ORIGINAL RESEARCH



Thumbs up, thumbs down: non-verbal human-robot interaction through real-time EMG classification via inductive and supervised transductive transfer learning

Jhonatan Kobylarz¹ · Jordan J. Bird² · Diego R. Faria² · Eduardo Parente Ribeiro¹ · Anikó Ekárt²

Received: 11 October 2019 / Accepted: 27 February 2020 © The Author(s) 2020

Abstract

In this study, we present a transfer learning method for gesture classification via an inductive and supervised transductive approach with an electromyographic dataset gathered via the Myo armband. A ternary gesture classification problem is presented by states of 'thumbs up', 'thumbs down', and 'relax' in order to communicate in the affirmative or negative in a non-verbal fashion to a machine. Of the nine statistical learning paradigms benchmarked over 10-fold cross validation (with three methods of feature selection), an ensemble of Random Forest and Support Vector Machine through voting achieves the best score of 91.74% with a rule-based feature selection method. When new subjects are considered, this machine learning approach fails to generalise new data, and thus the processes of Inductive and Supervised Transductive Transfer Learning are introduced with a short calibration exercise (15 s). Failure of generalisation shows that 5 s of data per-class is the strongest for classification (versus one through seven seconds) with only an accuracy of 55%, but when a short 5 s per class calibration task is introduced via the suggested transfer method, a Random Forest can then classify unseen data from the calibrated subject at an accuracy of around 97%, outperforming the 83% accuracy boasted by the proprietary Myo system. Finally, a preliminary application is presented through social interaction with a humanoid Pepper robot, where the use of our approach and a most-common-class metaclassifier achieves 100% accuracy for all trials of a '20 Questions' game.

 $\textbf{Keywords} \ \ Gesture \ classification \cdot Human-robot \ interaction \cdot Electromyography \cdot Machine \ learning \cdot Transfer \ learning \cdot Inductive \ transfer \ learning \cdot Myo \ armband \cdot Pepper \ robot$

Jhonatan Kobylarz and Jordan J. Bird are co-first authors.

✓ Jordan J. Bird birdj1@aston.ac.uk

> Jhonatan Kobylarz jhonatankobylarz@gmail.com

Diego R. Faria d.faria@aston.ac.uk

Eduardo Parente Ribeiro edu@eletrica.ufpr.br

Published online: 07 March 2020

Anikó Ekárt a.ekart@aston.ac.uk

- Department of Electrical Engineering, Federal University of Parana, Curitiba, Brazil
- School of Engineering and Applied Science, Aston University, Birmingham, UK

1 Introduction

Within a social context, the current state of Human-Robot Interaction is arguably most often concerned with the domain of verbal, spoken communication. That is, the transcription of spoken language to text, and further Natural Language Processing (NLP) in order to extract meaning; this framework is oftentimes multi-modally combined with other data, such as the tone of voice, which too carries useful information. With this in mind, a recent National GP Survey carried out in the United Kingdom found that 125,000 adults and 20,000 children had the ability to converse in British Sign Language (BSL) (Ipsos 2016), and of those surveyed, 15,000 people reported it as their primary language. With those statistics in mind, this shows that those 15,000 people only have the ability to directly converse with approximately 0.22% of the UK population. This argues for the importance of non-verbal communication, such as through gesture.



To answer in the affirmative, negative, or to not answer at all are three very important responses when it comes to meaningful conversation, especially in a goal-based scenario. In this study, a ternary classification experiment is performed towards the domain of non-verbal communication with robots; the electromyographic signals produced when performing a thumbs up, thumbs down, and resting state with either the left or right arms are considered, and statistical classification techniques are benchmarked in terms of validation, generalisation to new data, and transfer learning to better generalise to new data in order to increase reliability to within the realms of classical speech recognition. That is, to reach interchangeable accuracies between the two domains and thus enable those who do not have the ability of speech to effectively communicate with machines.

The main contributions of this work are as follows:

- An original dataset is collected from five subjects for three-class gesture classification.¹ A ternary classification problem is thus presented; thumbs up, thumbs down, and relaxed.
- A feature extraction process retrieved from previous work is used to extract features from electromyographic waves, the process prior to this has only been explored in electroencephalography (EEG) and in this work is adapted for electromyographic gesture classification.²
- Multiple feature selection algorithms and statistical/ ensemble classifiers are benchmarked in order to derive a best statistical classifier for the ground truth data.
- Multiple best-performing models attempt to predict new and unseen data towards the exploration of generalisation, which ultimately fails. Findings during this experiment show that 15 s (5 s per class) performs considerably better than 3, 6, 9, 12, 18, and 21 s of data. Model generalisation only slightly outperforms random guessing.
- Failure of generalisation is then remedied through the suggestion of a calibration framework via inductive and supervised transductive transfer learning. Inspired by the findings of the experiment described in the previous point, models are then able to reach extremely high classification ability on further unseen data presented post-calibration. Findings show that although a confidence-weighted Vote of Random Forest and Support Vector Machine performed better on the original, full dataset, the Random Forest alone outperforms this method for calibration and classification of unseen data (97% vs. 95.7% respectively).

Available online, https://github.com/jordan-bird/eeg-feature-generation/ Last Accessed: 25/02/2020.





Fig. 1 The MYO EMG Armband (Thalmic Labs)

 Finally, a real-time application of the work is preliminary explored. Social interaction is enabled with a humanoid robot (Softbank's Pepper) in the form of a game, through gestural interaction and subsequent EMG classification of the gestures in order to answer yes/no questions while playing 20 Questions.

In order to present the aforementioned findings in a structured manner, exploration and results are presented in chronological order, since a failed generalisation experiment is then remedied with the aid of the findings through limitation. The remainder of this article is structured as follows: firstly, important state-of-the-art work within the field of gesture recognition and electromyography are presented in Sect. 2, along with important background information regarding Feature Selection and Machine Learning techniques explored within this study. Section 3 then outlines the processes followed towards dataset acquisition, feature extraction, experimental methodologies, as well as important hyperparameters and hardware information required for replicability of the experiments. Results and discussion are then presented in Sect. 4, followed by a preliminary application of the findings in Sect. 5. Finally, possible future works are discussed in Sect. 6 with regards to the limitations of this work and a final conclusion of the findings presented.

2 Background

In this section, state-of-the-art literature in electromyographic gesture classification are considered. Additionally, a short overview of the statistical techniques are given.

¹ Available online, https://www.kaggle.com/birdy654/emg-gesture-classification-thumbs-up-and-down/ Last Accessed: 25/02/2020.

2.1 EMG gesture classification and calibration

The MYO Armband, as shown in Fig. 1, is a device comprised of 8 electrodes ergonomically designed to read electromyographic data from on and around the arm by an embedded chip within the device. Researchers have noted the MYO's quality as well as its ease of availability to both researchers and consumers (Rawat et al. 2016), and is thus recognised as having great potential in EMG-signal based experiments. In this section, notable state-of-the-art literature is presented within which the MYO armband has successfully provided EMG data for experimentation.

The Myo Armband was found to be accurate enough to control a robotic arm with 6 Degrees of Freedom (DoF) with similar speed and precision to the controlling subject's movements (Widodo et al. 2018). In this work, researchers found an effective method of classification through the training of a novel Convolutional Neural Network (CNN) architecture at a mean accuracy of 97.81%. A related study, also performing classification with CNN successfully classified 9 physical movements from 9 subjects at a mean accuracy of 94.18% (Mendez et al. 2017); it must be noted, that in this work, the model was not tested for generalisation ability. This has shown to be important in this study, since the strongest method for classification of the dataset was ultimately weaker than another model when it came to transfer of ability to unseen data.

Researchers have noted that gesture classification with Myo has real-world application and benefits (Kaur et al. 2016), showing that physiotherapy patients often exhibit much higher levels of satisfaction when interfacing via EMG and receiving digital feedback (Sathiyanarayanan and Rajan 2016). Likewise in the medical field, Myo has shown to be competitively effective with far more expensive methods of non-invasive electromyography in the rehabilitation of amputation patients (Abduo and Galster 2015), and following this, much work has explored the application of gesture classification for the control of a robotic hand (Ganiev et al. 2016; Tatarian et al. 2018). Since the armband is worn on the lower arm, the goal of the robotic hand is to be teleoperated by non-amputees and likewise to be operated by amputation patients in place of the amputated hand. Work from the United States has also shown that EMG classification is useful for exercises designed to strengthen the glenohumeral muscles towards rehabilitation in Baseball (Townsend et al. 1991).

Recently, work in Brazilian Sign Language classification via the Myo armband found high classification ability of results through a Support Vector Machine on a 20-class problem (Abreu et al. 2016). Researchers noted 'substantial limitations' in the form of realtime classification application and generalisation, with models performing sub-par on unseen data. For example, letters A, T, and U had worthless

classification abilities of 4%, 4%, and 5% respectively. This work aims to set out to both train models, and also explore methods of generalisation to new, unseen data in real-time. The Myo armband's proprietary framework, through a short exercise, boasts up to an 83% real-time classification ability. Although seemingly relatively high, this margin of error that is a statistical risk in 17% of cases prevents the Myo from being deployed in situations where such a rate of error is unacceptable and considered critical. Though it may be considered acceptable to possibly miscommunicate 17% of the time in sign language dictation, this error rate would unacceptable, for example, for the control of a drone where a physical risk is presented. Thus, the goal of many works is to improve this ability. In terms of real-time classification, there are limited works, and many of them suggest a system of calibration during short exercises (similarly to the Myo framework) in order to fine-tune a Machine Learning model. In (Benalcázar et al. 2017), authors suggested a solution of a ten second exercise (5, 2 s activities) in order to gain 89.5% real-time classification accuracy. This was performed through K-Nearest Neighbour (KNN) and the Dynamic Time Warping (DTW) algorithms. EMG has also been applied to other bodily surfaces for classification, for example, to the face in order to classify emotional response based on muscular activity(Tan et al. 2012).

In 2017, researchers found that certain early layers of a CNN could be applied to unseen subjects when further training is performed on subsequent layers of the network on new subject data (Côté-Allard et al. 2019). This study showed not only that a physical task ('pick up the cube') could be completed on average in less time than with joystick hardware, but that the transfer learning process allowed for 97.81% classification accuracy of the EMG data produced by the movements of 17 individual subjects. It must be noted, that this deep learning technique (along with some aforementioned) is heavy in terms of resource usage (Shi et al. 2016), and thus, in this study, classical statistical methods are explored which require far fewer resources to train and classify data. This paradigm is followed in order to allow autonomous machines (usually operating a single CPU) the ability to perform training, calibration, and classification without the need for comparatively more expensive GPU capabilities, or access to a cloud system with similar means.

Discrimination of affirmative and negative responses in the form of *thumbs up* and *thumbs down* was shown to be possible in a related study (Huang et al. 2015b), within which the two actions were part of a larger eight-class dataset which achieved 87.6% on average for four individual subjects. Linear Discriminant Analysis (LDA) was used to classify features generated by a sliding window of 200ms in size with a 50ms overlap technique similar to that followed in this work; the features were mean absolute value, waveform length, zero crossing and sign slope change for the



EMG itself and mean value and standard deviation observed by the accelerometer. In (Huang et al. 2015a), researchers followed a similar process of the classification of minute thumb movements when using an Android mobile phone. Results showed that accuracies of 89.2% and 82.9% are achieved for a subject holding a phone and not holding a phone respectively when 2 s of EMG data is classified with a K-Nearest Neighbour (KNN) classification algorithm. A more recent work explored the preliminary applications of image enhancement to surface electromyographs showing their potential to improve the classification of muscle characteristics(ul Islam et al. 2019).

Calibration in the related works, where performed, are through the process of *Inductive Transfer Learning* (ITL) and Supervised Transductive Transfer Learning (STTL). According to (Pan and Yang 2009) and (Arnold et al. 2007), ITL is the process satisfied when the source domain labels are available as well as the target labels, this is leveraged in the calibration stage, in which the gesture being performed is known. STTL is the process in which the source domain labels are available but the target is not, this is the validation stage in this study, when a calibrated model is benchmarked on further unknown data during application of a calibrated model. Transfer learning is the process of knowledge transfer from one learned task to another (Zhuang et al. 2019), in this study, it is shown to be difficult to generalise a model to new subjects and thus application of a model to new data is considered a task to be solved by transfer learning; transfer learning often shows strong results in the application of gesture classification in related state-of-the-art works (Liu et al. 2010; Goussies et al. 2014; Costante et al. 2014; Yang et al. 2018; Demir et al. 2019).

Numerous open issues arising from this literature review can be observed, and this is experiment seeks to address said issues:

- 1. Often, only one method of Machine Learning is applied, and thus different statistical techniques are rarely compared as benchmarks on the same dataset.
 - In this work, many statistical techniques of feature selection and machine learning are applied in order to explore the abilities of each in EMG classification.
- Very little exploration of generalisation has been performed, researchers usually opt to present classification ability of a dataset and there is a distinct lack of exploration when unseen subjects are concerned. This is important for real-world application.
 - In this work, models attempt to classify data gathered from new subjects and experience failure. This is further remedied by the suggestion of a short calibration task, in which the generalisation then succeeds through the process of inductive transfer learning and transductive transfer learning.

- 3. When applications are presented, there is often a lack of exposition in the real-time results for that application.
 - In this work, where real-world, real-time applications are concerned, classification abilities are given at each step where required. This is important for exploration of ability, and thus, exploration of areas for future work.

2.2 Selected feature selection algorithms

Feature selection is the process of reducing a dataset's dimensionality in order to reduce the complexities of machine learning algorithms while still effectively maintaining effective classification ability (Dash and Liu 1997; Guyon and Elisseeff 2003). Thus, the main goal of feature selection is to disregard worthless attributes that have no bearing on class, and if stricter rules are in place, to also disregard those with very little classification ability which is not considered worth their contribution to model complexity. In this section, the chosen feature selection algorithms employed within this study are described.³

Information Gain is the scoring of an attribute's classification ability in regards to comparing a change in entropy when said attribute is used for classification (Kullback and Leibler 1951). The entropy measured for a specific attribute is given as:

$$E(s) = -\sum_{k} p_k \times log(p_k). \tag{1}$$

That is, the Entropy E is the sum of the probability mass function of the value p times by its negative logarithm. The change in entropy (*Information Gain*) when different attributes are observed for classification thus allow for scoring of ability.

Symmetrical Uncertainty is a method of dimensionality reduction by comparison of two attributes in regards to classification entropy and Information Gain given a pair (Gel'Fand and Yaglom 1959; Piao et al. 2019). This allows for comparative scores to be applied to attributes within the vector. For attributes *X* and *Y*, Symmetrical Uncertainty is given as:

$$SymmU(X,Y) = 2 \times \frac{(IG(X|Y))}{E(X) + E(Y)},$$
(2)

where Entropy *E* and Information Gain *IG* are calculated as previously described.



³ For the One Rule Feature Selection process, please see Sect. 2.3.

2.3 Selected machine learning algorithms

A Machine Learning (ML) algorithm, in general terms, is the process of building an analytical or predictive model with inspiration from labelled (known) data (Bishop 2006; Michie et al. 1994). The process of *classification* is to develop rules to label unseen (validation) data based on seen (training) data. This section details the general background of the learning models selected in this study. A wide range of models are chosen in order to explore the differing abilities of multiple statistical techniques.

One Rule classification is an extremely simplistic process in order to generate a best-fit ruleset based on one attribute. A single attribute is identified as the best for classification, and rules are generated based upon it, that is, effective splits to disseminate the data object (eg. for an attribute a, IF a > 10 THEN Class = Y, IF a > 10 THEN Class = Z)

Decision Trees are tree-like branched data structures, where at each node, a conditional control statement is used to provide a rule based on attribute values where an end node without connections represents a class (Pal 2005). Classification follows a process of cascading the data objects from start to end of the tree and their predicted class is given as the one reached. Fitness of a tree layout is given as the entropy within the end nodes and their classified instances⁴. A Random Decision Tree (RDT) with parameter *K* will select K random attributes at each node and develop splitting rules based on them (Prasad et al. 2006). The model is simple since no pruning is performed and thus an overfitted tree is produced to classify all input data points, therefore cross-validation is used to create an average of the best performing random trees, or with a testing set of unseen data.

Support Vector Machines (SVM) classify data points by optimising a data-dimensional hyperplane to most aptly separate them, and then classifying based on the distance vector measured from the hyperplane (Cortes and Vapnik 1995). Optimisation follows the goal of the average margins between points and the separator to be at the maximum possible value. Generation of an SVM is performed through Sequential Minimal Optimisation (SMO), a high-performing algorithm to generate and implement an SVM classifier (Platt 1998). To perform this, the large optimisation problem is broken down into smaller sub-problems, these can then be solved linearly. For multipliers a, reduced constraints are given as:

$$0 \le a_1, a_2 \le C, y_1, a1 + y_2, a_2 = k,$$
(3)

where there are data classes y and k are the negative of the sum over the remaining terms of the equality constraint.

Naive Bayes is a probabilistic model given by *Bayes' The-orem* which aims to find the posterior probability for a number of different hypotheses, then select the hypothesis with the highest probability. The posterior probability is given by:

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)} \tag{4}$$

Where P(h|d) is the probability of hypothesis h given the data d, P(d|h) is the probability of data d given that the hypothesis h is true. P(h) is the probability of hypothesis h being true and P(d) = P(d|h)P(h) is the probability of the data. The algorithm assumes each probability value as conditionally independent for a given target (ergo naive), calculated as P(d1|h)P(d2|h) and so on. Despite its simplicity, related work has shown its effectiveness in some complex problems (Wood et al. 2019), showing that Naive Bayes classification achieves 96% in negative predicted value with the Wisconsin breast cancer data set.

Bayesian Networks are graphic probabilistic models that satisfy the local Markov property, and are used for computation of probability. This network is a Directed Acyclic Graph (DAG) in which each edge is a conditional dependency, and each node corresponds to a unique random variable and is conditionally independent of its non-descendants. Thus the probability of an arbitrary event $N = (n_1, ..., n_k)$ can be computed as $P(X) = \prod_{i=1}^k P(X_i|X_i, ..., X_{i-1})$.

Logistic Regression is a process of symmetric statistics where a numerical value is linked to a probability of event occurring, ie. the number of driving lessons to predict pass or fail (Walker and Duncan 1967). In a two class problem within a dataset containing i number of attributes and β model parameters, the log odds l is derived via $l = \beta_0 + \sum_{i=0}^{x} \beta_i + x_i$ and the odds of an outcome are shown through $o = b^{\beta_0 + \sum_{i=0}^{x} \beta_i + x_i}$ which can be used to predict an outcome based on previous observation.

Voting allows for multiple trained models to act as an ensemble through democratic or weighted voting. Each model will vote on their outcome (prediction) by way of methods such as simply applying a single vote or voting by weight of probability experienced from training and validation. The final decision of the model is the class receiving the highest number of votes or weighted votes, and is given as the outcome prediction. A Random Decision Forest (RDF) is an example of a voting model. A specified number of n RDTs are generated on randomly selected subsets of the input data (Bootstrap Aggregation), and produce an overall prediction by presenting the majority vote (Ho 1995).



⁴ For details on Information Gain, please see Sect. 2.2.

3 Method

In this section, the methodology of the experiments in this study are described. Initially, data is acquired prior to the generation of a full dataset through feature extraction. Machine Learning paradigms are then benchmarked on the dataset, before the exploration of real-time classification of unseen data.

The experiments performed in this study were executed on a AMD FX-8520 eight-core processor with a clock speed of 3.8 GHz. In terms of software, the algorithms are executed via the Weka API (implemented in Java). The machine learning algorithms are validated through a process of k-fold cross validation, where k is set to 10 folds. The voting process is to vote by average probabilities of the models, since two models are considered and thus a democratic voting process would result in a tie should the two models disagree.

3.1 Data acquisition

The Myo Armband records EMG data at a rate of 200 Hz via 8 dry sensors worn on the arm, and it also has a 9-axis Inertial Measurement Unit (IMU) performing at a sample rate of 50 Hz. For this study, data acquisition is performed with 5 subjects, which are three males and two females (aged 22–40). For model generalisation, 4 more subjects ware taken into account, of which two of them are new subjects and two are performing the movements again. The gestures performed were, thumbs up, thumbs down, and resting (a neutral gesture in which the subject is asked to rest their hand). For training, 60 s of forearm muscle activity data was recorded for each arm (two minutes, per subject, per gesture). In the case of benchmark data, the muscle waves were recorded in intervals of 1–7 s each.

3.2 Feature extraction

In this study, time series are considered through a sliding window technique in order to generate statistics and thus extract features or attributes from the 8-dimensional data. Related work in biological signal processing argues for the need of feature extraction prior to data mining(Mendoza-Palechor et al. 2019; Seo et al. 2019) This is performed due to wave data being complex and temporal in nature and thus single points are difficult to classify (since they depend on both past and future events). The feature extraction process in this study is based on previous works with electroencephalographic signals (Bird et al. 2018, 2019)⁵, which have been

Available online, https://github.com/jordan-bird/eeg-feature-generation/ Last Accessed: 25/02/2020



noted to bare some similarity to EMG signals (Grosse et al. 2002). A general overview of the process is as follows:

Initially, a sliding window of length 1s at an overlap of 0.5s divides the data into short wave segments.

For each time window, the following is performed:

- Considering the full time window, the following statistics are measured:
 - The mean and standard deviation of the wave.
 - The skewness and kurtosis of each signal (Zwillinger and Kokoska 2000).
 - The maximum and minimum values.
 - The sample variances of each signal, plus the sample covariances of all pairs of waves (Montgomery and Runger 2010).
 - The eigenvalues of the covariance matrix (Strang 2006).
 - The upper triangular elements of the matrix logarithm of the covariance matrix (Chiu et al. 1996).
 - The magnitude of the frequency components of each signal by Fast Fourier Transform (FFT) (Van Loan 1992).
 - The frequency values of the ten most energetic components of the FFT, for each signal.
- Considering the two 0.5s windows produced due to offset (overlap of two 1s windows resulting in 0.5s windows):
 - The change in both the sample means and in the sample standard deviations between the 1st and 2nd 0.5s windows.
 - The change in both the maximum and minimum values between the first and second 0.5s windows.
- Considering the two 0.25 s quarter windows produced due to offset:
 - The mean of each each quarter-window.
 - All paired differences of means between the quarterwindows.
 - The maximum (minimum) values of each quarterwindow, plus all paired differences of maximum (minimum) values between the quarter-windows.

Change in attributes is also treated as a feature, in which each window is passed the previous extracted value vector sans maximum, mean, and minimum values of quarter windows. The first window does not receive this vector since no window preceded it.

Feature extraction thus produced a dataset of 2040 numerical attributes from the 8 electrodes, of which there are 159 megabytes of data produced from the five subjects. A minor original contribution is also presented in the form

Table 1 A comparison of the three attribute selection experiments

Method	No. attributes selected	Max score	Min score	
One rule	2000	64.39	30.51	
Information gain	1898	0.62	0.004	
Symmetrical uncertainty	1898	0.32	0.003	

Note that Scoring methods are Unique and thus not Comparable between the Three

4 Results

In this section, the preliminary results from the experiments are given. Firstly, the chosen machine learning techniques are benchmarked in order to select the most promising method for the problem presented in this study. Secondly, generalisation of models to unseen data is benchmarked before a similar experiment is performed within which transfer learning is leveraged to enable generalisation of models to new data through calibration to a subject.

Table 2 10-fold classification ability of both single and ensemble methods on the datasets

Dataset	Single Model Accuracy (%)				Ensemble Model Accuracy (%)				
	OneR	RT	SVM	NB	BN	LR	RF	Vote (best two)	Vote (best three)
OneR	61.33	74.03	87.14	64.32	69.9	60.76	91.30	91.74	74.67
InfoGain	61.49	75.39	87.11	64.13	69.9	61.45	91.7	91.30	75.13
Symmetrical uncertainty	61.48	74.37	87.11	64.13	69.9	61.55	91.36	91.4	75.16
Whole dataset	61.33	74.09	87.14	64.32	69.9	x	91.3	91.71	74.72

Voting does not include random tree due to the inclusion of random forest

of the application of these features to EMG data, since they have only been shown to be effective thus far in EEG signal processing.

3.3 Machine learning and benchmarking towards real-time classification

Following data acquisition and feature extraction, multiple ML models are benchmarked in order to compare their classification abilities on the EMG data. The particularly strong models are then considered for generalisation and real-time classification.

In this work, two approaches towards real-time classification are explored. Small datasets are recorded sequentially from four subjects, varying from lengths of 1 s, from 1 to 7 s per class. These then constitute seven datasets per person {3,6..21}.

Initially, the best four models observed by the previous experiments are used to classify these datasets in order to derive the ideal amount of time that an action must be observed before the most accurate classification can be performed.

Following this, a method of calibration through transfer learning is also explored. The result from the aforementioned experiment (the ideal amount of observation time) is taken forward and, for each person, appended to the full dataset recorded for the classification experiments. Each of the chosen ML techniques are then retrained and used to classify further unseen data from said subject.

4.1 Feature selection and machine learning

Table 1 shows the results of attribute selection performed on the full dataset of 2040 numerical attributes. One Rule feature selection found that the majority of attributes held strong One Rule classification ability, as is often expected (Ali and Smith 2006). Information Gain and Symmetrical Uncertainty produced slightly smaller datasets both of 1898, and it must be noted that the two datasets are comprised of differing attributes.

In Table 2, the full matrix of benchmarking results are presented. An interesting pattern occurs throughout all datasets, both reduced and full; an SVM is always the best single classifier, scoring between 87.11 and 87.14%. Additionally, a voting ensemble of Random Forest and SVM always produce the strongest classifiers at results of between 91.3 and 91.74%. Interestingly, the One Rule dataset is slightly less complex than the full dataset but produces a slightly superior result. The Information Gain and Symmetrical Uncertainty datasets are far less complex, and yet are only behind the best One Rule score by 0.44% and 0.34% respectively. Logistic Regression on the whole dataset fails due to its high resource requirements, but is observed to be viable on the datasets that have been reduced.



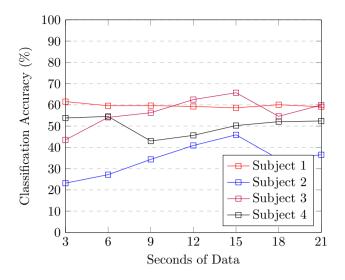


Fig. 2 Benchmarking of vote (Best Two) model generalisation ability for unseen data segments per subject, in which generalisation has failed due to low classification accuracies

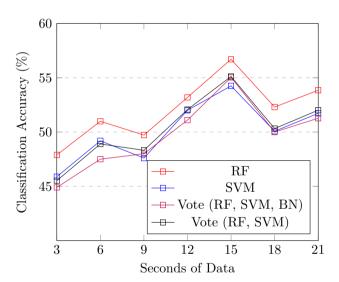
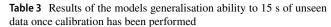


Fig. 3 Initial pre-calibration mean generalisation ability of models on unseen data from four subjects in a three-class scenario. Time is given for total data observed Equally for three classes. Generalisation has failed

4.2 Benchmarking requirements for realtime classification

In this section, very short segments of unseen data are collected from four subjects in order to attempt to apply the previously generated models to new data. That is, to experiment on the generalisation ability or lack thereof of the models on the 5-subject dataset. Generalisation initially fails, but with the least catastrophic model in mind,



Model	Generalisation Ability (%)
Single models	
OneR	63
RT	91.86
SVM	94
NB	53.35
BN	66.05
LR	90.1
Ensemble models	
RF	97
Vote (RF, SVM)	95.7
Vote (RF, SVM, BN)	87.8

Table 4 Confusion matrix for the random forest once calibrated by the subject for 15 s when used to predict unseen data

Prediction			Ground Truth
Rest	Up	Down	
300	0	1	Rest
0	324	1	Up
0	19	376	Down

Counts have been compiled from all subjects. Class imbalance occurs in real-time due to bluetooth sampling rate

leading the focus to calibration of a 'user' in ideally short amounts of time via transfer learning.

When the best model from Table 2 is used, the ensemble vote of average probabilities between a Random Forest and SVM fails in being able to classify unseen data. Observe Fig. 2, in which 15 s of unseen data performs, on average, in excess of any other amount of data, but yet still only reaches a mean classification ability of 55.12% (which is unacceptable for a ternary classification problem).

In Fig. 3, the mean classification ability of other highly performing models from the previous experiment are given when unseen data are attemptedly classified. Likewise to the Vote model observed in Fig. 2, generalisation has failed for all models. Two interesting insights emerge from the failed experiments; firstly, 15 s of data (5 s per class) most often leads to the best limited generalisation as opposed to both shorter and longer experiments. Furthermore, the ability of the Random Forest can be seen to exceed all of the other three methods, suggesting that it is superior (albeit limited) when generalisation is considered.

As previously described, calibration is attempted through a short experiment. Due to the findings aforementioned, 15 s of known data (that is, requested during 'setup') is collected.



These labelled data are then added to the training data, in order to expand knowledge at a personal level. Once this is performed, and the models are trained, they are then benchmarked with a further unseen dataset of 15 s of data, again, 5 s per class. No further training of models are performed, and they simply attempt to classify this unseen data. Table 3 shows the abilities of all previously benchmarked models once the short calibration process is followed, with far greater success than observed in the previous failed experiments, where those previous were benchmarked. As was conjectured from said failed experiments, the Random Forest showed to be the most successful calibration experiment for generalisation towards a new subject. The error matrix for the best model is given in Table 4. The most difficult task was the prediction of 'thumbs down', which, when a subject had a particularly smaller arm would sometimes be classified as a resting state. Observed errors are extremely low, and thus future work to explore this is suggested in Sect. 6.

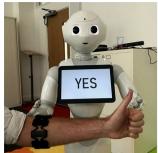
5 Applications in human-robot interaction

In this section, an application of the framework is presented in a HRI context. The Random Forest model observed to be the best model for generalisation in Sect. 4.2 is calibrated for 5 s per class in regards to the benchmark results, then enabling the subject to interact non-verbally with machines via EMG gesture classification. Note that only preliminary benchmarks are presented, and Sect. 6 details potential future work in this regard, that is, these preliminary activities are not considered the main contributions of this work which were presented in Sect. 4.

5.1 20 Questions with a humanoid robot opponent

20Q, or 20 Questions, is a digital game developed by Robin Burgener based on the 20th Century American parlor game of the same name and rules; it is a situational puzzle. Through Burgener's algorithm, computer opponents play via the dissemination and subsequent strategy presented by an Artificial Neural Network (Burgener 2006, 2003). In the game between man and machine, the player thinks of an entity and the opponent is able to ask 20 yes/no questions. Through elimination of potential answers, the opponent is free to guess the entity that the player is thinking of. If the opponent cannot guess the entity by the end of the 20 questions, then the player has won.

In this application the 20 Questions game is played with a humanoid robot, Softbank Robotics' *Pepper*. Initially, the subject is calibrated with 15 s of data (5 per class) added to the full dataset, due to the findings in this work. Following this, for every round of questioning, the robot will listen to 5 s of data from the player, perform feature generation,



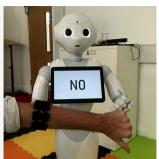


Fig. 4 Softbank Robotics' pepper robot playing 20 Questions with a human through real-time EMG signal classification

Table 5 Statistics from two games played by two subjects each

Subject	Yes avg. confidence (accuracy) (%)	No avg. confidence (accuracy) (%)	Avg. confidence (accuracy) (%)	EMG predictions confidence (accuracy) (%)
1	96.9	96.5	96.7	100
2	97	97	97	100

Average Accuracy is given as per-data-object, correct EMG predictions are given as overall decisions

and finally will consider the most commonly predicted class from all data objects produced in order to derive the player's answer. This process can be seen in Fig. 4 in which feedback is given during data classification. Two players each play two games each with the robot. Thus, the model used is a calibrated Random Forest (through inductive and transductive transfer learning) and a simple meta-approach of the most common class.

As can be seen in Table 5, results from the four games are given as average accuracy on a per-data-object basis, but the results of the game operate on the final column, *EMG Predictions Accuracy*, this is the measure of correct predictions of thumb states by the most common prediction of all data objects generated over the course of data collection and feature generation. As can be observed, the high accuracies of per-object classification contribute towards perfect classification of player answers, all of which were at 100%.

6 Future work and conclusion

In the calibration experiment, error rates were found to be extremely low. Accuracy measurements exceeded the original benchmarks and thus further experimentation is required to explore this. Calibration was performed for a limited group of four subjects, further experimentation should explore a more general affect when a larger group of participants are considered.



Towards the end of this work, preliminary benchmarks are presented for potential application of the inductive and supervised transductive transfer learning calibration process. The 20 Questions game with a Pepper Robot was possible with 15 s of calibration data and 5 s of answering time per question, and predictions were at 100% for two subjects in two different experimental runs. Further would could both explore more subjects as well as attempt to perform this task with shorter answering time, ie. a deeper exploration into how much data is enough for a confident prediction. For example, rather than the simplistic most common class Random Forest approach, a more complex system of meta-classification could prove more useful as the pattern of error may be useful also for prediction; if this were so, then it stands to reason that confident classification could be enabled sooner than the 5 s mark. Additionally, when a a best-case paradigm is confirmed, the method could then be compared to other sensory techniques such as image/video classification for gesture recognition. Furthermore, should said method be also viable, then a multi-modal approach could also be explored in order to fuse both visual and EMG data.

This article shows that the proposed transfer learning system is viable to be applied to the ternary classification problem presented. Future work could explore the robustness of this approach to problems of additional classes and gestures in order to compare how results are affected when more problems are introduced.

To finally conclude, this experiment firstly found that a voting ensemble was a strong performer for classification of gesture but failed to generalise to new data. With the inductive and transductive transfer learning calibration approach, the best model for generalisation of new data was a Random Forest technique which achieved very high accuracy. After gathering data from a subject for only 5 s, the model could confidently classify the gesture at 100% accuracy through the most common class Random Forest classifier. Since very high accuracies were achieved by the transfer learning approach in this work when compared to the state-of-theart related works and the proprietary MYO system, future applications could be enabled with our approach towards a much higher resolution of input than is currently available with the MYO system.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

- Abduo M, Galster M (2015) Myo gesture control armband for medical applications. https://www.semanticscholar.org/paper/Myo-Gesture-Control-Armband-for-Medical-Abduo-Galster/3b5ed355b09beecb7b2b6bbd23fead44b50374c6
- Abreu JG, Teixeira JM, Figueiredo LS, Teichrieb V (2016) Evaluating sign language recognition using the myo armband. In: 2016 XVIII Symposium on Virtual and Augmented Reality (SVR), IEEE, pp 64–70
- Ali S, Smith KA (2006) On learning algorithm selection for classification. Applied Soft Computing 6(2):119–138
- Arnold A, Nallapati R, Cohen WW (2007) A comparative study of methods for transductive transfer learning. In: ICDM Workshops, pp 77–82
- Benalcázar ME, Motoche C, Zea JA, Jaramillo AG, Anchundia CE, Zambrano P, Segura M, Palacios FB, Pérez M (2017) Real-time hand gesture recognition using the myo armband and muscle activity detection. In: 2017 IEEE Second Ecuador Technical Chapters Meeting (ETCM), IEEE, pp 1–6
- Bird JJ, Manso LJ, Ribeiro EP, Ekárt A, Faria DR (2018) A study on mental state classification using eeg-based brain-machine interface. In: 2018 International Conference on Intelligent Systems (IS), IEEE, pp 795–800
- Bird JJ, Faria DR, Manso LJ, Ekárt A, Buckingham CD (2019) A deep evolutionary approach to bioinspired classifier optimisation for brain-machine interaction. Complexity. https://doi.org/10.1155/2019/4316548
- Bishop CM (2006) Pattern recognition and machine learning. Springer, Berlin
- Burgener R (2003) 20q twenty questions
- Burgener R (2006) Artificial neural network guessing method and game. US Patent App. 11/102,105
- Chiu TY, Leonard T, Tsui KW (1996) The matrix-logarithmic covariance model. Journal of the American Statistical Association 91(433):198–210
- Cortes C, Vapnik V (1995) Support-vector networks. Machine learning 20(3):273–297
- Costante G, Galieni V, Yan Y, Fravolini ML, Ricci E, Valigi P (2014) Exploiting transfer learning for personalized view invariant gesture recognition. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 1250–1254
- Côté-Allard U, Fall CL, Drouin A, Campeau-Lecours A, Gosselin C, Glette K, Laviolette F, Gosselin B (2019) Deep learning for electromyographic hand gesture signal classification using transfer learning. IEEE Transactions on Neural Systems and Rehabilitation Engineering 27(4):760–771
- Dash M, Liu H (1997) Feature selection for classification. Intelligent data analysis 1(1–4):131–156
- Demir F, Bajaj V, Ince MC, Taran S, Şengür A (2019) Surface emg signals and deep transfer learning-based physical action classification. Neural Computing and Applications 31(12):8455–8462
- Ganiev A, Shin HS, Lee KH (2016) Study on virtual control of a robotic arm via a myo armband for the selfmanipulation of a hand amputee. Int J Appl Eng Res 11(2):775–782
- Gel'Fand I, Yaglom A (1959) Calculation of amount of information about a random function contained in another such function. Eleven Papers on Analysis, Probability and Topology 12:199
- Goussies NA, Ubalde S, Mejail M (2014) Transfer learning decision forests for gesture recognition. The Journal of Machine Learning Research 15(1):3667–3690
- Grosse P, Cassidy M, Brown P (2002) Eeg-emg, meg-emg and emgemg frequency analysis: physiological principles and clinical applications. Clinical Neurophysiology 113(10):1523–1531



- Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
- Ho TK (1995) Random decision forests. In: Proceedings of the third international conference on document analysis and recognition, IEEE, vol 1, pp 278–282
- Huang D, Zhang X, Saponas TS, Fogarty J, Gollakota S (2015a) Leveraging dual-observable input for fine-grained thumb interaction using forearm emg. In: Proceedings of the 28th annual ACM symposium on user interface software and technology, ACM, pp 523–528
- Huang Y, Guo W, Liu J, He J, Xia H, Sheng X, Wang H, Feng X, Shull PB (2015b) Preliminary testing of a hand gesture recognition wristband based on emg and inertial sensor fusion. In: International conference on intelligent robotics and applications, Springer, pp 359–367
- Ipsos M (2016) Gp patient survey-national summary report. NHS England, London
- ul Islam I, Ullah K, Afaq M, Chaudary MH, Hanif MK (2019) Spatiotemporal semg image enhancement and motor unit action potential (muap) detection: algorithms and their analysis. J Ambient Intell Humaniz Comput 10(10):3809–3819
- Kaur M, Singh S, Shaw D (2016) Advancements in soft computing methods for emg classification. Int J Biomed Eng Technol 20(3):253–271
- Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22(1):79–86
- Liu J, Yu K, Zhang Y, Huang Y (2010) Training conditional random fields using transfer learning for gesture recognition. In: 2010 IEEE international conference on data mining, IEEE, pp 314–323
- Mendez I, Hansen BW, Grabow CM, Smedegaard EJL, Skogberg NB, Uth XJ, Bruhn A, Geng B, Kamavuako EN (2017) Evaluation of the myo armband for the classification of hand motions. In: 2017 International conference on rehabilitation robotics (ICORR), IEEE, pp 1211–1214
- Mendoza-Palechor F, Menezes ML, Sant'Anna A, Ortiz-Barrios M, Samara A, Galway L (2019) Affective recognition from eeg signals: an integrated data-mining approach. J Ambient Intell Humaniz Comput 10(10):3955–3974
- Michie D, Spiegelhalter DJ, Taylor C et al (1994) Machine learning. Neural Stat Classif 13:1–298
- Montgomery DC, Runger GC (2010) Applied statistics and probability for engineers. Wiley, New York
- Pal M (2005) Random forest classifier for remote sensing classification. Int J Remote Sens 26(1):217–222
- Pan SJ, Yang Q (2009) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359
- Piao M, Piao Y, Lee JY (2019) Symmetrical uncertainty-based feature subset generation and ensemble learning for electricity customer classification. Symmetry 11(4):498
- Platt J (1998) Sequential minimal optimization: a fast algorithm for training support vector machines. https://www.microsoft.com/ en-us/research/wp-content/uploads/2016/02/tr-98-14.pdf
- Prasad AM, Iverson LR, Liaw A (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9(2):181–199

- Rawat S, Vats S, Kumar P (2016) Evaluating and exploring the myo armband. In: 2016 International conference system modeling and advancement in research trends (SMART), IEEE, pp 115–120
- Sathiyanarayanan M, Rajan S (2016) Myo armband for physiotherapy healthcare: a case study using gesture recognition application. In: 2016 8th International conference on communication systems and networks (COMSNETS), IEEE, pp 1–6
- Seo J, Laine TH, Sohn KA (2019) Machine learning approaches for boredom classification using eeg. J Ambient Intell Humaniz Comput 10(10):3831–3846
- Shi S, Wang Q, Xu P, Chu X (2016) Benchmarking state-of-the-art deep learning software tools. In: 2016 7th international conference on cloud computing and big data (CCBD), IEEE, pp 99–104
- Strang G (2006) Linear algebra and its applications. Brooks Cole, London
- Tan JW, Walter S, Scheck A, Hrabal D, Hoffmann H, Kessler H, Traue HC (2012) Repeatability of facial electromyography (emg) activity over corrugator supercilii and zygomaticus major on differentiating various emotions. J Ambient Intell Humaniz Comput 3(1):3–10
- Tatarian K, Couceiro MS, Ribeiro EP, Faria DR (2018) Stepping-stones to transhumanism: An emg-controlled low-cost prosthetic hand for academia. In: 2018 International conference on intelligent systems (IS), IEEE, pp 807–812
- Townsend H, Jobe FW, Pink M, Perry J (1991) Electromyographic analysis of the glenohumeral muscles during a baseball rehabilitation program. Am J Sports Med 19(3):264–272
- Van Loan C (1992) Computational frameworks for the fast Fourier transform. SIAM 10:10
- Walker SH, Duncan DB (1967) Estimation of the probability of an event as a function of several independent variables. Biometrika 54(1–2):167–179
- Widodo MS, Zikky M, Nurindiyani AK (2018) Guide gesture application of hand exercises for post-stroke rehabilitation using myo armband. In: 2018 international electronics symposium on knowledge creation and intelligent computing (IES-KCIC), IEEE, pp 120–124
- Wood A, Shpilrain V, Najarian K, Kahrobaei D (2019) Private naive bayes classification of personal biomedical data: application in cancer data analysis. Comput Biol Med 105:144–150
- Yang S, Lee S, Byun Y (2018) Gesture recognition for home automation using transfer learning. In: 2018 International conference on intelligent informatics and biomedical sciences (ICIIBMS), IEEE, vol 3, pp 136–138
- Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, Xiong H, He Q (2019) A comprehensive survey on transfer learning. arXiv:191102685
- Zwillinger D, Kokoska S (2000) CRC standard probability and statistics tables and formulae. Chapman and Hall, Boca Raton

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

