An Investigation of Multimodal EMG-EEG Fusion Strategies for Upper-Limb Gesture Classification

Michael Pritchard[1*], Felipe Campelo[2], and Harry Goldingay[3]

[1]Department of Applied AI and Robotics, Aston University, Birmingham, B4 7ET UK *(Corresponding Author)*
[2]School of Engineering Mathematics and Technology, University of Bristol, Bristol, BS8 1QU UK
[3]Aston Centre for Artificial Intelligence Research and Application, Aston University, Birmingham, B4 7ET UK

E-mail: m.pritchard2@aston.ac.uk, f.campelo@bristol.ac.uk, h.j.goldingay1@aston.ac.uk

May 2025

Abstract. Objective: 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 methodological rigour, the extent to which generalisation is demonstrated, and the granularity of gestures classified. This work evaluates three architectures for multimodal fusion of EMG & EEG data in gesture classification, including a novel Hierarchical strategy, in both subject-specific and subject-independent settings.
Approach: We propose an unbiased methodology for designing classifiers centred on Automated Machine Learning through Combined Algorithm Selection & Hyperparameter Optimisation (CASH); the first application of this technique to the biosignal domain. Using CASH, we introduce an end-to-end pipeline for data handling, algorithm development, modelling, and fair comparison, addressing established weaknesses among biosignal literature.
Main results: EMG-EEG fusion is shown to provide significantly higher subject-independent accuracy in same-hand multi-gesture classification than an equivalent EMG classifier. Our CASH-based design methodology produces a more accurate subject-specific classifier design than recommended by literature. Our novel Hierarchical ensemble of classical models outperforms a domain-standard CNN architecture. We achieve a subject-independent EEG multiclass accuracy competitive with many subject-specific approaches used for similar, or more easily separable, problems.
Significance: To our knowledge, this is the first work to establish a systematic framework for automatic, unbiased designing and testing of fusion architectures in the context of multimodal biosignal classification. We demonstrate a robust end-to-end modelling pipeline for biosignal classification problems which if adopted in future research can help address the risk of bias common in multimodal BCI studies, enabling more reliable and rigorous comparison of proposed classifiers than is usual in the domain. We apply the approach to a more complex task than typical of EMG-EEG fusion research, surpassing literature-recommended designs and verifying the efficacy of a novel Hierarchical fusion architecture.

Keywords: Biosignal Fusion, Multimodal Gesture Classification, Brain-Computer-Interface, Automated Machine Learning

1. Introduction

Much of society and the built environment are designed for an imagined "typical" individual and inaccessible to those falling outside this narrow definition. At least 1 in 190 of the U.S. population [1], a proportion believed similar worldwide, are believed to have experienced the loss of a limb. For many upper-limb amputees, accessibility support comes in the form of prostheses. While Electromyography (EMG)-controlled prosthetic arms are now widely available – in the United Kingdom, for example, the National Health Service offers such prostheses as standard [2] – many prostheses are limited in the control they allow. For instance, the OpenBionics Hero Arm, which is generally recognised as state-of-the-art among affordable, accessible, commercially available robotic prostheses, operates through Direct Control. Users can select from a set of pre-defined gestures, and the amplitude of their measured residual limb EMG activity acts as the control signal for actuation of the selected gesture [3].

Pattern Recognition systems – those employing machine learning algorithms to classify between defined gestures from measured biological data – have been the focus of extensive research but rarely seen deployment in the "real world". This is despite much research indicating their suitability [4, 5]. One barrier to their adoption is the need for data upon which to train the underlying models. The majority of research considers subject-specific classification, training "Bespoke" models using only a given individual's data. Less studied are subject-independent paradigms, wherein a "Generalist" model has no access to the target user's data before prediction.

In either setting, existing studies suffer from several problems. Firstly, the majority work solely with EMG data, whereas amputees may not have the same amount or type of EMG data available. One potential solution is to draw on Electroencephalography (EEG) alongside EMG in a multimodal system but, while this has been tried in literature, there is no established technique for fusing these data and fair like-for-like comparisons between proposed fusion methods, and against non-fusion "unimodal" approaches, are often lacking, leading to questions over the extent of the merit of such strategies. Studies (both unimodal and multimodal) also often make assumptions about modelling choices based on literature precedent, which may not have strong empirical justification. Frequently, studies give insufficient detail over their design processes to rule out the potential of biased comparisons [6]. Further, many studies fail to clearly articulate their data handling strategy. It is easy to mishandle biosignal data, leading to data leakage and invalid results [7,8]. Although this has been discussed in the literature [9], there is still no standard strategy in the domain. Finally, datasets used in biosignal studies, and especially in works on EMG-EEG fusion, are often simple, based on gestures which are easily differentiable, e.g. hand vs foot movement [10]. This is not representative of potential real-world applications [11].

There is clearly a need for a more thorough evaluation of biosignal fusion techniques for upper-limb gesture classification. To do so requires establishing a robust methodology for unbiased determination of classifier design, enabling fair comparison between competing strategies, and the application of a data handling protocol which minimises the risk of data leakage undermining validity.

In this paper we address the issues described above through several main contributions, which together present a more reliable process for developing and evaluating biosignal-based classifiers, including those based on multimodal biosignal fusion, than is typical of the domain. We provide an extensive evaluation of EMG–EEG fusion strategies, for both bespoke and generalist settings, which includes the development and validation of new architectures for fusing multimodal data which can outperform unimodal EMG. To enable this evaluation, we propose an unbiased methodology for designing and comparing classification approaches which is novel to the domain. We use this approach to confirm the efficacy of some literature-precedented approaches to the problem, and show that literature inferences do not provide performant designs in certain important cases. Through the application of this methodology, we also demonstrate an end-to-end data handling strategy that addresses issues of bias and leakage identified in the literature. Notably, we present these in the context of a dataset more complex than those used in most other studies in the field of EMG-EEG fusion, having multiple participants and gestures which are not trivially distinguishable.

2. Related Work

There is no clear consensus in the literature about which machine learning algorithms are best suited for classifying EMG and EEG data. This choice can be highly dependent on characteristics of the dataset and of the features used, the ultimate predictive goal, and the preprocessing steps deployed. Despite this, there are some trends in the published literature. For EMG data, SVMs are one of the leading choices [12–15] but other algorithms such as LDA [16, 17], Random Forests [18, 19], Naïve Bayes [20, 21], k-NN [22, 23] and QDA [24] are also commonly used. For EEG data, the use of LDA-based classifiers is a long-standing

preference [6], but many of the algorithms listed above are also commonly applied, notably RFs [25] and SVMs [26–28]. Studies on joint EMG-EEG classification also use a wide range of ML algorithms including both LDA and QDA [29, 30], RF [18, 19], and kNN [18, 30].

Not only does algorithm choice vary much among biosignal research, but their configuration does also. Among SVMs the Radial Basis Function (RBF) kernel is the *de facto* standard [31], although some studies also use a linear kernel [12, 13]. There is a large heterogeneity of approaches used to adjust the SVM regularisation and scale parameters, including manual tuning, automated optimisation, and derivation from dataset properties, with a consequent diversity of results [12, 14, 26, 27]. Similarly varied hyperparameter determination approaches and results are reported in works deploying kNN classifiers [12, 22, 23, 32] or Random Forests [25, 32], and works using LDAs or QDAs frequently omit any such details [16, 30, 33, 34]. It is even not uncommon for key implementation details such as hyperparameter values to go unspecified [13, 18, 19, 35].

Deep learning has also seen successful use in biosignal research [36]. However, literature does not indicate that deep learning models consistently outperform alternative, lower-cost options. In reviewing EEG classification, Lotte et al. [11] note that while popular, Deep Learning architectures do not appear to present notable benefit over alternatives, and can be hindered by excessively long training times. For EMG, Phinyomark et al. [37] highlight their need for very large datasets to avoid overfit, which may not be always available. Dolopikos et al. [38] corroborate this, finding a voting ensemble of classical models, and indeed individual RFs and SVMs, to outperform a Deep Neural Network. Shallower architectures have similarly been found not to notably outperform SVMs and LDA models in both EMG and EEG [17, 39], and to face speed issues [12] and a sensitivity to electrode repositioning [40].

With no clear consensus as to the most suitable modelling options, researchers and developers must make choices. This however introduces a number of challenges. 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 and datasets [41], and over-reliance on literature trends risks underexploration of potential performant approaches. Transparency and the avoidance of bias in designing and comparing candidate classification systems has been noted as lacking among biosignal literature [6, 11]. Such "*biased comparisons, with* [...] *unjustified choices of parameters* [which prevent] *us from ruling out manual tuning of these parameters with knowledge of the test set*" [11] highlight the need for fair, unbiased explorations of candidate classification approaches.

The approaches by which multimodal data are used in a Hybrid Brain-Computer Interface also vary significantly among published works [42]. They are rarely directly compared against one another or with unimodal approaches, further highlighting the need for more thorough approaches to the design and assessment of biosignal fusion systems. In reviewing the literature, we identified four major approaches that are commonly used for biosignal fusion: sequential contribution, decoupled contribution, feature-level fusion and decision-level fusion.

Signals are sometimes utilised on a **sequential** basis – a "gated" paradigm wherein some data modalities contribute only when certain characteristics have been observed in others. Commonly, movement intention is predicted from EEG data, and downstream system components either characterise movement properties [43, 44] or modify the kinematics of robotic actuation [45, 46]. Others extend this principle to a "cascaded" design for multi-gesture classification: one datatype places samples in broad categories and, conditional on these decisions, different downstream models predict specific classes with data of the other modality [47, 48].

In certain studies multiple sensor modalities used in parallel serve separate roles, with their data being used for wholly **decoupled** tasks. In some the different modalities' tasks are unrelated, e.g. control of different robotic joints [49] or providing distinct control mechanism options which can be cycled between [50]. Others use data of one modality to monitor or error-correct a classifier based on the other, measuring physical resistance against exoskeletal movement [51] or the error-related potential (ErrP) neural signal [52] to identify erroneous predictions.

Feature-Level fusion, a type of early fusion [53], is the combining of data modalities after informative features have been extracted but prior to model fitting. This fusion strategy has been applied to a variety of problems including with amputee data, with studies reporting improvements in model accuracy [29] or competitive results with EMG-only models [54]. Other studies of early fusion systems show promising results but do not compare the resulting systems' performance with unimodal EEG based classifiers [18, 30], making it difficult to ascertain the extent to which the early fusion itself contributed to performance.

Finally, **Decision-Level fusion**, also known as late fusion [53], involves the processing and parallel classification of each data modality by independent models, with the resulting predictions combined in an

ensemble. Strategies for this can be categorised into two broad groups: those which implement a static rule or formula for combining predictions, and those wherein an additional ML algorithm is stacked onto the end of the classification process as a meta-model.

In multiclass problems, predictions of different data modalities' classifiers are best combined in the form of classwise probability estimates, providing richer data than simply a class label. Distributions can be fused with simple rules [19], but more common is for a system-level probability distribution to be calculated, through averaging those predicted by the constituent classifiers. Leeb et al. [55]'s seminal work on biosignal fusion found taking the arithmetic mean of EMG and EEG models' probability estimates could significantly outperform those single-mode classifiers individually. Other studies have experimented with weighting the distributions resulting from each modality either through predefined weights [54, 56] or based on estimates of the respective unimodal classifiers' reliability [32, 57].

The work of Cui et al. [19] is as far as we are aware the only prior investigation comparing a range of biosignal fusion strategies for the classification of distinct movements: walking, cycling, and repeated stepping up to and down from a raised surface (it should be noted that Tryon et al. [54, 56] do assess multiple strategies, but to classify variations in force of the same fundamental motion, not distinct movements). Their work is a rare example among biosignal fusion literature of a stacked meta-model being used. Under this strategy a single classification algorithm received the classwise probability estimates returned by the constituent EMG, EEG, and Magnetomyography (MMG) models and trained to combine those to predict the movement class. This classification-based fusion strategy outperformed rule-based methods near-universally and, particularly with an SVM meta-model, was frequently more accurate than unimodal systems [19].

3. Methodology

As Section 2 demonstrates the recommendations of biosignal classification literature are diverse and often inconsistent. There is no consensus on best practice for the design of EMG or EEG classifiers nor is there an established optimal technique for their combination in a multimodal system. Where popular approaches emerge the strength of evidence for their adoption can fall short due to various methodological limitations. While many studies make efforts to employ some measures for ensuring the quality of results, the requisite standards are neither found consistently nor together. We present a strategy which addresses these flaws in methodology. It applies and extends a number of techniques for sound classifier design & comparison which, taken together, establish new standards for research in the field. These are detailed further throughout this section but summarised as follows.

Measures which aim to ensure unbiased, valid results:

- Avoiding bias or favouritism in algorithm selection, and demonstrating transparently that hyperparameters were not tuned with knowledge of test data, by designing classifiers with Automated Machine Learning
- Validating findings on data of wholly unseen "Holdout" subjects to avoid "over-hyping" [9] in optimisation, and transparently reporting where intermediary results are derived from data seen in optimisation
- Avoiding "temporal" leakage between time-correlated datapoints [8] by grouping data belonging to a given execution of a physical gesture together in any random data splitting
- Avoiding temporal leakage in the embedding of data by avoiding signal processing techniques which require advance knowledge of an entire waveform, such as non-causal filters, wherever possible
- Ensuring train/test separation is not violated by using only training data for standardisation & selection of features, and applying identified transformations naïvely to testing data.

Measures which aim to ensure transparency and reproducibility:

- Reporting in detail hyperparameter values and optimisation ranges for machine learning algorithms, and justifying their choice with respect to prevailing literature
- Comparing novel models/architectures to those typical of current literature directly (i.e. applying both to the same problem, on the same dataset)
- Reporting all relevant results, including those of small effect size, and testing for statistical significance where appropriate.

We start this section by fully specifying the problem we are addressing, then proceed to describe the candidate fusion architectures and classifiers tested. To avoid making modelling choices without sufficient evidence for their suitability we consider a range of options with literature precedent for fusion architectures, ML algorithms, and hyperparameters. We then detail the data preprocessing and feature engineering steps employed, and the approach we used to guarantee fairness in the assessment and comparison of the candidate approaches.

3.1. Problem specification

The public biosignal dataset used in this research was collected by Jeong et al. [34]. It comprises Electromyographic, Electroencephalographic, and Electrooculographic signals recorded from 25 participants, all right-handed and all inexperienced with Brain-Computer-Interfaces, performing a range of upper-limb activities. 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, **Lateral grasp** to pick up a credit card, and **Spherical grasp** to pick up a cricket ball. Full details of the gestures and data collection procedures are provided in the original source [34] and will not be reproduced here, but a summary is provided below.

Twenty-five subjects performed each of these three gestures 50 times for 4 seconds each, in a randomised order. Each subject 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. EEG data were sampled at 2500Hz from 50 electrodes arranged in accordance with the International 10-10 system [58]. In this work, as is well-precedented in BCI studies [59–61], the EEG were trimmed to 20 channels situated near brain regions relevant to the planning, execution, and sensation of movements: Electrodes FC1-6, Cz-6, and CPz-6 were retained as highlighted in Figure 1. For EMG collection, six electrodes were positioned over the *extensor digitorum*, *extensor carpi ulnaris*, *flexor carpi radialis*, *flexor carpi ulnaris*, *triceps brachii*, and *biceps brachii* muscles of the right forearm (shown in Figure 4 of [34]).

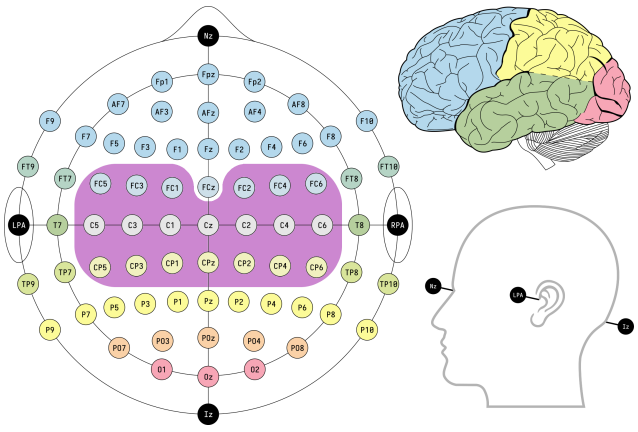


Figure 1. Electroencephalography (EEG) electrode channels used (shaded purple) according to International 10-10 system [62]. Adapted from [63]; originally published under CC0 1.0.

We investigated classifiers considering two use cases. The first is subject-independent: we assume no access to our intended user during development, thus requiring a system which an unseen user could use out-of-the-box. We simulate this use case by allowing only the use of non-subject data (i.e. data from subjects other than the one on whom we test a system) to train and configure our classifier. We call systems developed in this way "Generalist" systems. In our second use case we assume that we have access to our intended user to train and configure a subject-specific system. We model this by allowing 67% of each subject's data to be used to develop a system in addition to the non-subject data allowed for "generalist" systems. Specifically we use non-subject data to help configure our system, and exclusively subject data to train it. We then tested such systems on the remaining 33% of the subject's data. We call systems developed in this way "Bespoke" systems.

3.2. Candidate Fusion Architectures

3.2.1. Feature-Level Fusion In the Feature-Level Fusion architecture, illustrated in Figure 2, features derived from EEG and EMG data are merged prior to their classification by a single model, as explored in works including [29, 30, 56]. Data of the two modalities are first processed, and statistical features are extracted, independently. The method of feature selection then defines two subtypes of this architecture. In the separate-selection variant, informative features are selected from EMG and EEG featuresets independently, then joined into a single set for model training. This ensures the final feature ensemble exploits both data modalities. However, should EMG features be included which are highly correlated with selected EEG features or vice versa, it may risk a reduction in the overall information captured. In joint-selection, the features of both data types are firstly merged, and features selected from this combined set. This approach can avoid including highly correlated features, so can reduce duplication of information, but may select different numbers of features from each data modality and thus might unduly exploit one less than the other.

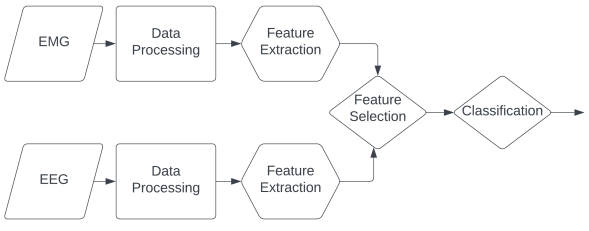


Figure 2. Feature-Level Fusion Architecture

3.2.2. Decision-Level Fusion The Decision-Level Fusion architecture, shown in Figure 3, encompasses a range of methods which use predictions made by parallel independent EMG and EEG classifiers to arrive at a final decision. Following the approaches discussed in section 2 we consider both rule-based and metaclassifier-based late fusion approaches.

Here, after Tryon et al. [54] among others, we trialled the **mean** alongside fixed **weightings in favour of each data type**: EMG predictions being weighted at w_{EMG}=0.75 and EEG at w_{EEG}=0.25, and vice versa. A **"tunable" weighted average** variant was included wherein the distribution of weights over data modalities was itself a hyperparameter that could be optimised. We also included the **maximum rule** seen in works such as [19], which selects the distribution with the highest probability in its respective highest-scoring class, favouring the predictions of confident models.

For the stacking approach we trialled two linear meta-model candidates (**linear SVM** and **LDA**) and one nonlinear model (**Random Forest**). To train these meta-models we employed a 3-fold cross validation procedure with the system's training data. The classwise probabilities returned by the base EMG and EEG classifiers in this cross validation were used to train the meta-model, and the base component classifiers subsequently retrained on all 3 folds of their respective training data.

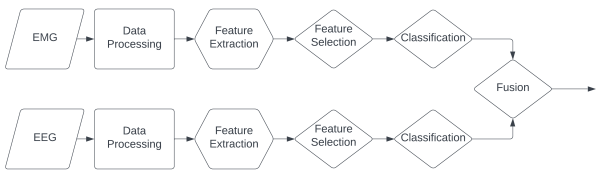


Figure 3. Decision-Level Fusion Architecture

3.2.3. Hierarchical Fusion The two-stage classification architecture described as *Hierarchical* is believed novel, at least in this domain, though it takes inspiration in part from the principles of cascaded approaches in literature (see Section 2) wherein data modalities serve distinct *consecutive* roles. Here by contrast we make predictions with data of one modality, and it is these predictions we combine with the featureset of the other data modality. We thus incorporate principles of stacked generalisation, though contrary to conventional stacking wherein a meta-model learns from multiple base classifiers' outputs, we inject the outputs of one model to a "higher-ranking" model before the latter makes its prediction. This is performed on a probabilistic basis. The lower-level model is used to predict the probability distribution of any given sample over the four classes. This distribution is then used to supplement that sample's entry in the higher-level model's (post-feature-selection) featureset. Higher-ranking classifiers were trained using an equivalent K-fold cross-validation to that used for meta-models in Decision-Level Fusion (Section 3.2.2).

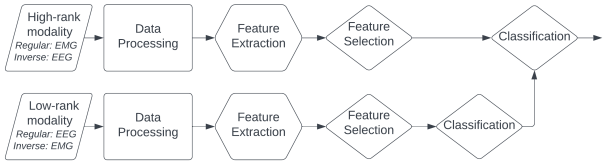


Figure 4. Novel Hierarchical Fusion Architecture

We trialled both possible arrangements of data modalities in this Hierarchical architecture, illustrated in Fig. 4. The case wherein class probabilities predicted by an EEG model were used to supplement EMG data is referred to hereafter simply as *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 orientation, wherein the model considering primarily EMG data outranks its EEG counterpart, is thus considered the architecture's default configuration. The opposite, wherein probability distributions obtained from an EMG model are joined with EEG data is hence named *Inverse Hierarchical*.

3.3. Candidate Classifiers and Hyperparameters

The candidate classifiers for use on the EMG and EEG data (or joint dataset in Feature-Level fusion) were selected from those with precedent in literature. For each classifier we also considered a number of hyperparameters and associated tuning ranges which were conditional on their associated model being constituent in a system. The overall modelling space is presented in Table 1; the full rationale of our hyperparameter ranges and other implementation details is provided in the Supplementary Material, Section 1.

Table 1. Joint algorithm-hyperparameter space for a single classifier.

Algorithm	Hyperparameter	Tuning Range
RF	# Trees	10 - 100 (steps of 5)
KNN	k	1 - 25
LDA	Solver	SVD; Eigen; LSQR
	Shrinkage	0.0 - 1.0
QDA	Regularisation	0.0 - 1.0
GNB	Smoothing	1e-9 - 1.0 (log scaled)
SVM‡	C	0.1 - 100.0 (log scaled)
	γ	0.01 - 1.0 (log scaled)

In Decision-Level and Hierarchical fusion, separate classifiers are used for EMG and EEG data. Because hyperparameters well-suited for EMG classification may not be well-suited for EEG data and vice versa, we allow these separate classifiers to be configured independently. The configuration space for these fusion architectures thus comprises two instances of the space described in Table 1, as separate axes. In Decision-Level fusion the search space additionally contained a subsection defining hyperparameters used to choose and configure the Decision-Level algorithm, presented in Table 2.

Table 2. Joint algorithm-hyperparameter space for Decision-Level Fusion algorithms

Decision-Level Algorithm	Hyper-parameter	Tuning Range
Mean	-	(Static Hyperparameters)
EMG-Weighted	-	(Static Hyperparameters)
EEG-Weighted	-	(Static Hyperparameters)
Tunable Weighting	w_{EEG}	0.0 - 100.0
Maximum Rule	-	(No Hyperparameters)
Stacked Linear SVM	C	0.01 - 100.0 (log scaled)
Stacked LDA	Solver	SVD; Eigen; LSQR
	Shrinkage	0.0 - 1.0
Stacked RF	# Trees	10 - 100 (steps of 5)

3.4. Data Preprocessing and Feature Extraction

Data were preprocessed in Mathworks MATLAB R2020a [64] with a script adapted from those provided by Jeong et al. [65]. EEG signals were bandpass filtered from 2 - 30 Hz with a 4th-order Butterworth filter, and EMG were rectified and filtered from 10 - 500 Hz using a 5th-order Butterworth filter. We used the MATLAB *filter()* function here, rather than the zero-phase *filtfilt()*, since the latter is non-causal and requires advance knowledge of an entire waveform. This would risk information from later points in time leaking into earlier datapoints, and would also be impossible in a real-time system. Though our experiments were performed offline, we strived to avoid techniques which would be unviable for real-time classification. For the same reason, no specific steps were taken to remove any EEG artefacts associated with participants blinking. Conventional methods for this involve offline Independent Component Analysis [66] and related techniques; while real-time strategies have been proposed they can often introduce significant delay [67] or require advance knowledge of artefact characteristics [68].

Individual three-second gesture performances and subsequent rest periods were then extracted from the processed EMG and EEG. This demarcated the 50 performances of each grasp type and 150 rest periods; a pseudorandom sample of 50 rests was retained to balance the classes. The resultant dataset thus comprised a total of 200 gesture performances (including rests) per participant per session; 600 gestures per subject in total.

We extracted features from 1 second windows overlapped by 50%. After [69], features from each window were joined with those of its immediate successor to form datapoints corresponding to 1.5s. Thus from each 3s gesture performance four datapoints were extracted, of which each except the first shared one window with its predecessor and introduced 500ms of new data, resulting in 2400 observations per subject.

The specific ensemble of features used in this work is given in Supplementary Table 2 and includes time-domain, frequency-domain, and correlation-based features. It been used successfully in previous work classifying both EEG [69] and EMG [38, 70] data. These features were extracted from each time window of both EMG and EEG data§. Feature extraction was performed independently on the EMG and EEG datasets, enabling the two modalities to be handled separately throughout the modelling pipeline until merged per the fusion strategies outlined in section 3.2.

EMG featuresets were reduced to the 15% with the highest ANOVA F-values, identified from only the training data of a given modelling process, for a resultant 88 features. In EEG however, the greater number of sensors from which features were derived meant that a reduction by the same proportion would be insufficient to avoid overfit, and vastly outnumber EMG features. Instead, L1-norm based selection (using a linear SVM with $C = 0.005$) was used to retain a fixed number of features. In bespoke systems 40 attributes, the approximate square root of a subject's 1600 training samples, were retained. The square root of generalist datasets' length, being much larger, would be unsuitably high; these were instead reduced to 88 features for a consistent width with EMG. For parity the merged featuresets of Feature-Level fusion systems werethe same width as the total number of features available to other architectures: 128 in a bespoke case and 176 in a generalist. In the separate-selection subtype these were selected from the modalities independently as described; in joint-selection, they were selected from the combined set of all EMG and EEG features with the above L1-norm method.

Finally, the training data were standardised such that features had zero mean and unitary standard

‡ SVMs were included only in Bespoke systems for computational feasibility

§ Using a script adapted from `https://github.com/fcampelo/EEG_Classification_`, itself adapted from that of [69].

deviation, and the learned transformation used to map test data to the same space.

3.5. Combined Algorithm Selection And Hyperparameter (CASH) Optimisation

A suitable classifier design must be determined to develop gesture identification system, whether that be a pretrained generalist or a bespoke system trained on individual user data. By using a data-driven approach to determine modelling choices, we can avoid potential biases resulting from an arbitrary or unjustified selecting of models and hyperparameter values. Specifically, we use Automated Machine Learning [41] to find suitable configurations for each fusion architecture. That is, rather than select models solely on the bases of *a priori* assumptions or their dominance among literature, we optimise over the range of candidate models and hyperparameter values introduced in Section 3.3 to identify configurations resulting in the highest classification accuracy. Our architectures typically comprise multiple classifiers in an ensemble; we need to both select ML models for these components and set their hyperparameter values. We thus modelled the determination of classifiers and the tuning of hyperparameters together as a Combined Algorithm Selection and Hyperparameter (CASH) [71] optimisation problem. CASH (also known as Full Model Selection - FMS) is a specific problem within AutoML in which model choice and hyperparameters are optimised simultaneously. Distinct from a typical paradigm of hyperparameter optimisation wherein a range of values for one or more hyperparameters are considered for one specific model, in CASH the choice of the classifier algorithm is itself tunable in the configuration space. This represents to our knowledge the first application of CASH to the field of biosignal classification.

Of the available AutoML frameworks, we chose hyperopt [72] for its particular suitability to our architectures. Crucially, in an ensemble architecture like ours, the choice of algorithm for each classifier is effectively a separate axis in the search space. This allows for a better chance of discovering algorithm combinations which might not be investigated if components were optimised separately. Hyperparameters defining the properties of these different algorithm choices are conditional, being well-defined only when the applicable classifier is selected (i.e. when the "model choice" parameter has a certain categorical value). This tree-like structure of the optimisation space makes *hyperopt*'s use of Tree-Structured Parzen Estimators (TPE) [73] particularly well suited to the problem we are investigating in this paper.

TPE is a form of Bayesian optimisation in which, initially, a set of configurations $\mathcal{D} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$ are sampled at random from the configuration space; each $\boldsymbol{x_n} = \{x_{n,1}, ..., x_{n,D}\}$ is a complete set of model choices and hyperparameter values. These configurations are evaluated based on an objective function f. In this study, the objective function will be based on the quality of models trained using a given configuration as described in Section 3.8.

The configurations $\mathcal{D}$ are split into two subsets based on a quality threshold, $y*$, and the probability densities of the parameters in the resulting sets are modelled independently:

$$p(\boldsymbol{x}) = \begin{cases} l(\boldsymbol{x}), & f(\boldsymbol{x}) < y* \\ g(\boldsymbol{x}), & f(\boldsymbol{x}) \geq y* \end{cases}$$

Assuming that we are attempting to minimize f then, within *hyperopt*, $y*$ is chosen such that the best k observations fall below the threshold for some fixed value of k.

For their respective subsets of $\mathcal{D}$, the models l and g are formed by independently modelling the densities in each dimension within the configuration space using kernel density estimation [74]. Given a set of samples $\{\boldsymbol{x}_n\}_{n=1}^{N}$ to model, a joint distribution is defined as follows:

$$p(\boldsymbol{x}|\{\boldsymbol{x}_n\}_{n=1}^{N}) = \frac{1}{N}\prod_{d=1}^{D}\sum_{n=1}^{N} k_d(x_d, x_{n,d})$$

where $x_{n,d}$ is the d-th dimension of the n-th sample and k_d is the kernel function for the d-th dimension. Kernels may be chosen per dimension based on the type of hyperparameter to be modelled, with truncated Gaussian kernels being used within *hyperopt* [73]. Treating the dimensions of the configuration space independently in this way allows the modelling of the density in the context of conditional hyperparameters.

A new set of configurations is sampled from the distribution of the better group, $\mathcal{S} = \{x_s\}_{s=1}^{S} \sim l(\boldsymbol{x})^S$. Because evaluating a configuration can be expensive (in this study requiring the training and testing of multiple models) we choose only the most promising candidate configuration $\boldsymbol{x}*$ from the set:

$$\boldsymbol{x}* = argmax_{\boldsymbol{x}\in\mathcal{S}}\frac{l(\boldsymbol{x})}{g(\boldsymbol{x})}$$

Intuitively, the configuration selected as a result of this should have a high chance of being a member of the "better" group and a low chance of being a member of the "worse" one. This $\boldsymbol{x}*$ is evaluated using the objective function and added to the set of observations $\mathcal{D} := \mathcal{D} \cup \boldsymbol{x}*$. The process repeats until termination.

The candidate fusion architectures described in 3.2 are diverse, and it cannot be assumed that the same modelling choices would be optimal for all. CASH optimisation was thus performed independently for each fusion architecture, to enable fair comparison

between them on the basis of their respective "best-in-class" configurations.Likewise, bespoke and generalist settings of each architecture were optimised separately on the grounds of their distinct natures. For consistency and to enable fair comparison, in all cases optimisation processes were afforded a budget of 100 iterations. This figure is somewhat arbitrary, but exploratory experiments suggested it as capable of ensuring sufficiency while maintaining computational feasibility, typically completing a CASH optimisation routine within 2 hours for a Bespoke system and 20 for a Generalist (wherein training data are much larger). The cost of this computation while not negligible must be weighed against the costs in time and expertise required to design a suitable equivalent system manually – *without* the aid of an automated methodology such as ours.Though beyond the scope of this work to quantify, these cannot be dismissed. Nor can the challenges inherent to such manual design, both in the risk of experimenter bias previously mentioned, and the extent to which useful inferences can be drawn from a diverse and often limited literature base — as the following section highlights.

3.6. Benchmark Classifiers

We have motivated the use of CASH with reference to potential issues resulting from simply assuming that common modelling choices from the literature will be performant on new problems. To test the utility of this approach, we compared our CASH-derived classifiers to two benchmarks representing common trends in the biosignal literature. The first is based on classical ML models and uses modelling decisions inferred solely from synthesising literature on biosignal fusion. The second is a Deep Learning-based unimodal EMG classifier, representative of popular domain trends.

The rarity of direct comparisons between different fusion architectures in the biosignal classification literature makes evidence for such inferences scarce. From [19], unique in comparing fusion methods for multi-gesture classification, we derive our **classical ML bespoke benchmark** fusion algorithm. We draw on both [19] & [56] in the choice of its EMG component classifier, though in the absence of hyperparameter valueswe defer to their default values in *scikit-learn 0.24.2* (while [19] states that some hyperparameters were determined through cross-validation, neither optimisation ranges nor resultant values are provided). From [54], perhaps the only prior subject-independent comparison of fusion strategies, we derive our **classical ML generalist benchmark** fusion algorithm and EMG model. Again we must fall back on library defaults for hyperparameters, though as [54] notes implementation in *MATLAB* we refer to its documentation [75] to determine the SVM kernel. Though not ubiquitous in fusion, to dismiss the dominance of LDAs [6] in EEG classification would be to disregard literature from which we seek to draw inferences; these are hence chosen for both bespoke and generalist EEG component models, each using the *scikit-learn* default solver.

Table 3. Literature-Derived Baseline Systems

Design Element	Implementation	Source
Bespoke		
Fusion Algorithm	SVM metamodel	[19]
	RBF Kernel	[19]
	C = 1.0	*scikit-learn* 0.24.2 default
	$\gamma = \frac{1}{n\_features}$	*scikit-learn* 0.24.2 default
Component EMG	SVM	[19, 56]
	RBF Kernel	[19, 56]
	C = 1.0	*scikit-learn* 0.24.2 default
	$\gamma = \frac{1}{n\_features}$	*scikit-learn* 0.24.2 default
Component EEG	LDA	[6]
	SVD Solver	*scikit-learn* 0.24.2 default
Generalist		
Fusion Algorithm	Mean	[54]
Component EMG	SVM	[54]
	Linear Kernel	*MATLAB Statistics and Machine Learning Toolbox* default [75]
	C = 1.0	*scikit-learn* 0.24.2 default
Component EEG	LDA	[6]
	SVD Solver	*scikit-learn* 0.24.2 default

Notwithstanding the discussion in section 2 of the limitations of Deep Learning architectures for biosignal classification problems, they remain the focus of many studies, particularly in the EMG domain. Our work is motivated more strongly by efforts to explore and evaluate methodologies for multimodal fusion and classifier development than it is by any attempt to definitively claim superiority of a given model over those of prior research. Nevertheless, it is useful to compare the predictive abilities of the systems we develop to that of an established architecture of the type popular among recent research. Unimodal EMG systems being dominant among literature (and indeed state-of-the-art prostheses) serve as an appropriate standard against which to assess our systems. We use the Convolutional Neural Network architecture of Lehmler et al. [76] as our **deep learning based** benchmark, chosen because they test both subject-specific and subject-independent settings comparable to our bespoke and generalist paradigms.

To ensure a fair comparison, the experimental conditions of the EMG-CNN comparators were the same as those of our multimodal systems. They were tested on the same problem with the same dataset, using the same signal preprocessing steps and applying the same stringent data handling strategy outlined below in defense against leakage. Signals

were windowed in the same manner described in section 3.4, though to exploit the CNN's feature extraction capabilities the data were provided raw. The model architectures were exactly as described in [76], extracted from their code repository. The sole exception is that early stopping criteria for CNN training were removed in our Bespoke case. Our datasets for subject-specific modelling are much smaller than those of [76]; early stopping led training to end prematurely, which would give an unfair portrayal of the CNNs' performance.

3.7. Data Handling

We take particular care in our experiments to minimise modelling and assessment issues caused by mishandling of data, which would compromise the validity of our findings. We implemented more stringent data handling standards than commonly found in the related literature, including the strict preservation of train/test data separation at the exploratory data analysis, data preparation and feature engineering stages [7]. We also go beyond usual good practice in two important areas.

3.7.1. Avoiding leakage through model selection or over-optimisation Optimising a set of classifiers over, and selecting the highest-performing classifier on the basis of its performance on, a given dataset is itself a learning process. To draw conclusions from systems' ability to classify data upon which they were selected would introduce leakage by undermining train/test separation [7,77]. This would risk reporting artificially inflated performance measures and compromising the generalisability of inferences drawn from the results. Hosseini, Powell, et al. discuss the prevalence of such over-optimisation among biosignal literature [9], and highlight the need for findings to be verified on data wholly unseen by any modelling or optimisation procedures.

Here we go beyond these recommendations by splitting the data at the participant level: data from five participants were withheld throughout experimentation, excluded from all parts of exploration, preprocessing, modelling, and testing, and not accessed until such time as they were used exclusively to verify observations and test specific hypotheses. The data from these held-out participants (IDs 1, 6, 11, 16, and 21) is referred to hereafter as the **Holdout set**‖; the remaining 20 subjects are denoted the **Development set** upon which modelling decisions can be made. These subjects were selected, without any sight of their biosignal data, to preserve a consistent mean age (27.8 years) and proportion of female participants (40%) across both Development and Holdout sets.

‖ Described by Hosseini, Powell, et al. as "Lock-Box" [9]

3.7.2. Avoiding leakage through temporal dependencies To protect against temporal data leakage [8], wherever data were further split this was done at the level of gesture performances - i.e., all datapoints from any given execution of a gesture were grouped together. Data collected at consecutive time intervals of the same gesture performance were not distributed among the training and testing splits, thereby preventing time-series correlations in the data from violating the train/test independence requirement.

3.8. Experimental Procedure

Figure 5 illustrates our procedure for CASH-based development of a gesture classification system; Bespoke and Generalist versions of every architecture being each afforded an optimisation budget of 100 iterations. Our Bespoke systems were designed to be "portable" - every candidate set of modelling choices was trained and tested on a subject-specific basis for all 20 Development Set subjects, and the configuration which maximised the mean per-participant accuracy was selected. We developed our Generalist systems using leave-one-subject-out cross-evaluation: a candidate configuration was trained on 19 subjects and tested on the remaining one, for each Development Set subject in turn. Again the mean per-participant accuracy was the optimisation target.

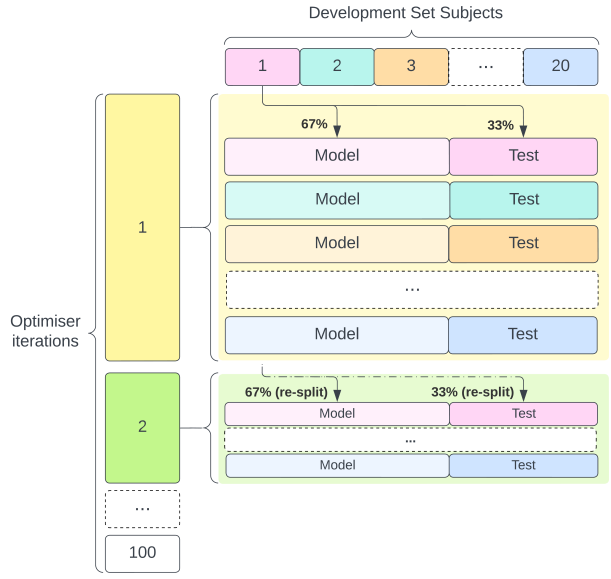
For a fair assessment of the merit of EMG-EEG fusion, performant multimodal systems were compared with equivalent single-mode systems, to test the null hypothesis that *a multimodal system's performance will be no greater than a comparable unimodal one* versus the one-sided alternative, i.e.:

$$H_0 : \mu_{fusion} - \mu_{unimodal} \leq 0 \quad (1)$$

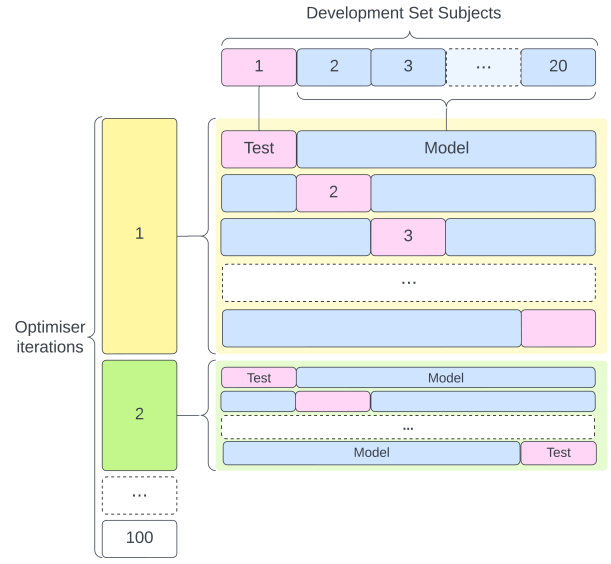
$$H_1 : \mu_{fusion} - \mu_{unimodal} > 0 \quad (2)$$

To ensure fairness in this comparison, we designed single-classifier EMG and EEG models with the same CASH procedure used for Fusion systems, each afforded the same optimisation budget and search space (Table 1).

Candidate fusion and unimodal architectures were selected according to their predictive power over the Development Set. To avoid over-optimisation, our formal tests were subsequently based on their ability to predict the wholly unseen Holdout Set. In the Bespoke case, Holdout tests were conducted as the mean of 100 trials, to mitigate the random effect of the 67/33 train/test splitting of Holdout subjects' data. This was not necessary for Generalist candidates, which were modelled using all 20 Development subjects' data and tested once on all the data of each Holdout in turn.



(a) Bespoke



(b) Generalist

Figure 5. Data flow within Combined Algorithm Selection & Hyperparameter (CASH) optimisation procedures for Bespoke & Generalist settings

The code used to run our experiments is found at `https://github.com/mgpritchard/emg-eeg-CASH`.

4. Results

4.1. Determining candidate models

Table 4 presents the peak accuracies achieved in CASH optimisation of each architecture¶.

¶ Note that both the Hierarchical and Inverse Hierarchical systems reporting a mean Generalist accuracy of 71.68% is not a typographical error.

Table 4. Peak mean Development Set accuracies achieved in CASH optimisation of Multimodal & Unimodal architectures.

Architecture	Mean ±SD Accuracy (%)	
	Bespoke	Generalist
Decision-Level	87.83 ±4.286	71.67 ±7.381
Feature-Level (Separate selection)	86.13 ±4.682	72.03 ±7.224
Feature-Level (Joint selection)	86.48 ±4.836	**72.30 ±7.329**
Hierarchical	**88.98 ±4.519**	71.68 ±7.196
Inverse Hierarchical	83.68 ±6.956	71.68 ±7.222
Unimodal EMG	**87.78 ±4.263**	**68.90 ±6.780**
Unimodal EEG	54.80 ±8.242	49.11 ±6.330

In Bespoke Fusion, the highest accuracy was reached by the novel Hierarchical architecture. In its winning configuration the lower-ranking EEG model was a QDA with a Regularisation value of 0.4559, and the higher-ranking supplemented EMG model was an SVM with $C = 19.4037$ and $\gamma = 0.0138$. It is noted here that the degree of difference between this architecture's performance and that of the next best-performing system (the Decision-Level fusion), and indeed the Unimodal EMG, was quite small in these experiments. A small effect size such as this could be caused by an uncontrolled covariate, such as the presence of blinks in the EEG data[+] being indicative of movement, though we did not observe notable differences in error distribution between the various systems. In a case like this where one data modality is generally a better predictor of the class than the other, a Hierarchical system may be better able to down-weight the worse modality than in conventional Decision-Level fusion; this is a property of our architecture worthy of further investigation. Such an apparently small performance difference between our chosen fusion system and its unimodal counterpart may well indicate fusion as insufficiently beneficial in some applications if a simpler EMG based system can suffice. That decision would be both problem- and context-dependent and thus not one for which we can be prescriptive. That our methodological framework can identify and measure this difference however, given the rarity of such direct comparisons among pre-existing literature, is evidence of its utility. The best-performing Generalist architecture was the Joint Selection subtype of Feature-Level Fusion, using a single LDA with the Least Squares Solution solver and a Shrinkage value of 0.1871.

Among Unimodal classifiers, the optimised EMG models were superior to the EEG in both Bespoke and Generalist paradigms. An EMG-SVM with $C =$

[+] These will have been partially suppressed by the bandpass filter but as described in 3.4 were not specifically removed.

Table 5. Classification accuracies (%) of CASH-derived candidate Multimodal and Unimodal systems, and Literature-informed Benchmarks, in predicting unseen Holdout dataset, in Bespoke and Generalist settings.

Bespoke (means ±std dev of 100 trials)				
Subject ID	Literature-derived: SVM fusion	Comparator: EMG CNN	CASH: Unimodal EMG	CASH: Hierarchical fusion
1	75.51 ±2.180	53.31 ±2.679	80.32 ±1.418	82.93 ±1.577
6	80.72 ±2.277	61.02 ±3.030	82.49 ±1.492	83.45 ±1.524
11	92.80 ±1.070	70.31 ±3.258	94.55 ±0.939	94.67 ±0.838
16	79.96 ±2.565	61.16 ±2.495	83.76 ±1.424	83.24 ±1.597
21	80.71 ±2.008	64.53 ±2.133	86.50 ±1.452	86.97 ±1.295
Mean	81.94 ±6.446	62.07 ±6.177	85.52 ±5.519	86.25 ±4.983
Generalist				
Subject ID	Literature-derived: Mean fusion	Comparator: EMG CNN	CASH: Unimodal EMG	CASH: Feat.-level fusion (joint selection)
1	68.29	57.17	59.33	66.33
6	73.38	68.08	70.88	74.75
11	83.17	74.71	79.33	82.33
16	69.83	68.29	69.71	72.46
21	70.38	67.96	66.38	71.21
Mean	73.01 ±5.972	67.24 ±6.315	69.13 ±7.264	73.42 ±5.857

4.1725 and $\gamma = 0.0126$ was selected in the Bespoke case, and for the Generalist an EMG-LDA using the Eigenvalue Decomposition solver with Shrinkage equal to 0.0744. That the selected unimodal systems represent algorithm choices popular among literature is fortunate. It means our candidate EMG-EEG fusion architectures can be compared with domain-standard approaches to gesture classification, while maintaining the fairness of the design of these competitors which our CASH-based methodology enables.

Though we leave detailed analyses of the stability of choices made by CASH in optimising biosignal classifiers for future work to investigate in depth, we note here that in all four of these selected candidate architectures, the same ML algorithms were used (albeit with varying hyperparameter values) in the 10 highest-scoring observations made throughout their respective CASH optimisation routines. The only exception was the Bespoke Hierarchical system's lower-ranking EEG model, for which a QDA was indeed chosen in the top 5 optimisation iterations but the next most accurate 5, while still using SVMs for their top-level EMG models, included KNNs, a QDA, and a GNB. The CASH-identified configurations of non-winning architectures are provided in the Supplementary Material, Section 3.

Though not the primary focus of investigation, the subject-independent EEG accuracy achieved here should be highlighted. At 49% it is well above the chance level — which for a Generalist, wherein all 150 of a subject's gesture performances were used for testing, has an upper confidence interval of 29% at the α=0.05 confidence level [78] — and remarkably is competitive with many subject-specific attempts at multi-gesture EEG classification: 51% in [79], 44% in [33], 50% in [34], though in 2-grasp classification [60] reports 76%. Fazli et al.'s comparable leave-one-subject-out approach did exceed our Generalist accuracy at 73% [80], but on a simpler left-vs-right-hand task than the four same-hand gestures of this work. This is a strong indicator of the potential of CASH as a design methodology for EEG-BCI development.

4.2. Comparative tests with Holdout Data

The selected candidate multimodal and unimodal systems' accuracies over the 5 Holdout Subjects are shown in in Table 5 for both Bespoke and Generalist cases. Also in Table 5 are the accuracies of our two literature-derived benchmarks (see Section 3.6): multimodal systems using classical ML algorithms designed by manual synthesis of the EMG-EEG fusion literature, and a unimodal deep-learning-based competitor representing popular trends in the wider body of work on biosignal classification.

The latter enriches the scope of comparison between our multimodal systems and domain-leading unimodal classification approaches. Our Unimodal EMG CASH procedures identified the SVM and LDA, both popular choices in literature, as performant models for comparison. The inclusion of this deep-learning-based competitor accounts for a family of algorithms which, due to the reservations we outline in Section 2, were excluded from consideration and could not have been generated by our CASH approach.

5. Discussion

5.1. Value of Fusion

In both Bespoke and Generalist settings, the observed differences between our candidate fusion and unimodal systems' accuracies indicate the capability of our

CASH-based methodology in identifying performant multimodal designs. To test the hypotheses in Equation 1 we performed paired samples t-tests, which indicated that in the Bespoke case this difference was not statistically significant (t = 1.3771, p = 0.1203). Nevertheless to have attained an equivalent accuracy to the "domain-standard" EMG-SVM approach demonstrates the potential of the novel Hierarchical Fusion architecture. In the Generalist case however this difference is statistically significant at the 95% confidence level (t = 5.5761, p = 0.0025). The optimised Generalist Feature-level Fusion system offered a distinct improvement over the equivalently-optimised Generalist Unimodal EMG system in classifying the four same-hand gestures in this task. EEG data, even if of somewhat low reliability in isolation, is thus demonstrated as meriting incorporation to hand-gesture classifiers in zero-calibration contexts.

5.1.1. Exploring the impact in low-fidelity EMG settings Considering the likelihood of reduced EMG fidelity among amputee populations, it should be noted this improvement may well be more distinct for that group. Indeed while further investigation would be needed to confirm as such, it appears plausible that the observable difference in Bespoke accuracy, while not found significant here, could in fact be so for amputees.

Amputee data, particularly multimodal data, is scarce, and to gather sufficient data to fully investigate this would be beyond the scope of the current study. We can however artificially modify our EMG data to carry less information. This will of course not be an equivalent substitute for amputee data, but can enable a provisional exploration of the impact of EMG-EEG fusion in situations where EMG is less informative, and whether our CASH-based methodology can be effectively applied.

The characteristics of amputee EMG in the residual limb are distinct from that of non-amputees and though methods have been proposed [81] there is not yet a domain-standard technique for transforming able-bodied subjects' EMG to emulate amputee data. For the purposes of this small-scale experiment, we opt degrade the able-bodied subjects' EMG signals in two ways: attenuation and noising, both precedented for the artificial corruption of EMG data [82, 83]. We attenuated the filtered, rectified EMG by 50%, and subsequently injected Gaussian white noise with MATLAB's *awgn()* function at a Signal-to-Noise Ratio (SNR) of 5 decibels (dB). This SNR was an arbitrarily chosen value below the 20dB typical of an able-bodied subject but well above the extremes of <1.5dB seen in paralysis [83, 84]. Figure 6 illustrates an example of both original and degraded EMG signals.

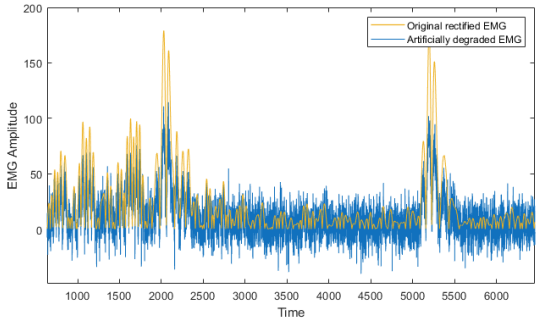


Figure 6. Illustrative comparison of both original processed EMG (yellow) and artificially corrupted EMG (blue).

We then re-performed our experimental procedure using this degraded EMG alongside the original EEG data. Only the Bespoke setting was considered as it was here that the performance difference between unimodal and multimodal systems was not found significant. We used our CASH-based methodology to determine classifier configurations for both Unimodal EMG and Hierarchical Fusion (the best-performing Bespoke fusion architecture), the result of which is seen in Table 6.

Table 6. System configurations identified through CASH optimisation with artificially degraded EMG and corresponding peak mean Development Set accuracy (±SD)

System	Configuration	Accuracy (%)
Unimodal EMG	SVM C: 7.5395 Gamma: 0.0129	86.59 ± 4.502
Hierarchical Fusion	EEG: kNN k: 24 EMG: SVM C: 2.7866 Gamma: 0.0169	87.35 ± 4.942

We tested the identified configurations on every subject in the reserved Holdout dataset, again taking the mean of 100 trials each to account for the randomness introduced in data splitting. We also assessed the accuracy of the Literature-derived Comparator fusion system when faced with this degradation of EMG signal quality. The results in Table 7 indicate that a Hierarchical fusion approach continued to offer consistently more accurate gesture classification than an EMG-based classifier when EMG data were degraded. The effect size is generally low, though it is notable that the greatest improvement was found for the poorest-performing individual (Subject 1); the same trend can be observed in our main results with unmodified EMG in Table 5. This may suggest that the multimodal fusion of EMG & EEG data is not of universal benefit in Bespoke systems but has more potential value in subjects for whom EMG-based performance alone is substandard. Of

Table 7. Classification accuracies (%) of CASH-derived Bespoke Unimodal and Multimodal systems, and Literature-derived Benchmark, in predicting unseen Holdout dataset with artificially degraded EMG.

Subject ID	Unimodal EMG	Hierarchical Fusion	Literature-derived: SVM fusion
1	77.63 ± 1.934	79.96 ± 1.706	73.96 ± 1.992
6	79.96 ± 1.585	80.72 ± 1.625	79.00 ± 2.289
11	93.14 ± 0.989	93.28 ± 1.170	91.29 ± 1.437
16	81.72 ± 1.604	82.20 ± 1.561	78.80 ± 2.375
21	86.60 ± 1.286	86.63 ± 1.423	81.85 ± 1.927
Mean	83.81 ± 6.169	84.56 ± 5.518	80.98 ± 6.423

course, while the attenuation and noise injection we performed here enabled this cursory investigation, they do not reflect the genuine modifications to EMG following amputation, and further confirmatory testing with amputee subjects would be needed for a more conclusive evaluation.

That the CASH-derived fusion system routinely outperformed the comparator fusion system inferred from literature reinforces the value of our methodology for biosignal classifier design. Despite the poorer data quality potentially diminishing the applicability of literature-recommended modelling choices, our approach enabled a performant classifier configuration to be identified, evidencing its robustness.

5.1.2. Comparing Fusion Architectures To enable comparison between Fusion architectures' predictive abilities, all architectures - not only the two identified in Section 4.1 as best-performing on the Development set - were deployed to predict the reserved Holdout set, each using their respective CASH-optimised configuration (see Supplementary Tables 5 & 6). We performed all-vs-all pairwise comparisons between the fusion architectures with Tukey's Honest Significant Difference test as implemented in the *R* package *multcomp version 1.4-26*. Figure 7 presents the simultaneous 95% confidence intervals of these pairwise comparisons between architectures' mean accuracy values, for Bespoke and Generalist classification respectively.

Among Bespoke systems, the Hierarchical and Decision-Level architectures jointly outperformed all other strategies, evidencing the capability of our novel approach. In the Generalist case, though the previously selected Feature-Level Fusion did attain a higher Holdout Set accuracy than others by a small magnitude, the performance differences between architectures were not statistically significant.

5.2. Assessing Literature-derived System Designs

The use of CASH to create optimised fusion systems allows us to assess the quality of systems designed only on the basis of literature recommendations. Table 5 compares the accuracies achieved by the candidate CASH-identified systems with the alternatives derived solely from literature inferences in Section 3.6 above. As discussed in Section 3.8, a fair test requires the CASH-derived candidates here to be selected without any knowledge of the Holdout data. They are thus the same Hierarchical and Feature-Level systems chosen based on their Development Set performance in Section 4.1, i.e., ignoring our later assessment of other architectures on the Holdout data.

In the Bespoke case, a paired one-sided t-test between the CASH-determined Fusion system and that derived from literature reports a t-statistic of 4.0222, at a p-value of 0.0079, with the 95% confidence interval on the mean of paired differences calculated as 0.0431 ± 0.0228. This result shows that a subject-specific fusion system designed solely on the basis of inferences from biosignal literature would underperform the best identifiable system by a significant margin, evidencing the limits of reliance on literature recommendations as a design method.

In the Generalist case however, though a positive difference in means was observed the t-test does not indicate a statistically significant difference between the CASH-determined system and that based on literature insights ($t = 0.5032$, $p = 0.3207$). This gives confidence to the literature-precedented design choices of Section 3.6 for subject-independent contexts.

The distinction in significance between the two classification paradigms could be in part due to inherent limitations in the usefulness of literature recommendations for Bespoke classification. By contrast to the "portable" nature of the Bespoke CASH optimisation here, wherein a single design was found which maximised mean subject-specific accuracy over all Development Subjects, many works optimising subject-specific models do so separately on a per-subject basis. Such studies' findings while legitimate are thus liable to be over-optimised to the particular individuals in their datasets. In the subject-independent case such specialisation is precluded; any recommendations arising are plausibly better able to generalise to new studies and thus more competitive here.

As demonstrated in Table 5, under both paradigms our CASH-identified multimodal systems

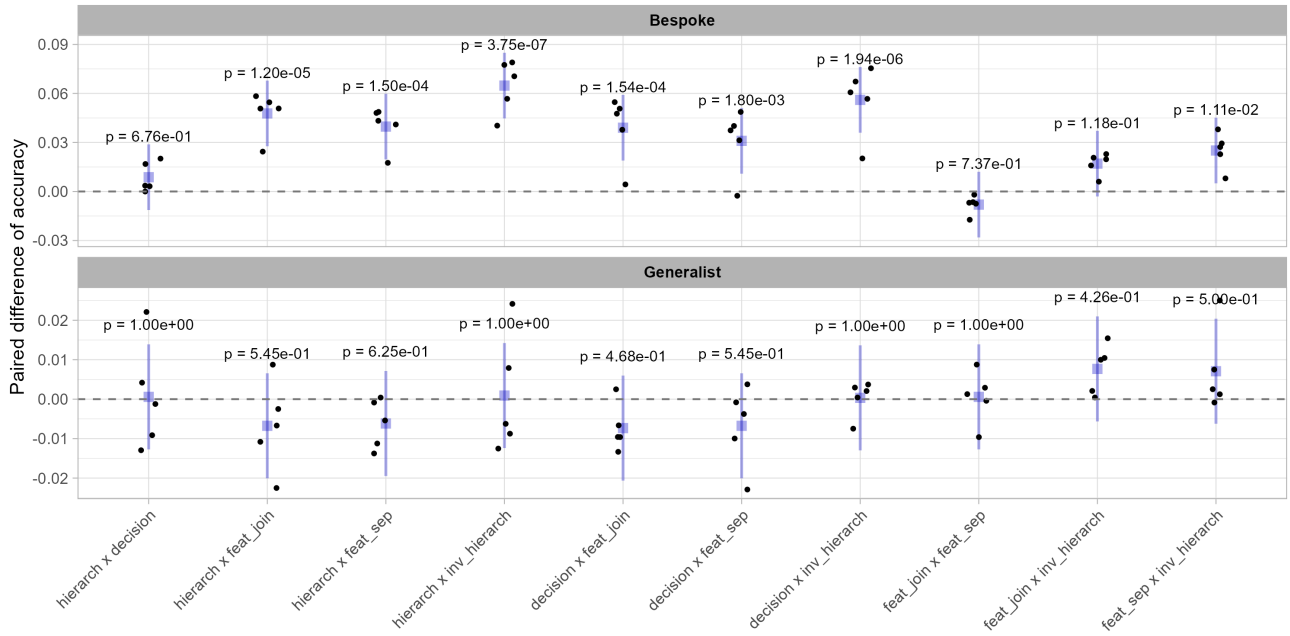


Figure 7. Simultaneous 95% Confidence Intervals & corrected p-values for the paired differences of accuracy for CASH-configured Bespoke and Generalist systems of various fusion architectures on the Holdout set, with markers corresponding to differences for individual Holdout subjects.

significantly outperformed the comparator CNNs ($p \ll 0.01$ in both Bespoke and Generalist). This is of course not to posit that no Neural Network architecture exists capable of greater accuracy than our CASH-derived designs for this task, but to give a like-for-like benchmark against a domain-popular architecture representing a prevailing trend in literature sources.

5.3. Value of a CASH-based design pipeline

The proposed CASH-based methodology offers numerous benefits, providing an unbiased design tool to a field often lacking in sufficient transparency on how design choices are reached; easing the burden of literature analysis for those designing gesture classifiers; and giving confidence in the likelihood of identifying adequate design choices where literature inferences are limited.

As an **investigative tool for researchers** our methodology allows a robust, unbiased assessment of frequently used design and solution methodologies within the domain. This has enabled us to show that, while fusion-based systems outperform unimodal systems in each of our use cases, the difference is only significant in a Generalist setting. Given the additional burden to users of gathering EEG data, it is valuable for designers of such classifiers to have confidence that a Bespoke unimodal system can achieve comparable performance to a fusion-based system.

Our methodology has also allowed us to demonstrate that design trends observed in the literature are not always optimal. In the Bespoke case, a CASH-derived system was able to significantly outperform both a literature-derived fusion system and a comparator CNN. In the Generalist case the literature-derived system was competitive with the CASH-derived one. The similarity in accuracies between the two systems gives us strong evidence that the recommendations in the literature are indeed optimal for this case given the architectures considered.

As a **tool for designing classifiers**, our CASH-based approach has shown consistently strong performance, particularly in multimodal applications. In each use case studied, a CASH-derived fusion system achieved the highest average performance on the unseen Holdout dataset. A CASH-derived fusion system significantly outperformed each of the alternative systems considered (literature-derived fusion, EMG-based CNN, CASH-derived unimodal classifier) in at least one of the two cases considered. This demonstrates the suitability of CASH for contexts wherein literature-based recommendations have not been robustly justified. Our results highlight the caution needed when deferring design decisions to literature recommendations which may - due to insufficiently detailed reporting, inherent limits in generalisability across datasets, the inadvertent overlooking of performant configurations in favour of popular algorithms, or other limitations - fall short of expected performance, and the merit of an unbiased

design tool. That CASH-derived systems consistently attained competitive accuracies to literature-based ones even when not significantly superior also indicates its suitability as a reliable strategy for contexts in which system designers themselves do not have intimate knowledge of prevailing literature.

We additionally note that the multi-gesture accuracies achieved here, at 82% in the Bespoke case and 73% in the Generalist, exceed those of Ruiz-Olaya et al.'s recent work on the same dataset [85]. Though notably drawing on fewer EEG channels, in [85] a move-vs-rest accuracy of 70% was obtained but individual gestures could be distinguished with only 54% accuracy in binary grasp-vs-grasp classification, and multiclass tasks were not attempted.

6. Conclusion

Our work demonstrates the efficacy of Combined Algorithm Selection and Hyperparameter Optimisation as the central aspect of a reliable, fair methodology for the design of biosignal gesture classification systems, and our approach's ability to outperform traditional classifier design processes in important cases. To reduce the issues of bias and insufficient rigour often observed in the field, we recommend CASH as a means of increasing transparency of model design, leading to more fair and methodologically sound comparisons. We also demonstrate an end-to-end development pipeline which goes beyond domain standards of good data handling practice to prevent data leakage, and by holding-out a population subset for verification allows the benefits of CASH to be leveraged without reporting inflated, over-optimised results which fail to generalise.

We recognise that there are avenues for our methodology to be refined further. While it reduces the risk of bias in biosignal classifier design it may still be susceptible to such bias if not applied carefully. The joint algorithm-hyperparameter search space must itself be well defined with a sufficient scope, and not manipulated to favour practitioners' preferred modelling choices. It is for this reason we recommend candidate algorithms and hyperparameter ranges should be transparently detailed and justified. The wide-ranging and often conflicting recommendations of the literature may allow even highly suboptimal configurations to be argued for however; future extensions of our methods may seek to devise more prescriptive guidelines for constructing such a modelling search space. We also note that the Tree-Structured Parzen Estimator, being a greedy optimisation algorithm, can miss very narrow local minima if attempting to optimise over a particularly complex space. There could be benefit to exploring the impact of modifying the total optimisation budget, the size of the initial set of randomly sampled configurations, and the number of candidates sampled in each iteration, on not only the resultant accuracy of a CASH-derived biosignal classifier but the time taken for the optimisation to converge and the stability of its path.

Nevertheless, our use of the CASH-based methodology allowed us to conduct a more extensive comparison of biosignal fusion strategies than has been done before in the domain, in the context of a multi-participant dataset more complex and less trivially separable than those frequently seen in gesture classification research. We achieved a multi-gesture classification accuracy of 73% in wholly subject-independent EMG-EEG fusion, and our subject-independent Unimodal EEG accuracy at 49% was notably competitive with those reached for simpler problems on subject-specific bases by prior EEG studies.

Our novel Hierarchical fusion strategy reached 86% accuracy in a subject-specific multiclass (three grasps & rest) paradigm, significantly outperforming a state-of-the-art Deep Learning EMG architecture and performing equivalently to a "domain-standard" Unimodal EMG model developed with the same CASH-based procedure. This new approach also proved capable of greater accuracy than all but one of a diverse range of competing fusion architectures, further evidencing its capability and motivating its consideration in future research on the topic. There would also be merit in investigating the Hierarchical architecture's effectiveness in other domains beyond that of biosignal classification, and exploring further the distinctions in behaviour between it and conventional Decision-Level fusion. It could even in principle be extended to combine more data sources in a hierarchy with a greater number of "ranks"; this could be useful for embedding priority in systems which fuse many input modalities.

Though much research is conducted on biosignal-based gesture classification, methodological limitations and a sparsity of robust like-for-like comparisons between competing modelling choices means the evidence base for designing such systems is low. Our work not only offers a thorough, unbiased comparison of a range of fusion strategies for a highly application-relevant task, but also proposes a sound, transparent, and robust methodology for end-to-end development. By incorporating Automated Machine Learning alongside a range of measures extending best practice for data handling and classifier design & comparison we present a framework with which we hope to set new standards for biosignal fusion research going forwards.

Acknowledgments

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.

References

[1] Kathryn Ziegler-Graham, Ellen J MacKenzie, Patti L Ephraim, Thomas G Travison, and Ron Brookmeyer. Estimating the prevalence of limb loss in the united states: 2005 to 2050. *Arch. Phys. Med. Rehabil.*, 89(3):422–429, March 2008.

[2] NHS England. Nhs offers life-changing bionic arms to all amputees, November 2022.

[3] Open Bionics. *Hero Arm User Guide - Version 100583_01_0*, February 2019.

[4] Sophie M. Wurth and Levi 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*, 11(1):91, May 2014.

[5] Todd A. Kuiken, Laura A. Miller, Kristi Turner, and Levi 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*, 4:1–8, 2016.

[6] 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*, 4(2):R1, January 2007.

[7] Robert Whelan and Hugh Garavan. When optimism hurts: Inflated predictions in psychiatric neuroimaging. *Biological Psychiatry*, 75(9):746–748, 2014. Mechanisms of Aging and Cognition.

[8] Ren Li, Jared S. Johansen, Hamad Ahmed, Thomas V. Ilyevsky, Ronnie B. Wilbur, Hari M. Bharadwaj, and Jeffrey Mark Siskind. The perils and pitfalls of block design for eeg classification experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):316–333, 2021.

[9] Mahan Hosseini, Michael Powell, John Collins, Chloe Callahan-Flintoft, William Jones, Howard Bowman, and Brad Wyble. I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. *Neuroscience & Biobehavioral Reviews*, 119:456–467, December 2020.

[10] Han Yuan and Bin He. Brain–computer interfaces using sensorimotor rhythms: Current state and future perspectives. *IEEE Transactions on Biomedical Engineering*, 61(5):1425–1435, 2014.

[11] 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*, 15(3), April 2018.

[12] Ilja Kuzborskij, Arjan Gijsberts, and Barbara 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*, pages 4931–4937, 2012.

[13] Manfredo Atzori, Arjan Gijsberts, Claudio Castellini, Barbara Caputo, Anne-Gabrielle Mittaz Hager, Simone Elsig, Giorgio Giatsidis, Franco Bassetto, and Henning Müller. Electromyography data for non-invasive naturally-controlled robotic hand prostheses. *Scientific Data*, 1(1):140053, December 2014.

[14] Ali Ameri, Ernest N. Kamavuako, Erik J. Scheme, Kevin B. Englehart, and Philip A. Parker. Support vector regression for improved real-time, simultaneous myoelectric control. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(6):1198–1209, 2014.

[15] Claudio Castellini, Angelo Emanuele Fiorilla, and Giulio Sandini. Multi-subject/daily-life activity emg-based control of mechanical hands. *Journal of NeuroEngineering and Rehabilitation*, 6(1):41, November 2009.

[16] Radhika Menon, Gaetano Di Caterina, Heba Lakany, Lykourgos Petropoulakis, Bernard A. Conway, and John 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*, 25(10):1832–1842, 2017.

[17] Erik Scheme and Kevin Englehart. Electromyogram pattern recognition for control of powered upper-limb prostheses: state of the art and challenges for clinical use. *J. Rehabil. Res. Dev.*, 48(6):643–659, 2011.

[18] Maged S. Al-Quraishi, Irraivan Elamvazuthi, Tong Boon Tang, Muhammad Al-Qurishi, S. Parasuraman, and Alberto Borboni. Multimodal fusion approach based on eeg and emg signals for lower limb movement recognition. *IEEE Sensors Journal*, 21(24):27640–27650, 2021.

[19] 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*, 11(4):889–899, 2017.

[20] Zhiyuan Lu, Xiang Chen, Qiang Li, Xu Zhang, and Ping Zhou. A hand gesture recognition framework and wearable gesture-based interaction prototype for mobile devices. *IEEE Transactions on Human-Machine Systems*, 44(2):293–299, 2014.

[21] Jordan J. Bird, Michael Pritchard, Antonio Fratini, Anikó Ekárt, and Diego 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*, 6(2):3498–3504, February 2021.

[22] Marco E. Benalcázar, Cristhian Motoche, Jonathan A. Zea, Andrés G. Jaramillo, Carlos E. Anchundia, Patricio Zambrano, Marco Segura, Freddy Benalcázar Palacios, and María Pérez. Real-time hand gesture recognition using the myo armband and muscle activity detection. In *2017 IEEE Second Ecuador Technical Chapters Meeting (ETCM)*, pages 1–6, 2017.

[23] Kang Soo Kim, Heung Ho Choi, Chang Soo Moon, and Chi Woong 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*, 11(3):740–745, 2011.

[24] Xun Chen and Z. Jane Wang. Pattern recognition of number gestures based on a wireless surface emg system. *Biomedical Signal Processing and Control*, 8(2):184–192, 2013.

[25] David Steyrl, Reinhold Scherer, Josef Faller, and Gernot R. Müller-Putz. Random forests in non-invasive sensorimotor rhythm brain-computer interfaces: a practical and convenient non-linear classifier. *Biomedical Engineering / Biomedizinische Technik*, 61(1):77–86, 2016.

[26] Xinyi Yong and Carlo Menon. Eeg classification of different imaginary movements within the same limb. *PLOS ONE*, 10(4):1–24, 04 2015.

[27] Mojgan Tavakolan, Zack Frehlick, Xinyi Yong, and Carlo Menon. Classifying three imaginary states of the same

upper extremity using time-domain features. *PLOS ONE*, 12(3):1–18, 03 2017.

[28] Ke Liao, Ran Xiao, Jania Gonzalez, and Lei Ding. Decoding individual finger movements from one hand using human eeg signals. *PLOS ONE*, 9(1):1–12, 01 2014.

[29] Xiangxin Li, Oluwarotimi Williams Samuel, Xu Zhang, Hui Wang, Peng Fang, and Guanglin Li. A motion-classification strategy based on semg-eeg signal combination for upper-limb amputees. *Journal of NeuroEngineering and Rehabilitation*, 14(1):2, January 2017.

[30] Heba Ibrahim Aly, Sherin Youssef, and Cherine Fathy. Hybrid brain computer interface for movement control of upper limb prostheses. In *2018 International Conference on Biomedical Engineering and Applications (ICBEA)*, pages 1–6, 2018.

[31] Arjan Gijsberts, Manfredo Atzori, Claudio Castellini, Henning Müller, and Barbara Caputo. Movement error rate for evaluation of machine learning methods for semg-based hand movement classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(4):735–744, 2014.

[32] Michael Pritchard, Abraham Itzhak Weinberg, John A R Williams, Felipe Campelo, Harry Goldingay, and Diego R Faria. Dynamic fusion of electromyographic and electroencephalographic data towards use in robotic prosthesis control. *Journal of Physics: Conference Series*, 1828(1):012056, February 2021.

[33] Patrick Ofner, Andreas Schwarz, Joana Pereira, and Gernot R. Müller-Putz. Upper limb movements can be decoded from the time-domain of low-frequency eeg. *PLOS ONE*, 12(8):1–24, 08 2017.

[34] Ji-Hoon Jeong, Jeong-Hyun Cho, Kyung-Hwan Shim, Byoung-Hee Kwon, Byeong-Hoo Lee, Do-Yeun Lee, Dae-Hyeok Lee, and Seong-Whan Lee. Multimodal signal dataset for 11 intuitive movement tasks from single upper extremity during multiple recording sessions. *Gigascience*, 9(10), October 2020.

[35] Alois Schlögl, Felix Lee, Horst Bischof, and Gert Pfurtscheller. Characterization of four-class motor imagery eeg data for the bci-competition 2005. *Journal of Neural Engineering*, 2(4):L14, August 2005.

[36] Wei Li, Ping Shi, and Hongliu Yu. Gesture recognition using surface electromyography and deep learning for prostheses hand: State-of-the-art, challenges, and future. *Frontiers in Neuroscience*, 15, 2021.

[37] Angkoon Phinyomark and Erik Scheme. Emg pattern recognition in the era of big data and deep learning. *Big Data and Cognitive Computing*, 2(3), August 2018.

[38] Christos Dolopikos, Michael Pritchard, Jordan J. Bird, and Diego R. Faria. Electromyography signal-based gesture recognition for human-machine interaction in real-time through model calibration. In Kohei Arai, editor, *Advances in Information and Communication. Future of Information and Communication Conference (FICC) 2021*, volume 1364 of *Advances in Intelligent Systems and Computing*, pages 898–914, Cham, April 2021. Springer International Publishing.

[39] D. Garrett, D.A. Peterson, C.W. Anderson, and M.H. Thaut. Comparison of linear, nonlinear, and feature selection methods for eeg signal classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):141–144, 2003.

[40] Marta Gandolla, Simona Ferrante, Giancarlo Ferrigno, Davide Baldassini, Franco Molteni, Eleonora Guanziroli, Michele Cotti Cottini, Carlo Seneci, and Alessandra Pedrocchi. Artificial neural network emg classifier for functional hand grasp movements prediction. *Journal of International Medical Research*, 45(6):1831–1847, September 2017.

[41] Moncef Garouani, Adeel Ahmad, Mourad Bouneffa, Mohamed Hamlich, Gregory Bourguin, and Arnaud Lewandowski. Using meta-learning for automated algorithms selection and configuration: an experimental framework for industrial big data. *Journal of Big Data*, 9(1):57, April 2022.

[42] Thilina Dulantha Lalitharatne, Kenbu Teramoto, Yoshiaki Hayashi, and Kazuo 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*, 4(2):147–154, 2013.

[43] E. Rocon, J.A. Gallego, L. Barrios, A.R. Victoria, J. Ibánez, D. Farina, F. Negro, J.L. Dideriksen, S. Conforto, T. D'Alessio, G. Severini, J.M. Belda-Lois, L.Z. Popovic, G Grimaldi, M. Manto, and J.L. Pons. Multimodal bci-mediated fes suppression of pathological tremor. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pages 3337–3340, 2010.

[44] Yuhuan Du, Xiaodong Zhang, Yang Wang, and Tong Mu. Design on exoskeleton robot intellisense system based on multi-dimensional information fusion. In *2012 IEEE International Conference on Mechatronics and Automation*, pages 2435–2439, 2012.

[45] Andrea Sarasola-Sanz, Nerea Irastorza-Landa, Eduardo López-Larraz, Carlos Bibián, Florian Helmhold, Doris Broetz, Niels Birbaumer, and Ander 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)*, pages 895–900, 2017.

[46] Abid Hossain Khan, Imran Nawaz 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*, 8(3):941–948, December 2015.

[47] Neha Hooda, Ratan Das, and Neelesh Kumar. Fusion of eeg and emg signals for classification of unilateral foot movements. *Biomedical Signal Processing and Control*, 60:101990, July 2020.

[48] Ozan Ozdenizci, Sezen Yagmur Gunay, Fernando Quivira, and Deniz Erdogmug. Hierarchical graphical models for context-aware hybrid brain-machine interfaces. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2018:1964–1967, July 2018.

[49] Isuru Ruhunage, Chamika Janith Perera, Kalinga Nisal, Janaka Subodha, and Thilina Dulantha 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)*, pages 3169–3174, October 2017.

[50] Jinhua Zhang, Baozeng Wang, Cheng Zhang, Yanqing Xiao, and Michael Yu Wang. An eeg/emg/eog-based multimodal human-machine interface to real-time control of a soft robot hand. *Frontiers in Neurorobotics*, 13:7, 2019.

[51] Kazuo Kiguchi and Yoshiaki 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*, pages 2711–2716, 2012.

[52] K. Förster, Andrea Biasiucci, Ricardo Chavarriaga, José del R. Millán, Daniel Roggen, and Gerhard Tröster. On the use of brain decoded signals for online user adaptive gesture recognition systems. In *International Conference on Pervasive Computing*, pages 427–444, Berlin, Heidelberg, 2010.

[53] Maciej Pawłowski, Anna Wróblewska, and Sylwia Sysko-Romańczuk. Effective techniques for multimodal data fusion: A comparative analysis. *Sensors (Basel)*,

23(5):2381, February 2023.

[54] Jacob Tryon, Evan Friedman, and Ana Luisa Trejos. Performance evaluation of eeg/emg fusion methods for motion classification. In *16th IEEE International Conference on Rehabilitation Robotics (ICORR)*, pages 971–976, Toronto, Canada, June 2019. IEEE.

[55] Robert Leeb, Hesam Sagha, Ricardo Chavarriaga, and José del R. Millán. Multimodal fusion of muscle and brain signals for a hybrid-bci. In *32nd Annual International Conference of the IEEE EMBS*, pages 4343–4346, Buenos Aires, Argentina, August 2010. IEEE.

[56] Jacob Tryon and Ana Luisa Trejos. Classification of task weight during dynamic motion using eeg–emg fusion. *IEEE Sensors Journal*, 21(4):5012–5021, 2021.

[57] Zihao Wang and Ravi Suppiah. Upper limb movement recognition utilising eeg and emg signals for rehabilitative robotics. In Kohei Arai, editor, *Advances in Information and Communication*, pages 676–695, Cham, 2023. Springer Nature Switzerland.

[58] Valer Jurcak, Daisuke Tsuzuki, and Ippeita Dan. 10/20, 10/10, and 10/5 systems revisited: Their validity as relative head-surface-based positioning systems. *NeuroImage*, 34(4):1600–1611, 2007.

[59] Andreas Schwarz, Patrick Ofner, Joana Pereira, Andreea Ioana Sburlea, and Gernot R Müller-Putz. Decoding natural reach-and-grasp actions from human EEG. *Journal of Neural Engineering*, 15(1):016005, December 2017.

[60] Iñaki Iturrate, Robert Leeb, Ricardo Chavarriaga, and Joseé del R Millán. Decoding of two hand grasping types from eeg. In *6th International Brain-Computer Interface Meeting*. Verlag der Technischen Universität Graz, May 2016.

[61] Jerrin Thomas Panachakel, Nandagopal Netrakanti Vinayak, Maanvi Nunna, A. G. Ramakrishnan, and Kanishka Sharma. An improved eeg acquisition protocol facilitates localized neural activation, March 2020.

[62] George H. Klem, Hans Otto Lüders, Herbert H. Jasper, and Christian 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)*, 52:3–6, 1999.

[63] Laurens R Krol. Eeg 10-10 system with additional information, November 2020. CC0 1.0.

[64] The Mathworks, Inc., Natick, Massachusetts, USA. *MATLAB version 9.8 (R2020a)*, 2020.

[65] Ji-Hoon Jeong, Jeong-Hyun Cho, Kyung-Hwan Shim, Byoung-Hee Kwon, Byeong-Hoo Lee, Do-Yeun Lee, Dae-Hyeok Lee, and Seong-Whan Lee. Supporting data for "multimodal signal dataset for 11 intuitive movement tasks from single upper extremity during multiple recording sessions", 2020.

[66] Tzyy-Ping Jung, Scott Makeig, Colin Humphries, Te-Won Lee, Martin J. McKeown, Vicente Iragui, and Terrence J. Sejnowski. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37(2):163–178, March 2000.

[67] Kuan-Ju Huang, Jui-Chieh Liao, Wei-Yeh Shih, Chih-Wei Feng, Jui-Chung Chang, Chia-Ching Chou, and Wai-Chi Fang. A real-time processing flow for ica based eeg acquisition system with eye blink artifact elimination. In *SiPS 2013 Proceedings*, pages 237–240, 2013.

[68] Joseph W Matiko, Stephen Beeby, and John Tudor. Real time eye blink noise removal from EEG signals using morphological component analysis. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2013:13–16, 2013.

[69] Jordan Bird, Luis Manso, Eduardo Ribeiro, Aniko Ekart, and Diego 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, September 2018.

[70] Jhonatan Kobylarz, Jordan J. Bird, Diego R. Faria, Eduardo Parente Ribeiro, and Anikó 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*, March 2020.

[71] Chris Thornton, Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. Auto-weka: Combined selection and hyperparameter optimization of classification algorithms, 2013.

[72] James Bergstra, Daniel Yamins, and David Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 115–123, Atlanta, Georgia, USA, June 2013. PMLR.

[73] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[74] Shuhei Watanabe. Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance, 2023.

[75] The MathWorks, Inc. *Support Vector Machine Classification — MATLAB & Simulink*, 2023. Accessed: October 2023.

[76] Stephan Johann Lehmler, Muhammad Saif ur Rehman, Glasmachers Tobias, and Ioannis Iossifidis. Deep transfer learning compared to subject-specific models for semg decoders. *Journal of Neural Engineering*, 19(5):056039, October 2022.

[77] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, August 2023.

[78] Gernot Müller-Putz, Reinhold Scherer, Clemens Brunner, Robert Leeb, and Gert Pfurtscheller. Better than random? a closer look on bci results. *International Journal of Bioelectromagnetism*, 10:52–55, 01 2008.

[79] Susanna Yu. Gordleeva, Sergey A. Lobov, Nikita A. Grigorev, Andrey O. Savosenkov, Maxim O. Shamshin, Maxim V. Lukoyanov, Maxim A. Khoruzhko, and Victor B. Kazantsev. Real-time eeg–emg human–machine interface-based control system for a lower-limb exoskeleton. *IEEE Access*, 8:84070–84081, 2020.

[80] Siamac Fazli, Florin Popescu, Márton Danóczy, Benjamin Blankertz, Klaus-Robert Müller, and Cristian Grozea. Subject-independent mental state classification in single trials. *Neural Networks*, 22(9):1305–1312, 2009. Brain-Machine Interface.

[81] 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)*, pages 33–36. IEEE, 2019.

[82] Robert Leeb, Hesam Sagha, Ricardo Chavarriaga, and José del R Millán. A hybrid brain–computer interface based on the fusion of electroencephalographic and electromyographic activities. *Journal of Neural Engineering*, 8(2):025011, March 2011.

[83] Stefano Tortora, Luca Tonin, Carmelo Chisari, Silvestro Micera, Emanuele Menegatti, and Fiorenzo Artoni. Hybrid human-machine interface for gait decoding

through bayesian fusion of eeg and emg classifiers. *Frontiers in Neurorobotics*, Volume 14 - 2020, 2020.

[84] W. Wang, A. De. Stefano, and R. Allen. A simulation model of the surface emg signal for analysis of muscle activity during the gait cycle. *Computers in Biology and Medicine*, 36(6):601–618, 2006.

[85] A. F. Ruiz-Olaya, C.F. Blanco-Diaz, C.D. Guerrero-Mendez, T.F. Bastos-Filho, and S. Jaramillo-Isaza. Enhancing classification of grasping tasks using hybrid eeg-semg features. In Jefferson Luiz Brum Marques, Cesar Ramos Rodrigues, Daniela Ota Hisayasu Suzuki, José Marino Neto, and Renato García Ojeda, editors, *IX Latin American Congress on Biomedical Engineering and XXVIII Brazilian Congress on Biomedical Engineering*, pages 182–191, Cham, 2024. Springer Nature Switzerland.