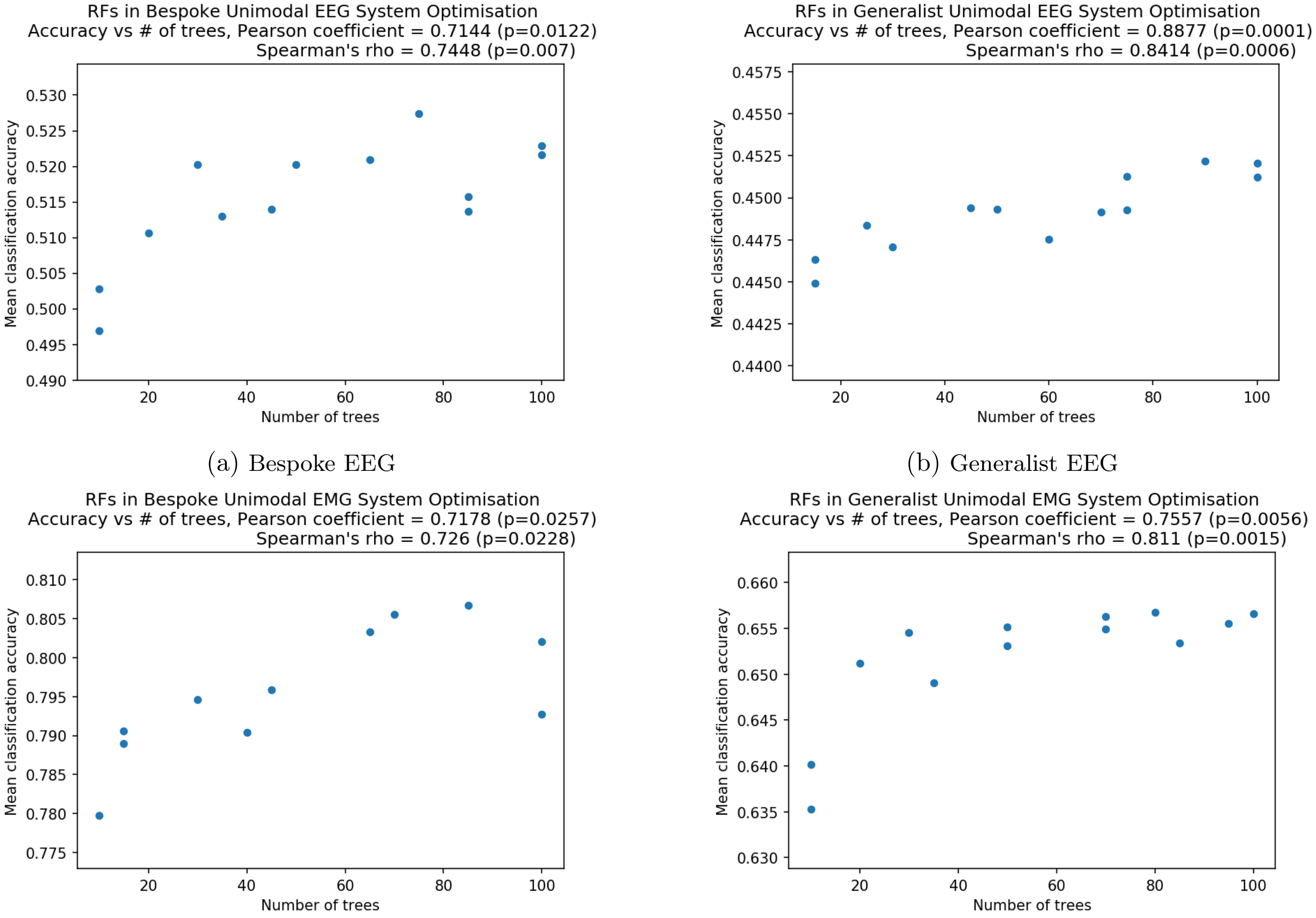
(c) Bespoke EMG

QDAs in Generalist Unimodal EMG System Optimisation
Accuracy vs Regularisation, Pearson coefficient = 0.4118 (p=0.367)
Spearman's rho = 0.3427 (p=0.5511)

(d) Generalist EMG

Figure A.11: Influence of Regularisation hyperparameter on Unimodal EEG (above) and EMG (below) QDAs' Development Set accuracy in CASH optimisation of Bespoke (left) and Generalist (right) cases.
Reported p-values adjusted by Bonferroni correction.

A.1.1.7 Number of trees in Random Forests



(a) Bespoke EEG

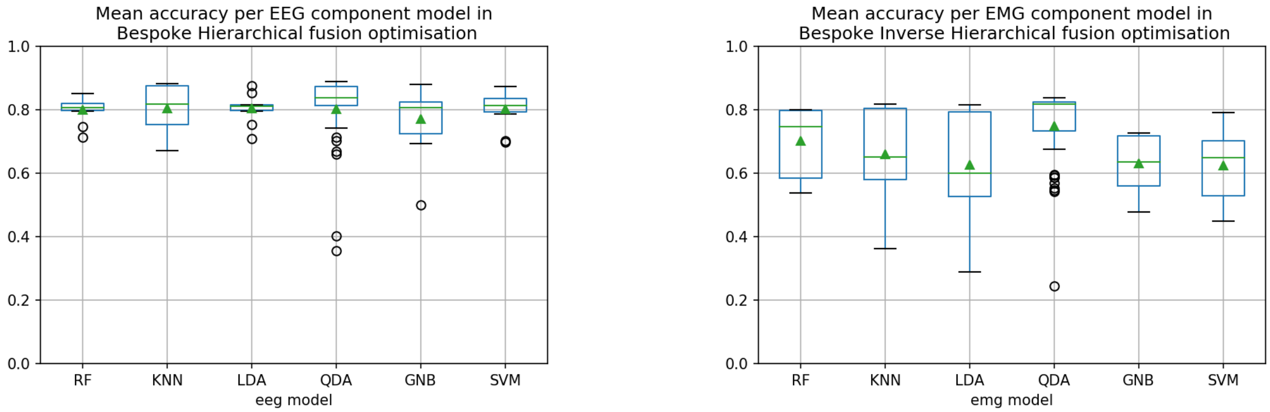
(b) Generalist EEG

(c) Bespoke EMG

(d) Generalist EMG

Figure A.12: Influence of number of trees (forest size) on Unimodal EEG (above) and EMG (below) Random Forests' Development Set accuracy in CASH optimisation of Bespoke (left) and Generalist (right) cases.
Reported p-values adjusted by Bonferroni correction.

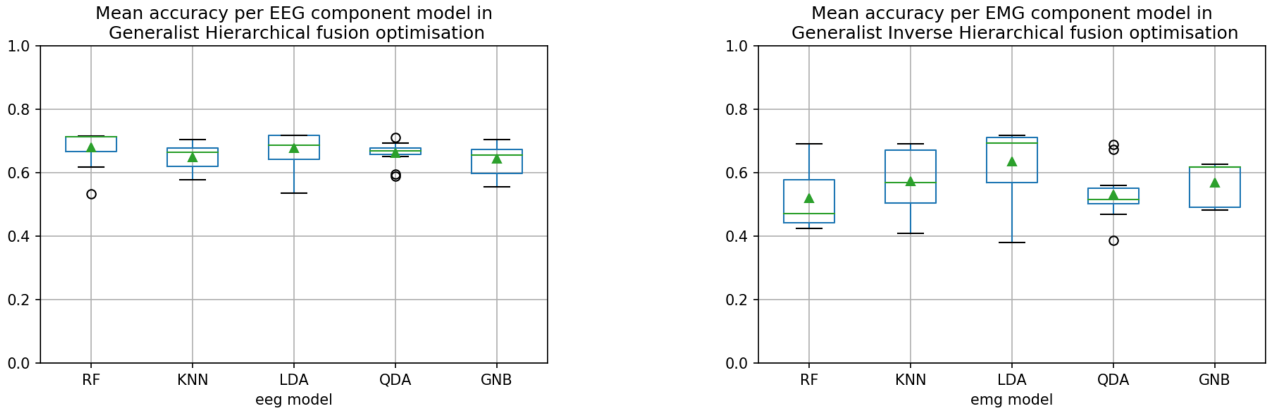
A.1.2 Lower-level model choices in Hierarchical & Inverse Hierarchical systems



(a) Hierarchical (where EMG is top rank)
ANOVA: f=0.2805, p=0.9227

(b) Inverse Hierarchical (where EEG is top rank)
ANOVA: f=3.6649, p=0.0045

Figure A.13: Mean accuracies across development set subjects achieved by different lower-level models in optimisation of Bespoke Hierarchical & Inverse Hierarchical systems



(a) Hierarchical (where EMG is top rank)
ANOVA: f=2.0657, p=0.0914

(b) Inverse Hierarchical (where EEG is top rank)
ANOVA: f=6.2277, p=0.0001

Figure A.14: Mean accuracies across development set subjects achieved by different lower-level models in optimisation of Generalist Hierarchical & Inverse Hierarchical systems

A.1.3 Decision-Level Fusion

A.1.3.1 Full pairwise comparisons of Decision-Level Fusion algorithms

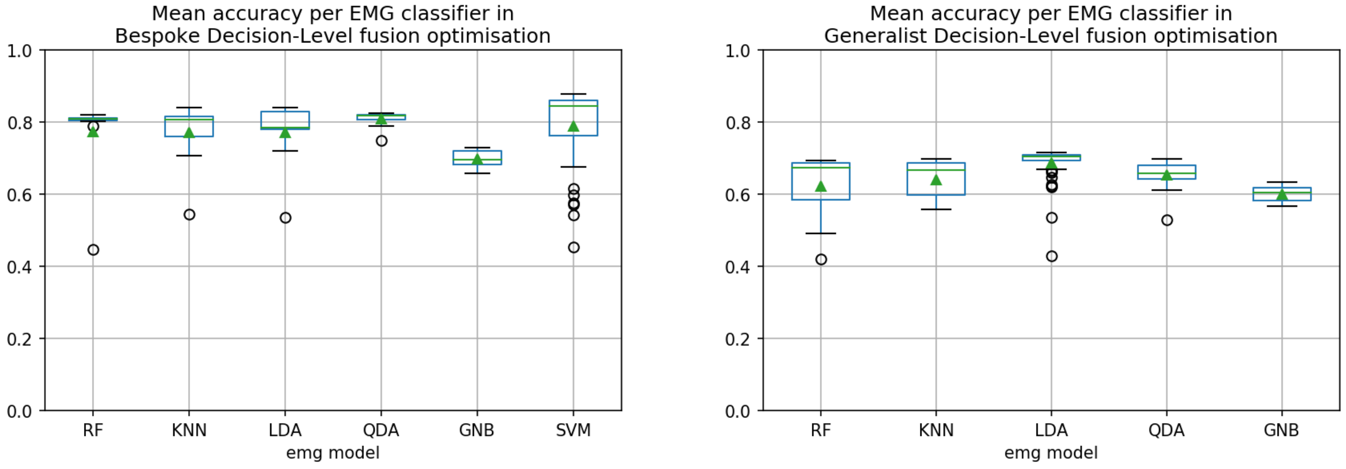
Hypothesis	Estimate	Std. Err	z value	p value
3:1 EEG – Max	-0.10617	0.03053	-3.477	**0.0114**
3:1 EMG – Max	0.03568	0.02319	1.539	0.7799
LDA – Max	0.06173	0.03083	2.002	0.4715
Mean – Max	0.02597	0.02916	0.891	0.9864
Tunable Weight – Max	-0.06436	0.02962	-2.173	0.3603
RF – Max	0.03811	0.02937	1.298	0.8962
SVM – Max	0.06088	0.02429	2.506	0.1872
3:1 EMG – 3:1 EEG	0.14185	0.03213	4.415	$<$ **0.001**
LDA – 3:1 EEG	0.16790	0.03842	4.371	$<$ **0.001**
Mean – 3:1 EEG	0.13214	0.03608	3.662	**0.0055**
Tunable Weight – 3:1 EEG	0.04181	0.03627	1.153	0.9426
RF – 3:1 EEG	0.14428	0.03610	3.997	**0.0016**
SVM – 3:1 EEG	0.16705	0.03240	5.156	$<$ **0.001**
LDA – 3:1 EMG	0.02605	0.03182	0.819	0.9917
Mean – 3:1 EMG	-0.00971	0.03050	-0.318	1.0000
Tunable Weight – 3:1 EMG	-0.10004	0.03119	-3.207	**0.0281**
RF – 3:1 EMG	0.00243	0.03112	0.078	1.0000
SVM – 3:1 EMG	0.02520	0.02575	0.979	0.9765
Mean – LDA	-0.03576	0.03678	-0.972	0.9773
Tunable Weight – LDA	-0.12609	0.03651	-3.454	**0.0124**
RF – LDA	-0.02362	0.03665	-0.645	0.9981
SVM – LDA	-0.00085	0.03205	-0.026	1.0000
Tunable Weight – Mean	-0.09033	0.03493	-2.586	0.1559
RF – Mean	0.01214	0.03507	0.346	1.0000
SVM – Mean	0.03492	0.03148	1.109	0.9531
RF – Tunable Weight	0.10247	0.03534	2.900	0.0704
SVM – Tunable Weight	0.12525	0.03140	3.989	**0.0016**
SVM – RF	0.02277	0.03103	0.734	0.9958

Table A.7: Tukey of all decision fusion algs for Bespoke

Hypothesis	Estimate	Std. Error	z value	p value
3:1 EEG – Max	-0.07725	0.01769	-4.367	$<$ **0.001**
3:1 EMG – Max	0.01667	0.01964	0.849	0.9896
LDA – Max	0.07235	0.01987	3.661	**0.0062**
Mean – Max	-0.01323	0.01576	-0.840	0.9903
Tunable Weight – Max	-0.02098	0.01844	-1.138	0.9460
RF – Max	0.02080	0.01770	1.175	0.9360
SVM – Max	0.02715	0.01609	1.688	0.6859
3:1 EMG – 3:1 EEG	0.09392	0.01764	5.325	$<$ **0.001**
LDA – 3:1 EEG	0.14998	0.01766	8.492	$<$ **0.001**
Mean – 3:1 EEG	0.06401	0.01393	4.595	$<$ **0.001**
Tunable Weight – 3:1 EEG	0.05626	0.01700	3.309	**0.0201**
RF – 3:1 EEG	0.09805	0.01610	6.089	$<$ **0.001**
SVM – 3:1 EEG	0.10440	0.01379	7.573	$<$ **0.001**
LDA – 3:1 EMG	0.05606	0.01933	2.899	0.0698
Mean – 3:1 EMG	-0.02991	0.01571	-1.904	0.5377
Tunable Weight – 3:1 EMG	-0.03766	0.01857	-2.028	0.4527
RF – 3:1 EMG	0.00413	0.01757	0.235	1.0000
SVM – 3:1 EMG	0.01048	0.01604	0.653	0.9980
Mean – LDA	-0.08597	0.01650	-5.210	$<$ **0.001**
Tunable Weight – LDA	-0.09372	0.01937	-4.838	$<$ **0.001**
RF – LDA	-0.05193	0.01816	-2.860	0.0779
SVM – LDA	-0.04558	0.01672	-2.726	0.1107
Tunable Weight – Mean	-0.00775	0.01477	-0.525	0.9995
RF – Mean	0.03404	0.01245	2.734	0.1083
SVM – Mean	0.04038	0.01221	3.307	**0.0199**
RF – Tunable Weight	0.04179	0.01686	2.478	0.1980
SVM – Tunable Weight	0.04813	0.01512	3.184	**0.0300**
SVM – RF	0.00635	0.01465	0.433	0.9999

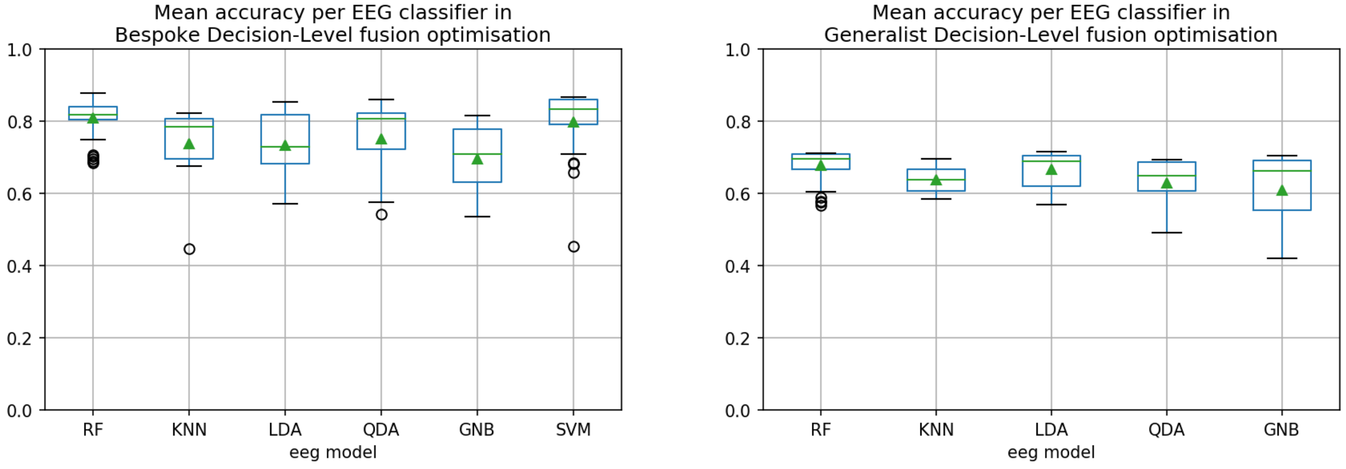
Table A.8: Tukey of all decision fusion algs for Generalist

A.1.3.2 Component EMG & EEG classifiers



(a) Bespoke (b) Generalist

Figure A.15: Mean Development Set accuracies achieved in CASH optimisation of Decision-Level systems grouped by EMG classifier



(a) Bespoke (b) Generalist

Figure A.16: Mean Development Set accuracies achieved in CASH optimisation of Decision-Level systems grouped by EEG classifier

Hypothesis	Estimate	Std. Err	z value	p value
GNB – SVM	-0.08720	0.02393	-3.643	**0.0036**
KNN – SVM	0.02429	0.02546	0.954	0.9291
LDA – SVM	0.01247	0.02637	0.473	0.9969
QDA – SVM	0.01748	0.02506	0.698	0.9813
RF – SVM	-0.02935	0.02627	-1.117	0.8695
KNN – GNB	0.11150	0.03092	3.606	**0.0042**
LDA – GNB	0.09967	0.03181	3.134	**0.0206**
QDA – GNB	0.10469	0.03010	3.478	**0.0065**
RF – GNB	0.05785	0.03213	1.800	0.4555
LDA – KNN	-0.01182	0.03215	-0.368	0.9991
QDA – KNN	-0.00681	0.03127	-0.218	0.9999
RF – KNN	-0.05364	0.03281	-1.635	0.5653
QDA – LDA	0.00501	0.03237	0.155	1.0000
RF – LDA	-0.04182	0.03388	-1.234	0.8137
RF – QDA	-0.04683	0.03222	-1.453	0.6853

(a) Bespoke

Hypothesis	Estimate	Std. Error	z value	p value
GNB – LDA	-0.10299	0.01244	-8.276	<**0.001**
KNN – LDA	-0.03978	0.01228	-3.240	**0.0103**
QDA – LDA	-0.02139	0.01239	-1.727	0.4117
RF – LDA	-0.06069	0.01211	-5.010	<**0.001**
KNN – GNB	0.06321	0.01534	4.121	<**0.001**
QDA – GNB	0.08160	0.01507	5.414	<**0.001**
RF – GNB	0.04229	0.01521	2.781	**0.0422**
QDA – KNN	0.01839	0.01452	1.266	0.7075
RF – KNN	-0.02092	0.01432	-1.461	0.5828
RF – QDA	-0.03930	0.01410	-2.788	**0.0413**

(b) Generalist

Table A.9: Full pairwise comparisons between EMG model choices in optimisation of Decision-Level Fusion systems, tested with Tukey's HSD.

Hypothesis	Estimate	Std. Err	z value	p value
GNB – RF	-0.10988	0.02720	-4.039	<**0.001**
KNN – RF	-0.05795	0.02556	-2.267	0.1989
LDA – RF	-0.05335	0.02915	-1.830	0.4335
QDA – RF	-0.07209	0.02863	-2.518	0.1131
SVM – RF	-0.02412	0.02009	-1.201	0.8293
KNN – GNB	0.05192	0.03274	1.586	0.5958
LDA – GNB	0.05652	0.03569	1.584	0.5970
QDA – GNB	0.03778	0.03613	1.046	0.8973
SVM – GNB	0.08575	0.02911	2.945	**0.0361**
LDA – KNN	0.00460	0.03441	0.134	1.0000
QDA – KNN	0.01414	0.03365	-0.240	0.9982
SVM – KNN	0.03383	0.02738	1.236	0.8114
QDA – LDA	-0.01874	0.03622	-0.517	0.9952
SVM – LDA	0.02923	0.02936	0.996	0.9152
SVM – QDA	0.04797	0.02930	1.637	0.5612

(a) Bespoke

Hypothesis	Estimate	Std. Err	z value	p value
GNB – LDA	-0.07840	0.01428	-5.488	<**0.001**
KNN – LDA	-0.01044	0.01428	-0.743	0.9446
QDA – LDA	-0.03496	0.01405	-2.639	0.0611
RF – LDA	0.01573	0.01325	1.571	0.5077
KNN – GNB	0.06796	0.01757	3.869	<**0.001**
QDA – GNB	0.04344	0.01585	2.741	**0.0464**
RF – GNB	0.09413	0.01469	6.410	<**0.001**
QDA – KNN	-0.02452	0.01594	-1.538	0.5287
RF – KNN	0.02617	0.01368	1.914	0.3024
RF – QDA	0.05069	0.01322	3.834	**0.0011**

(b) Generalist

Table A.10: Full pairwise comparisons between EEG model choices in optimisation of Decision-Level Fusion systems, tested with Tukey's HSD.

On Zero-Phase EEG Filtering

B.1 Confirmation of negligible impact of zero-phase EEG filtering

As discussed in 4.2.5 this work does not use zero-phase filtering to process biosignal data, as it would be unsuitable for the target applications of the research wherein real-time classification is required. For completeness, the impact of applying zero-phase filtering is briefly verified here, to give confidence that the work's findings are not unduly impacted by the Butterworth filter used.

Though efforts are not made to fully characterise the conventional Butterworth filter's group delay here, a cursory investigation of a randomly selected EEG recording was performed by finding the maximum cross-correlation coefficient between raw and Butterworth-filtered signals, i.e. the point at which they are best aligned. This suggested a delay of approximately 58ms — very low a proportion of the 1000ms windows used in feature extraction (see 4.3).

B.1.1 Modelling Performance

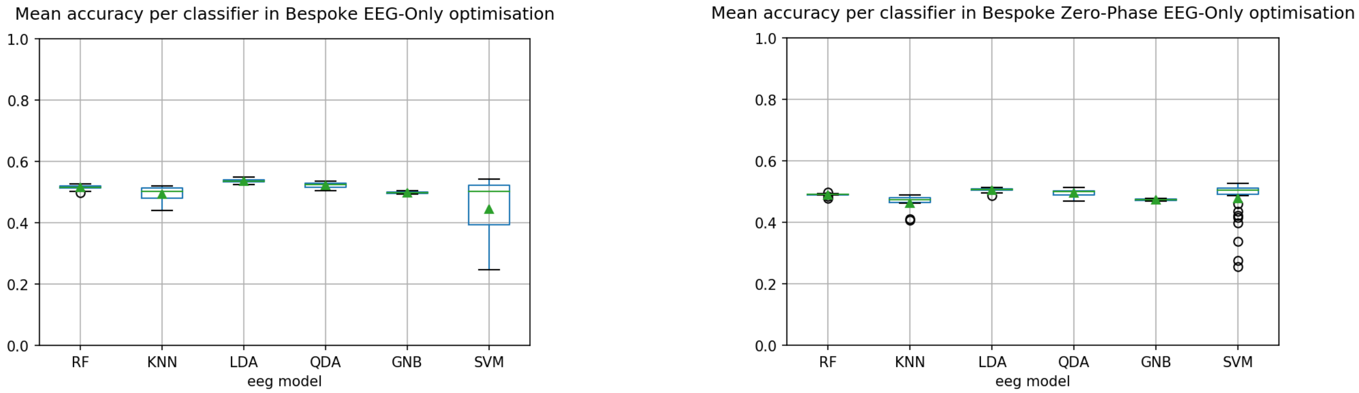
Firstly, any effects of the filtering technique on the modelling capabilities of the resultant EEG data are checked for. CASH optimisation routines for Unimodal EEG classifiers were performed equivalent to those of Chapter 5, using Development Set EEG data but bandpass filtered from 2 – 30 Hz with the zerp-phase MATLAB *filtfilt()* function, rather than the conventional *filter()*.

Table B.1 presents the accuracies achieved and the optimal configurations of Unimodal zero-phase EEG in both Bespoke and Generalist settings, compared against their "conventional" equivalents from Chapter 5 (see Tables 5.5 & 5.6). It can be seen that the difference in peak mean accuracy is low in both cases.

Setting	Conventional		Zero-Phase	
	Accuracy	Configuration	Accuracy	Configuration
Bespoke	54.80	Linear Discriminant Analysis Solver: Least Square Solution Shrinkage: 0.038	52.61	Support Vector Machine C: 1.8767 γ: 0.0281
Generalist	49.11	Linear Discriminant Analysis Solver: Least Squares Solution Shrinkage: 0.435	46.34	Linear Discriminant Analysis Solver: Eigenvalue Decomposition Shrinkage: 0.1265

Table B.1: Peak Development Set accuracy in Unimodal EEG CASH optimisation & corresponding configurations

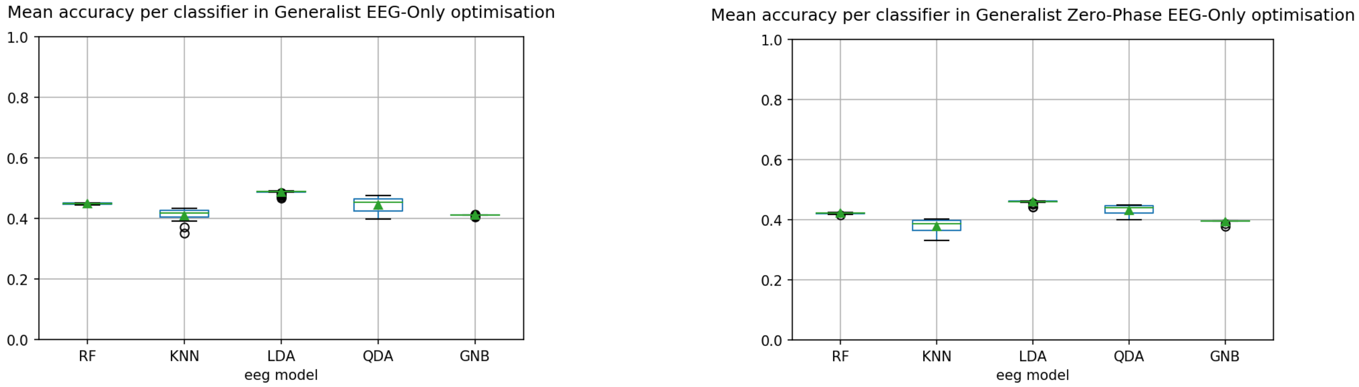
Figures B.1 and B.2 further demonstrate this equivalence in performance. The relative capabilities of the assorted candidate classification algorithms were similar across both the conventional Unimodal EEG experiments and the zero-phase EEG. It is thus considered highly unlikely that the performance of the Unimodal EEG classifiers developed in Chapter 5 was unduly influenced by the choice of filtering method.



(a) Conventional (reproduced from Figure 5.21)

(b) Zero-Phase

Figure B.1: Development Set accuracies achieved by different models in CASH optimisation of Bespoke Unimodal EEG system with conventional (left) & zero-phase (right) Butterworth filtering



(a) Conventional (reproduced from Figure 5.26)

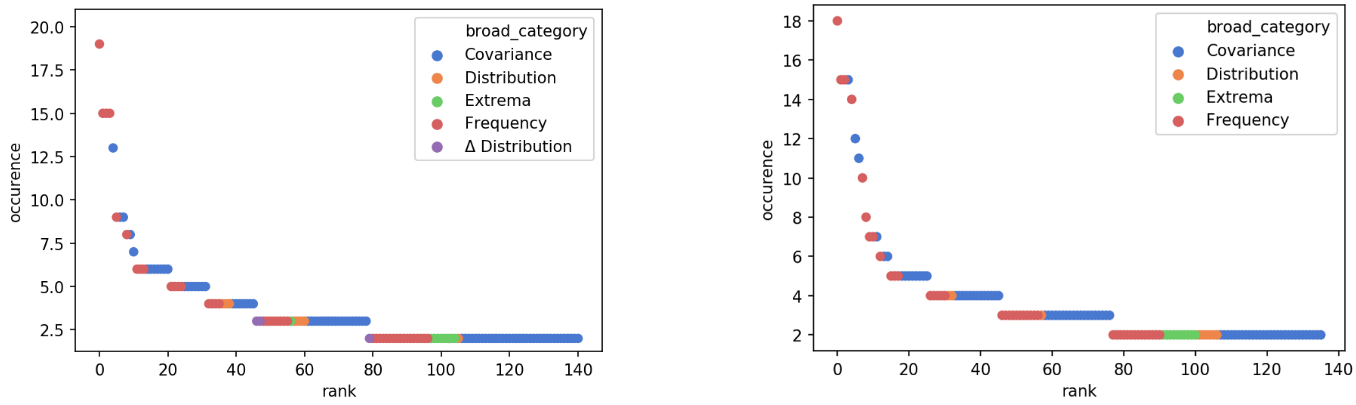
(b) Zero-Phase

Figure B.2: Development Set accuracies achieved by different models in CASH optimisation of Generalist Unimodal EEG system with conventional (left) & zero-phase (right) Butterworth filtering

B.1.2 Feature Informativity

Given that a Butterworth filter's phase delay can differ by frequency, it could be speculated that such delay had different impacts on the information carried in different frequency bands, and thus may have contributed to the findings in Section 5.5.7 regarding EEG feature informativity. This possibility is thus also investigated and discounted here.

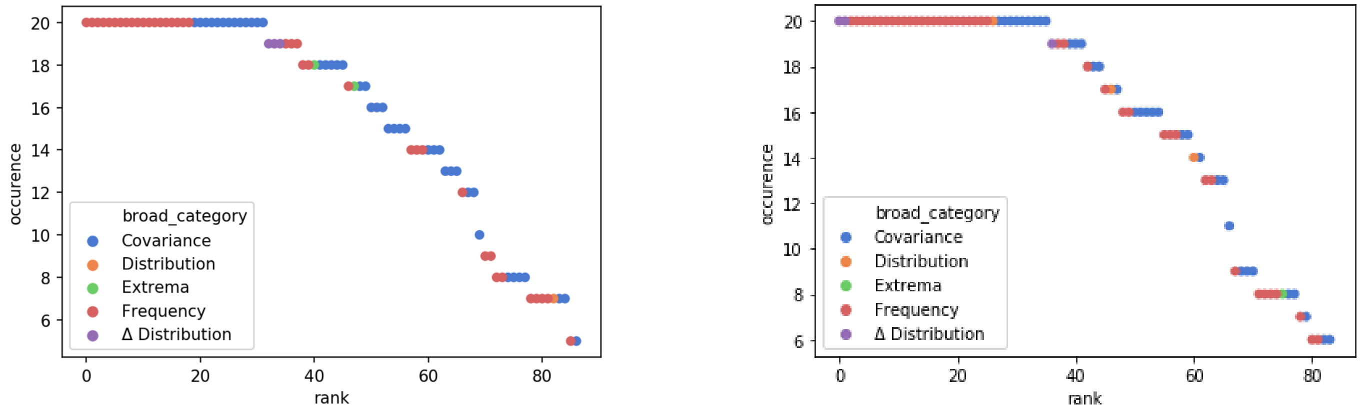
The similarity in sets of features commonly selected in Unimodal EEG systems utilising the different filtering techniques is presented in Figures B.3 & B.4, demonstrating that the same types of features were frequently found informative in both conventionally filtered and zero-phase EEG. Further, Tables B.2 & B.3 indicate that many of the same specific features were identified in both cases, including importantly the Delta frequency bandpower at electrodes 0, 3, 13, and 16 (10-10 sites FC5, C5, FC6, and C6), the significance of which is discussed in 5.5.3. It is thus not likely that the informativity of these features is an artificial by-product of the conventional Butterworth filter's phase delay.



(a) Conventional (reproduced from Figure 5.12)

(b) Zero-Phase

Figure B.3: Occurrence of EEG features selected in multiple subjects' Bespoke systems, grouped by feature category



(a) Conventional (reproduced from Figure 5.15)

(b) Zero-Phase

Figure B.4: Occurrence of EEG features selected in Generalist systems of at least 5 subjects, grouped by category

Feature	Occurrence Rate
sum delta 13	19
sum delta 16	15
sum delta 3	15
sum delta 0	15
log covM 0-13	13
sum delta 7	9
log covM 4-5	9
log covM 13-16	9
sum beta 5	8
log covM 5-5	8
log covM 0-3	7

(a) Conventional (data presented previously in Fig. 5.14)

Feature	Occurrence Rate
sum delta 13	18
sum delta 16	15
sum delta 0	15
log covM 0-13	15
sum delta 3	14
log covM 4-5	12
log covM 5-5	11
sum delta 7	10
sum theta 13	8
sum beta 5	7
sum beta 4	7
log covM 7-7	7

(b) Zero-Phase

Table B.2: EEG features selected by ≥ 7 subjects' Bespoke Unimodal EEG systems with conventional (left) & zero-phase (right) Butterworth filtering. Features uniquely selected in one processing technique or another in **bold**.

Feature			
logcovM 4-5	**logcovM 3-16**	logcovM 3-8	covM 2-2
logcovM 0-3	logcovM 5-5	logcovM 19-19	**logcovM 7-19**
covM 4-13	**logcovM 13-16**	logcovM 2-2	**logcovM 12-13**
logcovM 0-13	sum delta 0	**sum alpha 15**	sum alpha 13
sum alpha 10	sum delta 5	sum delta 16	sum beta 6
sum delta 13	sum delta 3	sum theta 3	sum beta 4
sum alpha 4	sum beta 9	sum beta 5	sum theta 16
sum delta 1	sum beta 11	sum beta 8	sum alpha 7

(a) Conventional (reproduced from Table 5.15)

Feature			
logcovM 5-5	logcovM 3-8	logcovM 0-13	**logcovM 6-10**
logcovM 4-5	logcovM 19-19	logcovM 2-2	covM 2-2
logcovM 0-12	**kurtosis 13**	sum beta 8	**sum alpha 18**
sum alpha 13	sum theta 16	sum delta 16	**sum delta 11**
sum alpha 4	**sum gamma 4**	sum beta 9	**sum delta 10**
sum beta 6	sum delta 0	sum alpha 10	sum beta 4
sum beta 5	**sum alpha 16**	sum alpha 7	sum delta 3
sum delta 1	sum beta 11	**sum beta 3**	sum delta 13
sum theta 3	sum delta 5	∆ std dev 13	∆ std dev 16

(b) Zero-Phase

Table B.3: EEG features selected in all 20 subjects' Generalist Unimodal EEG systems with conventional (above) and zero-phase (below) Butterworth filtering. Features uniquely selected in one processing technique or another in **bold**.

On Synthetic Augmentation: Selected extracts from "Synthetic Biological Signals Machine - Generated by GPT-2 Improve the Classification of EEG and EMG Through Data Augmentation" [1] (Bird, Pritchard, et al. in IEEE Robotics and Automation Letters ©2021 IEEE)

This work was a collaboration between myself and Dr. Jordan J Bird (as co-first authors), and Prof. Aniko Ekárt & Drs. Antonio Fratini and Diego Faria who provided supervision and gave feedback on the manuscript prior to corrections and submission to IEEE Robotics and Automation Letters.

Dr. Bird's PhD research interests focused more significantly on Electroencephalography than Electromyography and the focus of his contribution to the paper was in this area. The EEG experiments were included in his PhD thesis [376] and thus do not appear among the excerpts of the paper presented below.

Select passages, marked in bold for convenience, discuss aspects of the work which applied to both the EEG and EMG experiments; the reader is advised that these hence appear in both Dr. Bird's thesis and in the below extract.

Excerpts from: *Synthetic Biological Signals Machine-generated by GPT-2 improve the Classification of EEG and EMG through Data Augmentation*

Jordan J. Bird[1], Michael Pritchard[1], Antonio Fratini[2], Anikó Ekárt[3], Diego R. Faria[1*]

1 Introduction

When presenting their Generative Pretrained Transformer (GPT) model, researchers at OpenAI hypothesised that *language models are unsupervised multitask learners* [1]. At the current state-of-the-art this claim has been consistently argued through applications such as fake news identification [2], patent claims [3], and stock market analysis [4] to name just a few in a rapidly growing area of research. In this work, we follow those before us in exploring the capabilities of these models in a brand new field of application: the generation of bio-synthetic signals (in our case Electroencephalographic (EEG) and Electromyographic (EMG) activity). In detail, we aimed at exploring whether or not GPT-2's self-attention based architecture was capable of creating synthetic signals, and if those signals could improve the performance of classification models used on real datasets. Enabling better results for the deduction of a physical action or mental thought allows for a higher degree of certainty when it comes to an unseen subject. That is, for example in electromyographically controlled robotic prosthetic limbs, a more improved experience for the user of such a robotic device. Our scientific contributions and results suggest that:

1. It is possible to generate synthetic biological signals by tuning a language transformation model.
2. Classifiers trained on either real or synthetic data can classify one another with relatively high accuracy.
3. Synthetic data improves the classification of the real data both in terms of model benchmarking and classification of unseen samples.

2 Related Work and Background

In this section, we describe how previous work has demonstrated the benefits of augmenting biological signal datasets to improve classification results, since it has been noted that augmentation is a useful technique to overcome data scarcity in such domains [5]. A common approach is to generate synthetic signals by re-arranging components of real data. Lotte [6] proposed a method of *"Artificial Trial Generation Based on Analogy"* where three data examples x_1, x_2, x_3 provide examples and an artificial $x_{synthetic}$ is formed which is to x_3 what x_2 is to x_1. A transformation is applied to x_1 to make it more similar to x_2, the same transformation is then applied to x_3 which generates $x_{synthetic}$[1]. This approach was shown to improve performance of a Linear Discriminant Analysis classifier on three different datasets. Dai et al. [7] performed similar rearrangements of waveform components in both the time and frequency domains to add three times the amount of initially collected EEG data, finding that this approach could improve the classification accuracy of a Hybrid Scale Convolutional Neural Network. This work showed that data augmentation allowed the model to improve

[*] J.J. Bird and M. Pritchard are co-first authors

[1] Equations for Lotte's EEG generation technique can be found in [6]

the classification of data for individual subjects that were specifically challenging in terms of the model's classification ability. Dinarès-Ferran [8] decomposed EEG signals into Intrinsic Mode Functions and constructed synthetic data frames by arranging these IMFs into new combinations, demonstrating improvements of classification performance of motor imagery based BCIs while including these new signals. Other researchers have proposed data augmentation techniques commonly used in other domains such as image classification techniques with positive results. As an example Shovon et al. [9] applied conventional image augmentation techniques e.g. rotation, zoom, and brightness to spectral images formed from EEG analysis to increase the size of a public EEG dataset. This ultimately led to an improvement over the state-of-the-art.

Current research shows great impact can be derived from relatively simple techniques. For example, Freer [10] observed that introducing noise into gathered data to form additional data points improved the learning ability of several models which otherwise performed relatively poorly. Tsinganos et al. [11] studied the approaches of magnitude warping, wavelet decomposition, and synthetic surface EMG models (generative approaches) for hand gesture recognition, finding classification performance increases of up to +16% when augmented data was introduced during training. More recently, data augmentation studies have begun to focus on the field of deep learning, more specifically on the ability of generative models to create artificial data which is then introduced during the classification model training process. In 2018, Luo et al. [12] observed that useful EEG signal data could be generated by Conditional Wasserstein Generative Adversarial Networks (GANs) which was then introduced to the training set in a classical train-test learning framework. The authors found classification performance was improved when such techniques were introduced. Likewise, Zhang and Liu [13] applied similar Deep Convolutional GANs (DC-GAN) to EEG signals given that training examples are often scarce in related works. As with the previous work, the authors found success when augmenting training data with DC-GAN generated data. Zanini and Colombini [14] provided a state-of-the-art solution in the field of EMG studies when using a DC-GAN to successfully perform *style transfer* of Parkinson's Disease to bio-electrical signals, noting the scarcity of Parkinson's Disease EMG data available to researchers as an open issue in the field [14]. Many studies observed follow a relatively simple train/test approach to benchmarking models.

A limitation of many techniques is that they are not temporal in their generative natures. Each block of signal output has no influence on the next, and, as such, a continuous synthetic signal of unlimited length cannot therefore be generated. Our approach allows for infinite generation of temporal wave data given the nature of GPT-2; a continuous synthetic raw signal is generated by presenting some of the previous outputs as input for the next generation. We then benchmark the models through k-fold cross validation, where each fold has synthetic data introduced as additional training data. Moreover, for the first time in the field, we show the effectiveness of attention-based models at the signal level rather than generative based models at the feature-level for both training and unseen data. We then finally show that real-time gesture classification towards direct control of a robotic arm is improved following our data augmentation framework.

2.1 GPT-2 and Self-Attention Transformers

Self-Attention Transformers are based on calculating scaled dot-product attention units, and generate new data by learning to paying attention to previous data generated [15]. Scaled dot-product attention is calculated for each unit within the input vector, e.g. words in a sentence, or, in this case, signals in a stream. The attention units are input with a sequence and output embeddings of relevant tokens. Query (W_q), key (W_k), and value (W_v) weights are calculated as:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{1}$$

where the query is an entity within the sequence, keys are vector representations of the input, and the values are derived by querying against keys. The term self-attention comes from the fact that Q, K and V are received from the same source, and generation is an unsupervised. GPT-2 architecture follows the concept of Multi-headed Attention:

$$\begin{gathered} MultiHead(Q,K,V) = Concat(head_1, ..., head_h)W^O \\ head_i = Attention(QW_i^Q, KW_i^K, VW_i^V). \end{gathered} \tag{2}$$

That is, a deep structure of h_i attention heads in order to inter-connect multiple attention units. Fundamentally, the GPT and GPT-2 algorithms do not differ. The main advantages of GPT-2 are based on it being many times more complex than the GPT with 1.5 billion parameters and being trained on a large dataset of 8 million websites.

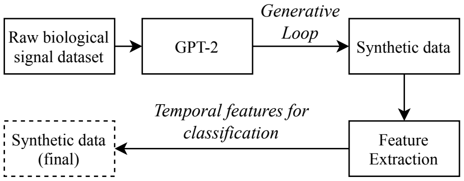


Figure 1: Initial training of the GPT-2 model and then generating a dataset of synthetic biological signals.

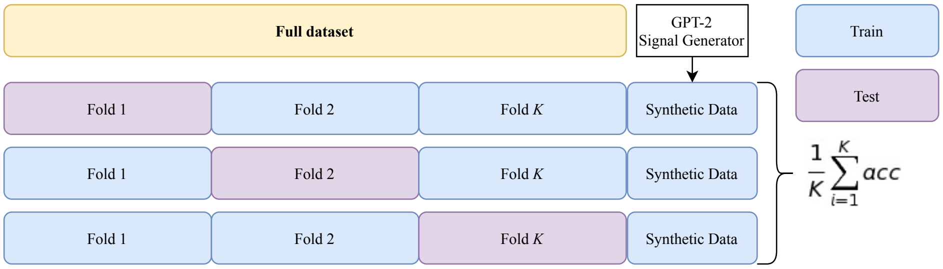


Figure 2: The standard K-Fold cross validation process with the GPT-2 generated synthetic data being introduced as additional training data for each fold.

3 Method

3.1 Data Collection, Pre-processing and Feature Extraction

The EMG dataset used in this study was initially acquired by Dolopikos et al. in [16]. EMG data corresponding to the opening and closing movements of the right hand were collected from fifteen able-bodied participants (9 male, 6 female, mean age 26) using a Thalmic Labs Myo armband. The participants performed the gestures after a cue from an instructor. The recorded data corresponding to the time before the onset of physical activity (muscular background tone) was extracted and compiled into a third "neutral" class. To assess contraction and relaxation of muscles, information can be extracted by the simple analysis of an EMG signal's smoothed rectified envelope [17]. The data was indeed first rectified and then low-pass filtered using a peak detection algorithm [18], interpolating between local maxima with a separation of at least 20 samples (equivalent to 0.1 seconds at the Myo's natural sample rate of 200Hz).

Whilst the data was provided to GPT-2 in its raw format, an ensemble of features was extracted from each dataset to enable classification. The feature set has previously proven effective, providing sufficient information to discriminate both between focused, relaxed, and neutral brains [19], and closed, open, and neutral hands [16]. Features are extracted from a sliding window of 1 second in length, at an overlap of 0.5 seconds. These windows are further sub-divided into halves and quarters, enabling extraction of the following ensemble of statistical features[2].

3.2 Generating and Learning from GPT-2 Generated Signals

GPT-2 models are initially trained on each class of data for 1,000 steps each. Then, for n classes, n GPT-2s are tasked with generating synthetic data and the class label is finally manually added to the generated data. This process can be observed in Figure 1 where the generative loop is prefixed by the latter half of the previously generated data[3]. The synthetic equivalent of 60 seconds of data per class are generated (30,000 rows per class of raw signal data).

To benchmark machine learning models, a K-fold cross validated learning process is followed and compared to the process observed in Figure 2 where training data is augmented by the synthetically-derived data at each fold of learning. The testing set does not contain any of the artificial signal data. **This**

[2] Feature extraction code available at`https://github.com/jordan-bird/eeg-feature-generation`

[3] Example code can be found at: `https://github.com/jordan-bird/Generational-Loop-GPT2`

process is performed for both the EEG and EMG experiments for six different models: Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbours (KNN, $K = 10$), Linear Discriminant Analysis (LDA), Logistic Regression (LR), and Gaussian Naïve Bayes (GNB). These statistical models are selected due to their differing nature, to explore the hypothesis with a mixed range of approaches. As was explored in [20], it was found that unseen signal classification can be improved through calibration via inductive and supervised transductive transfer learning. That is, tuning a model by providing a small amount of calibration data to the training set.

4 Observations and Results

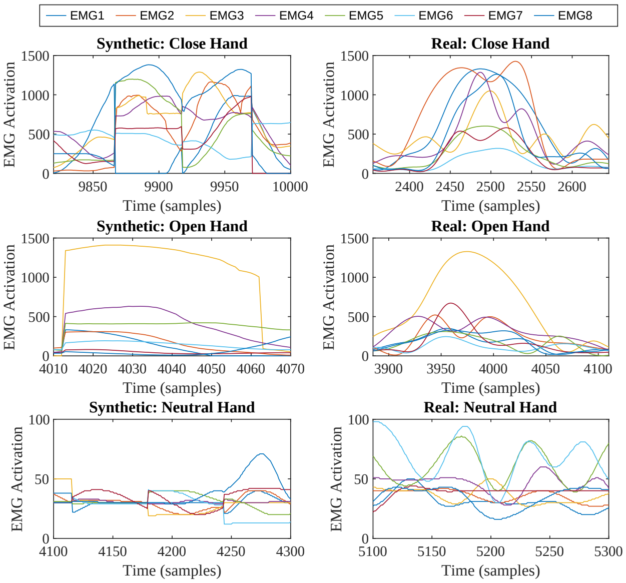


Figure 3: Comparison of GPT-2 generated (Left) and genuine recorded (Right) EMG data across "Closed", "Open", and "Neutral" hand classes.

Figure 3 compares samples of real and synthetic EMG data. **It was noted that all synthetic data was unique compared to the real data. Interestingly, natural behaviours, such as the presence of characteristic oscillations, can be observed within data, showing that complex natural patterns have been generalised by the GPT-2 model.** The GPT-2 generated waves are seemingly less natural than their human counterparts; although natural wave patterns do emerge, they are more erratic and prone to spiking unlike the signals recorded from a human forearm. The Power Spectral Densities, computed with Welch's method [21], of the GPT-2 generated and real data are presented in Figure 4. Across all classes the synthetic data has significantly more power in its high frequency components than the real data, despite the real EMG dataset having been low-pass filtered before being used to train GPT-2; this phenomenon is likely due in part to the aforementioned erratic nature of the synthetic EMG signals.

4.1 Classification of real-to-synthetic data and vice-versa

Table 1 shows the ability to classify real EMG data by learning from synthetic data and vice versa. The Naïve Bayesian model when trained on only real data can classify the synthetic data with 62.36% accuracy, whereas the K- Nearest Neighbours model can classify the real dataset with 78.24% accuracy when trained on only synthetic.

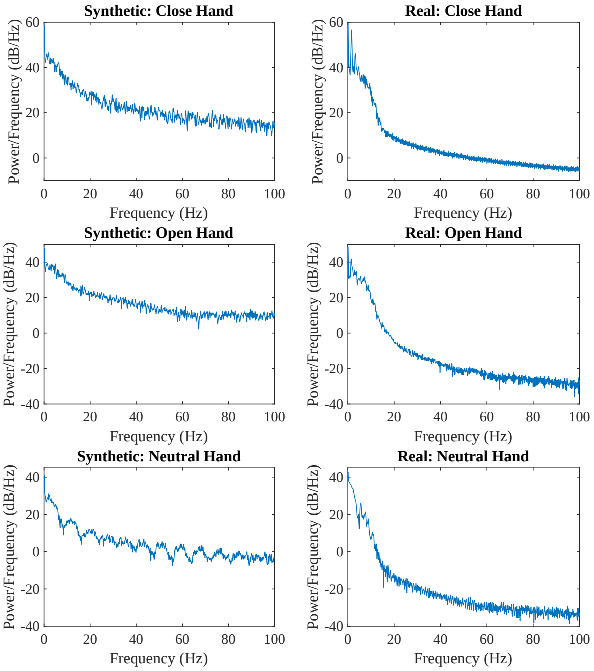


Figure 4: Comparison of Power Spectral Densities of GPT-2 generated (Left) and genuine recorded (Right) EMG data. For readability, only the PSD computed from electrode EMG1 is shown.

Table 1: Classification results when training on real or synthetic EMG data and attempting to predict the class labels of the other (sorted for real to synthetic).

Classifier	**Training and Prediction Data**	
	Real to Synthetic	***Synthetic to Real***
Gaussian Naïve Bayes	**62.36**	64.39
10 Nearest Neighbours	62.07	**78.24**
Random Forest	61.78	71.23
Linear Discriminant Analysis	50.00	60.69
Logistic Regression	37.36	71.71
Support Vector Machine	35.63	71.27

4.2 EMG Classification

Table 2: Comparison of the 10-fold classification of EMG data and 10-fold classification of EMG data alongside synthetic data as additional training data.

Classifier	Without GPT-2	With GPT-2 Data
Random Forest	93.62 (0.8)	**93.90 (0.59)**
Logistic Regression	93.75 (1.04)	93.86 (1.05)
Support Vector Machine	93.42 (0.89)	93.46 (0.94)
Linear Discriminant Analysis	91.95 (1)	92.59 (1.05)
10 Nearest Neighbours	91.23 (0.89)	91.11 (0.88)
Gaussian Naïve Bayes	77.73 (1.39)	74.46 (1.38)

Table 2 shows the results for EMG classification. The best model was the Random Forest which scored 93.9% (deviance 0.59) during the k-fold benchmarking process in which GPT-2 synthetic data was introduced as additional training data.

Table 3: EMG classification abilities of the models on completely unseen data with regards to both with and without synthetic GPT-2 data as well as prior calibration.

Classifier	Uncalibrated		Calibrated	
	Vanilla	***Synth.***	***Vanilla***	***Synth.***
Random Forest	67.33	69.31	74.26	75.25
Logistic Regression	60.40	87.13	60.40	87.13
Support Vector Machine	39.60	62.38	44.55	46.53
Linear Discriminant Analysis	65.35	67.33	86.14	79.21
10 Nearest Neighbours	75.25	75.25	78.22	78.22
Gaussian Naïve Bayes	95.05	94.06	96.04	**97.03**

Table 3 shows the abilities of the models when predicting the class label of completely unseen EMG data. Interestingly, the Gaussian Naïve Bayes model outperformed all others consistently. The best Gaussian Naïve Bayes model at predicting completely unseen data was when it was also trained with calibration and GPT-2 synthetic data alongside the dataset, at an accuracy of 97.03%.

4.3 Real-time EMG Prediction

The results in Figure 5 show the process of a user performing hand gestures for three minutes (124 data objects). The best-performing EMG prediction model was applied (Gaussian Naïve Bayes + GPT-2), which predicted real-time data with 89.5% accuracy. All of the erroneous predictions occurred during state transitions, which was expected given that models were trained on concrete gestures and had not been exposed to transitional behaviours of the arm muscles when shifting between gestures. The best predictive model on the dataset without GPT-2 augmentation scored 68.29% accuracy. The 95% Wilson confidence interval for the augmented model's accuracy was [82.89, 93.77], and for the non-augmentation model was [59.62,75.86]. No calibration was performed, that is, the models were never exposed to data from this user. Thus, GPT-2 biosignal data augmentation leads to a model which can classify data from unseen subjects with a higher rate of success. Figure 6 shows the confusion matrix for this experiment.

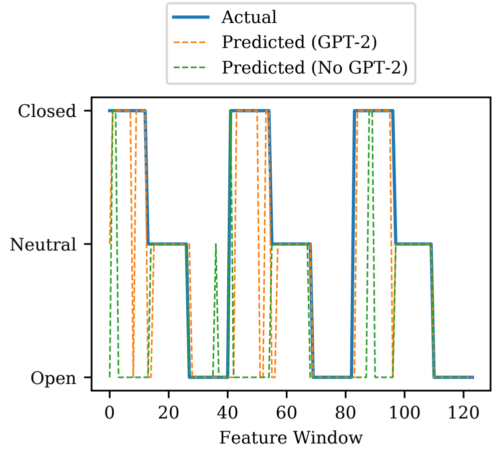


Figure 5: Real-time execution of gestures for three minutes predicted with the augmented EMG model (89.5%) and non-augmented EMG model (68.29%).

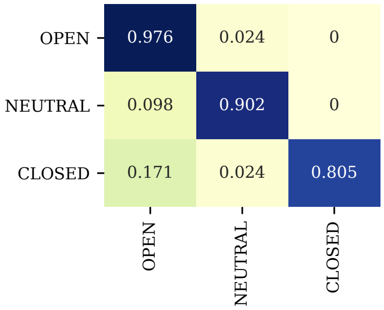


Figure 6: Confusion Matrix for real-time EMG classification

5 Conclusions and Future Work

To conclude, this study has presented multiple experiments with real and synthetic biological signals in order to ascertain whether classification algorithms can be improved by considering data generated by the GPT-2 model. Although the data are different, i.e., real and synthetic data were unique, a model trained on one of the two sets of signals can strongly classify the other and thus the GPT-2 model is able to generate relatively realistic data which holds useful information that can be learnt from for application to real signals. For EEG, an SVM trained on synthetic data could classify real data at 74.71% accuracy and a KNN algorithm could do the same for real EMG classification at 78.24% accuracy, training on only synthetic data. We then showed that several learning algorithms were improved for both EMG and EEG classification when the training data was augmented by GPT-2. The main argument of this work is that synthetic biosignals generated by an attention-based transformer hold useful information towards improving several learning algorithms for classification of real biological signal data. In future, larger datasets could be used and thus deep learning would be a realistic possibility for classification following the same process. Given that this work showed promise in terms of the model architecture itself, similar models could also be benchmarked in terms of their ability to create augmented training datasets e.g. BART, CTRL, Transformer-XL and XLNet. Another unoptimised level of detail is the amount of synthetic data that is added to the training set for augmentation, future work could explore the level of data needed for apt improvements to the models.

Our suggested model for EMG, the GNB approach trained with human-sourced GPT-2 generated synthetic signals, was powerful in terms of predictive ability and required relatively little computational resources given its simplistic nature. Additionally, the approach did not require further calibration, as many state-of-the-art approaches do (including the Myo software itself), instead correctly predicting the behaviours of a new subject from the point of wearing the device. Given these attributes, the model is apt for usage on-board within wearable EMG devices for real-time prediction of gesture.

References

[1] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” *OpenAI Blog*, vol. 1, no. 8, pp. 1–9, 2019.

[2] J. C. B. Cruz, J. A. Tan, and C. Cheng, “Localization of fake news detection via multitask transfer learning,” in *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 2596–2604, 2020.

[3] J.-S. Lee and J. Hsiang, “Patent claim generation by fine-tuning openai gpt-2,” *World Patent Information*, vol. 62, p. 101983, 2020.

[4] Y. Nishi, A. Suge, and H. Takahashi, “News articles evaluation analysis in automotive industry using gpt-2 and co-occurrence network,” *New Frontiers in Artificial Intelligence*, pp. 103–114, 2020.

[5] E. Lashgari, D. Liang, and U. Maoz, “Data augmentation for deep-learning-based electroencephalography,” *Journal of Neuroscience Methods*, p. 108885, 2020.

[6] F. Lotte, “Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain–computer interfaces,” *Proceedings of the IEEE*, vol. 103, no. 6, pp. 871–890, 2015.

[7] G. Dai, J. Zhou, J. Huang, and N. Wang, “HS-CNN: a CNN with hybrid convolution scale for EEG motor imagery classification,” *Journal of Neural Engineering*, vol. 17, jan 2020.

[8] J. Dinarès-Ferran, R. Ortner, C. Guger, and J. Solé-Casals, “A new method to generate artificial frames using the empirical mode decomposition for an eeg-based motor imagery bci,” *Frontiers in Neuroscience*, vol. 12, pp. 1–308, 2018.

[9] T. H. Shovon, Z. A. Nazi, S. Dash, and M. F. Hossain, “Classification of motor imagery eeg signals with multi-input convolutional neural network by augmenting stft,” in *2019 5th International Conference on Advances in Electrical Engineering (ICAEE)*, pp. 398–403, 2019.

[10] D. Freer and G.-Z. Yang, “Data augmentation for self-paced motor imagery classification with c-LSTM,” *Journal of Neural Engineering*, vol. 17, jan 2020.

[11] P. Tsinganos, B. Cornelis, J. Cornelis, B. Jansen, and A. Skodras, “Data augmentation of surface electromyography for hand gesture recognition,” *Sensors*, vol. 20, no. 17, p. 4892, 2020.

[12] Y. Luo and B.-L. Lu, “Eeg data augmentation for emotion recognition using a conditional wasserstein gan,” in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2535–2538, IEEE, 2018.

[13] Q. Zhang and Y. Liu, “Improving brain computer interface performance by data augmentation with conditional deep convolutional generative adversarial networks,” *arXiv preprint arXiv:1806.07108*, 2018.

[14] R. Anicet Zanini and E. Luna Colombini, “Parkinson’s disease emg data augmentation and simulation with dcgans and style transfer,” *Sensors*, vol. 20, no. 9, p. 2605, 2020.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[16] C. Dolopikos, M. Pritchard, J. J. Bird, and D. R. Faria, “Electromyography signal-based gesture recognition for human-machine interaction in real-time through model calibration,” in *SAI Future of Information and Communication Conference (FICC) 2021*, pp. 1–18, 2021.

[17] C. J. D. Luca, “The use of surface electromyography in biomechanics,” *Journal of Applied Biomechanics*, vol. 13, no. 2, pp. 135–163, 1997.

[18] The MathWorks, Inc., *Signal Processing Toolbox*, pp. 489–503. The MathWorks, Natick, MA, USA, 2020.

[19] J. J. Bird, L. J. Manso, E. P. Ribeiro, A. Ekart, and D. R. Faria, “A study on mental state classification using eeg-based brain-machine interface,” in *2018 International Conference on Intelligent Systems (IS)*, pp. 795–800, IEEE, 2018.

[20] J. Kobylarz, J. J. Bird, D. R. Faria, E. P. Ribeiro, and A. Ekárt, “Thumbs up, thumbs down: non-verbal human-robot interaction through real-time emg classification via inductive and supervised transductive transfer learning,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 12, p. 6021–6031, 2020.

[21] P. Welch, “The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms,” *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.