Chapter 12
Channel and Class Dependent Time-Series Embedding Using Partial Mutual Information Improves Sensorimotor Rhythm Based Brain-Computer Interfaces

Damien Coyle

Abstract. Mutual information has been found to be a suitable measure of dependence among variables for input variable selection. For time-series prediction mutual information can quantify the average amount of information contained in the lagged measurements of a time series. Information quantities can be used for selecting the optimal time lag, τ, and embedding dimension, Δ, to optimize prediction accuracy. Times series modeling and prediction through traditional and computational intelligence techniques such as fuzzy and recurrent neural networks (FNNs and RNNs) have been promoted for EEG preprocessing and feature extraction to maximize signal separability to improve the performance of brain-computer interface (BCI) systems. This work shows that spatially disparate EEG channels have different optimal time embedding parameters which change and evolve depending on the class of motor imagery (movement imagination) being processed. To determine the optimal time embedding for each EEG channel (time-series) for each class an approach based on the estimation of partial mutual information (PMI) is employed. The PMI selected embedding parameters are used to embed the time series for each channel and class before self-organizing fuzzy neural network (SOFNN) based predictors are specialization to predict channel and class specific data in a prediction based signal processing framework, referred to as neural-time-series-prediction-preprocessing (NTSPP). The results of eighteen subjects show that subject-, channel- and class-specific optimal time embedding parameter selection using PMI improves the NTSPP framework, increasing time-series separability. The chapter also shows how a range of traditional signal processing tools can be combined with multiple computational intelligence based approaches including the SOFNN and practical swarm optimization (PSO) to develop a more autonomous parameter optimization setup and ultimately a novel and more accurate BCI.

Damien Coyle
Intelligent Systems Research Centre,
University of Ulster, Derry, BT48 7JL, UK

W. Pedrycz & S.-M. Chen (Eds.): Time Series Analysis, Model. & Applications, ISRL 47, pp. 249–278.
DOI: 10.1007/978-3-642-33439-9_12

1 Introduction

The human brain contains approximately 10^{11} neurons interconnected through over 100 trillion synapses. Each neuron, containing many different compartments made up of many different chemicals and neurotransmitters, emits tiny electrical pulses every millisecond. The electroencephalogram (EEG), recorded from the scalp surface, is a measure of the aggregate activity of many post-synaptic-potentials (PSPs) of these neurons and includes information from many different brain sources along with background noise from other non-neural signals. EEG is therefore inherently complex and non-stationary, rendering it very difficult to associate a particular EEG time series pattern or dynamic with a specific mental state or thought.

Coupling EEG dynamics to a person's thoughts or intent, expressed in the form of mental imagery, is the objective of non-invasive brain-computer interface (BCI) technology. BCIs enable people to communicate with computers and devices without the need for neuromuscular control or the normal communication pathways and therefore have many potential applications [1]-[3]. BCI has applications in assistive technologies for the physically impaired [4][5], rehabilitation after stroke [7], awareness detection in disorders of consciousness (DoC) [6] and in non-medical applications such as games and entertainment [8]. Voluntarily modulation of sensorimotor rhythms (SMR) forms the basis of non-invasive (EEG-based) motor imagery (MI) BCIs. Planning and execution of hand movement are known to block or desynchronize neuronal activity which is reflected in an EEG bandpower decrease in mu band (8-12Hz). Inhibition of motor behaviour synchronizes neuronal activity [1]. During unilateral hand imagination, the preparatory phase is associated with a contralateral mu and central beta event related desynchronization (ERD) that is preponderant during the whole imagery process [9]-[11]. BCIs utilize a number of self-directed neurophysiological processes including the activation of sensorimotor cortex during motor imagery (MI). However, as outlined, the dynamical and non-stationary patterns in the time series must be dealt with to ensure information can be discriminated and classified precisely so that BCI technology is robust enough to be made available to those who need it most: those who are severely physically impaired due to disease or injury. Maximizing the capacity for computer algorithms to separate noise from source, distinguish between two or more different mental states or one mental task (intentional control (IC) state) from all other possible mental states (no control (NC) state) has been the goal of many BCI focused researchers for the past 20 years. Linear and non-linear approaches to classification have been applied to classifying the EEG signals [12]-[14]. Times series modeling and prediction through traditional and computational intelligence techniques such as fuzzy and recurrent neural networks (FNNs and RNNs) have been promoted for EEG preprocessing and feature extraction to maximize signal separability [15]-[24].

Coyle et al [22][23] have proposed an approach were specific self-organizing FNNs (SOFNN) are trained to specialize in predicting EEG time-series recorded from various electrode channels during different types of motor imagery (left/right

movement imagination). The networks become specialized on the dynamics of each time series and the relative difference in the predictions provided by the networks can produce information about the times series' that are being fed to the networks e.g., if two networks are specialized on two particular time series (left or right motor imagery) and unlabeled time series are fed to both networks, the network that produces the lowest prediction error can be indicative of the times series being processed and thus the information can be used to classify (or label) the unlabeled time series. This idea has been extended to include multiple time series, multiple classes and integrated with a range of other signal processing techniques to aid in the discrimination of sensorimotor based activations for BCI. A critical element in the neural time-series-prediction pre-processing (NTSPP) framework [21]-[24] is predictor (network) specialisation. This can be achieved through network optimization techniques and self-organising systems assuming that there are underlying differences in the time series being processed. Coyle et al [21] have shown in preliminary studies that subject specific time-embedding of the time-series can assist in specializing networks to improve BCI performance but that generally an embedding dimension, Δ=6 and a time lag, τ=1, works well for one-step-ahead EEG time series prediction.

The aim of this chapter is to show that spatially disparate EEG channels have different optimal time embedding parameters which change and evolve depending on the motor imagery or mental task being processed. To determine the optimal time embedding for each EEG channel (time-series) a recently proposed method based on the estimation of partial mutual information (PMI) is employed [25][26]. Mutual information has been found to be a suitable measure of dependence among variables for input variable selection and quantifies the average amount of common information contained in Δ measurements of a time series. Information quantities can be used for selecting the optimal time lag, τ, and embedding dimension, Δ, to optimize prediction accuracy. The PMI selected embedding parameters are used to embed the time series for each channel and class before SOFNN specialization is performed in the NTSPP framework. The results of eighteen subjects show that subject-, channel- and class-specific optimal time embedding parameter selection using PMI improves the NTSPP framework, increasing time-series separability and therefore overall BCI performance.

The following section describes the data used in the chapter to validate the proposed approach. Section 3 includes a description of the methods employed where section 3.1 describes the BCI including the NTSPP approach and other stages of signal processing such as spectral filtering, common spatial patterns, feature extraction and classification. Section 3.2 outlines the partial mutual information based input variable selection (PMIS) approach and the implications of applying this in the NTSPP framework for BCI. A description of how the BCI is setup and parameters are optimized is contained in Section 3.3. A discussion of the results and findings is presented in the remaining sections along with suggested future work for improvements to the proposed methodology.

2 Data Acquisition and Datasets

Data from 18 participants partaking in BCI experiments are used in this work. All datasets were obtained from the fourth international BCI competitions, BCI-IV, [27][28], which include datasets 2A and 2B [29]. Table 1 below provides a summary of the data.

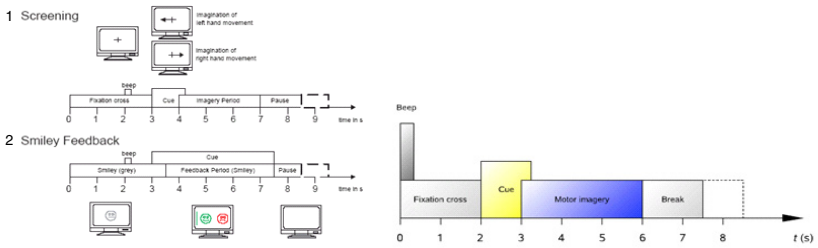
Table 1 Summary of datasets used from the International BCI competition IV

Competition	*Dataset*	*Subjects*	*Labels*	*Trials*	*Classes*	*Channels*
BCI-IV	2B	9	S1-9	720	2	3
BCI-IV	2A	9	S10-18	576	4	22

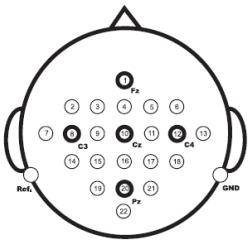
Dataset 2B - This data set consists of EEG data from 9 subjects (S1-S9). Three bipolar recordings (C3, Cz, and C4) were recorded with a sampling frequency of 250 Hz (downsampled to 125Hz in this work). The placement of the three bipolar recordings (large or small distances, more anterior or posterior) were slightly different for each subject (for more details see [29][31]). The electrode position Fz served as EEG ground. The cue-based screening paradigm (cf. Fig 1(a).1) consisted of two classes, namely the motor imagery (MI) of the left hand (class 1) and the right hand (class2). Each subject participated in two screening sessions without feedback recorded on two different days within two weeks. Each session consisted of six runs with ten trials each and two classes of imagery. This resulted in 20 trials per run and 120 trials per session. Data of 120 repetitions of each MI class were available for each person in total. Prior to the first motor imagery training the subject executed and imagined different movements for each body part and selected the one which they could imagine best (e.g., squeezing a ball or pulling a brake). For the three online feedback sessions four runs with smiley feedback were recorded whereby each run consisted of twenty trials for each type of motor imagery (cf. Fig 1(a) for details of the timing paradigm for each trial). Depending on the cue, the subjects were required to move the smiley towards the left or right side by imagining left or right hand movements, respectively. During the feedback period the smiley changed to green when moved in the correct direction, otherwise it became red. The distance of the smiley from the origin was set according to the integrated classification output over the past two seconds (more details can be found in [31]). The classifier output was also mapped to the curvature of the mouth causing the smiley to be happy (corners of the mouth upwards) or sad (cor ners of the mouth downwards). The subject was instructed to keep the smiley on the correct side for as long as possible and therefore to perform the correct MI as long as possible. A more detailed explanation of the dataset and recording paradigm is available [31]. In addition to the EEG channels, the electrooculogram (EOG) was recorded with three monopolar electrodes and this additional data can be used for EOG artifact removal [32] but was not used in this study.

Dataset 2A - This dataset consists of EEG data from 9 subjects (S10-S18). The cue-based BCI paradigm consisted of four different motor imagery tasks, namely

the imagination of movement of the left hand (class 1), right hand (class 2), both feet (class 3), and tongue (class 4) (only left and right hand trials are used in this investigation). Two sessions were recorded on different days for each subject. Each session is comprised of 6 runs separated by short breaks. One run consists of 48 trials (12 for each of the four possible classes), yielding a total of 288 trials per session. The timing scheme of one trial is illustrated in Fig 1(b). The subjects sat in a comfortable armchair in front of a computer screen. No feedback was provided but a cue arrow indicated which motor imagery to perform. The subjects were asked to carry out the motor imagery task according to the cue and timing presented in Fig 1(b). For each subject twenty-two Ag/AgCl electrodes (with inter-electrode distances of 3.5 cm) were used to record the EEG; the montage is shown in Fig 1(c) left. All signals were recorded monopolarly with the left mastoid serving as reference and the right mastoid as ground. The signals were sampled with 250 Hz (downsampled to 125Hz in this work) and bandpass filtered between 0.5 Hz and 100 Hz. EOG channels were also recorded for the subsequent application of artifact processing although this data was not used in this work. A visual inspection of all data sets was carried out by an expert and trials containing artifacts were marked.



(a) (b)



(c)

Fig. 1 (a) Timing scheme of the paradigm for recording dataset 2B; 1) the first two sessions provided training data without feedback, and 2) the last three sessions with smiley feedback. (b) Timing scheme of recording for dataset 2A; (c) electrode montage for recording dataset 2A; For dataset 2B electrodes positions were fine-tuned around positions c3, cz and c4 electrodes used to derive bipolar channels for each subject [31].

To summarize, in this work only electrodes positioned anteriorly and posteriorly to positions C3, Cz and C4 are used to derive 3 bipoloar channels. These channels are located over left, right hemisphere and central sensorimotor areas – areas which are predominantly the most active during motor imagery. In dataset 2A only 2 of the available 4 classes are used (left and right hemisphere). As outlined all data was downsampled to 125 Hz in this work. The data splits (training and testing) were the same as those used for the BCI Competition IV [30]. For dataset 2A, one session (2 classes consisting of 72 trials each) are used for training and the remaining session is used for final testing. For dataset 2B the first two sessions are not used, session 3, the first feedback session, is used for training (160 trials) and feedback sessions 4E and 5E are used for final testing. All parameter selection is conducted on the training data using cross validation as described in section 3.3 and the system setup is tested on the final testing sessions.

3 Methods

3.1 BCI Description

3.1.1 Neural-Time-Series-Prediction-Processing (NTSPP)

NTSPP, introduced in [21], is a framework specifically developed for preprocessing EEG signals associated with motor imagery based BCI systems. NTSPP increases class separability by predictive mapping and filtering the original EEG signals to a higher dimensional space using predictive/regression models specialized (trained) on EEG signals for different brain states i.e., each type of motor imagery. A mixture or combination of neural network-based predictors are trained to predict future samples of EEG signals i.e., predict ahead the state of the EEG. Networks are specialized on each class of signal from each EEG channel. Due to network specialization, prediction for one class of signals differ from the other therefore introducing discriminable characteristics into the predicted signal for each class of signal associated with a particular brain state. Features extracted from the predicted signals are more separable and thus easier to classify.

Consider two EEG times-series, x_i, $i \in \{1,2\}$ drawn from two different signal classes c_i, $i \in \{1,2\}$, respectively, assuming, in general, that the time series have different dynamics in terms of spectral content and signal amplitude but have some similarities. Consider also two prediction NNs, f_1 and f_2, where f_1 is trained to predict the values of x_1 at time $t+\pi$ given values of x_1 up to time t (likewise, f_2 is trained on time series x_2), where π is the number of samples in the prediction horizon. If each network is sufficiently trained to specialize on its respective training data, either x_1 or x_2, using a standard error-based objective function and a standard training algorithm, then each network could be considered an ideal

predictor for the data type on which it was trained[1] i.e., specialized on a particular data type. If each prediction NN is an ideal predictor then each should predict the time-series on which it was trained perfectly, leaving only error residual equiva lent to white or Gaussian noise with zero mean.

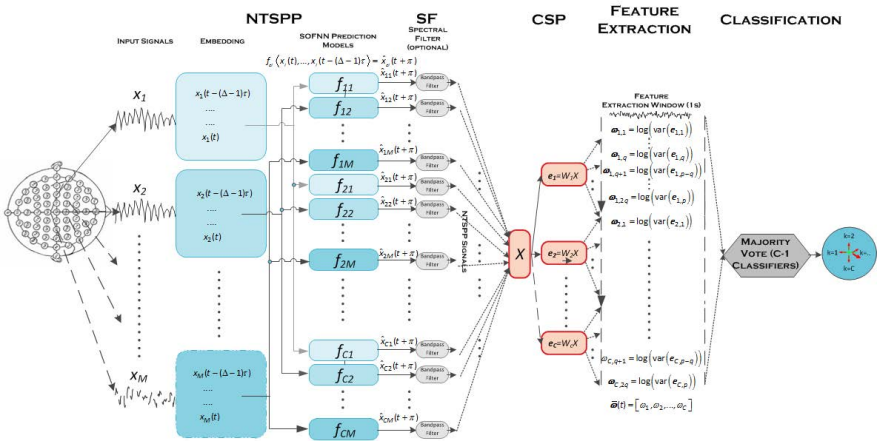


Fig. 2 An illustration of a generic multiclass or multichannel neural-time-series-prediction-preprocessing (NTSPP) framework with spectral filtering, CSP, feature extraction and classification

In such cases the expected value of the mean error residual given predictor f_1 for signal x_1 is $E[x_1–f_1(x_1)]=0$ and the expected power of the error residual, $E[x_1–f_1(x)]^2$, would be low (i.e., in relative terms) whereas, if x_2 is predicted by f_1 then $E[(x_2– f_1(x_2)] \neq 0$ and $E[(x_2–f_1(x_2)]^2$ would be high (i.e., again in relative terms). The opposite would be observed when x_i, $i \in \{1,2\}$, data are predicted by predictor f_2. Based on the above assumptions, a simple set of rules could be used to determine which signal class an unknown signal type, u, belongs too. To classify u one or both of the following rules could be used:-

1. *If* $E[u– f_1(u)] = 0$ & $E[u– f_2(u)] \neq 0$ *then* $u \in C_1$, *otherwise* $u \in C_2$.
2. *If* $E[u– f_1(u)]^2 < E[u– f_2(u)]^2$ *then* $u \in C_1$, *otherwise* $u \in C_2$.

These are simple rules and may only work successfully in cases where the predictors are ideal and specialized sufficiently. Due to the complexity of EEG data and its non-stationary characteristics, and the necessity to specify a NN architecture which approximates universally, predictors trained on EEG data will not consistently be ideal however; when trained on EEG with different dynamics e.g., left

[1] Multilayered feedforward NNs and adaptive-neuro-fuzzy-inference-systems (ANFIS) are considered universal approximators due to having the capacity to approximate any function to any desired degree of accuracy with as few as one hidden layer that has sufficient neurons [33][34].

and right MI, predictor NNs can introduce desirable characteristics in the predicted outputs which render the predicted signals more separable than the original signals and thus aid in determining which brain state produced the unknown signal. This predictive filtering modulates levels of variance in the predicted signals for data types and most importantly manipulates the variances differently for different classes of data. Instead of using only one signal channel, the hypothesis underlying the NTSPP framework is that, if two or more channels are used for each signal class and advanced feature extraction techniques and classifiers are used instead of the simple rules outlined above, additional advantageous information relevant to the differences introduced by the predictors for each class of signal can be extracted to improve overall feature separability thereby improving BCI performance.

In general, the number of time-series available and the number of classes governs the number of specialized predictor networks employed and the resultant number of predicted time-series from which to extract features, such that

$$P = M \times C \tag{1}$$

where P is the number of networks (=no. of predicted time-series), M is the number of EEG channels and C is the number of classes. For prediction,

$$\hat{x}_{ci}(t+\pi) = f_{ci}\left\langle x_i(t),...,x_i(t-(\Delta-1)\tau\right\rangle \tag{2}$$

where t is the current time instant, Δ is the embedding dimension and τ is the time delay, π is the prediction horizon, f_{ci} is the prediction network trained on the i^{th} EEG channel, x_i, i=1,..,M, for class c, c=1,..C, where C is the number of classes and $\hat{x}_{ci}$ is the predicted time series produced for channel i by the predictor for class c and channel i. An illustration of the NTSPP framework is presented in Fig. 2.

Many different predictive approaches can be used for prediction in the NTSPP framework [21][22][24]. In this work the self-organizing fuzzy neural network (SOFNN) is employed [23][36][37]. The SOFNN is a powerful prediction algorithm capable of self-organizing its architecture, adding and pruning neurons as required. New neurons are added to cluster new data that the existing neurons are unable to cluster while old, redundant neurons are pruned ensuring optimal network size, accuracy and training speed (cf. [23] for details of the SOFNN and recent improvements to the SOFNN learning algorithm and its autonomous hyperparameter-free application in BCIs).

Earlier work [21] has shown Δ=6 and τ=1 provide good performance in a two class MI-BCI however this chapter shows how NTSPP can be enhanced by selecting channel- and class-specific embedding parameters using partial mutual information selection as described in section 3.2. Firstly, the other signal processing components of the BCI are described.

3.1.2 Common Spatial Patterns (CSP)

CSP maximizes the ratio of class-conditional variances of EEG sources [38][39]. To utilise CSP, Σ_1 and Σ_2 are the pooled estimates of the covariance matrices for two classes, as follows:

$$\sum_c = \frac{1}{I_c} \sum_{i=1}^{I_c} X_i X_i^t \quad (c \in \{1,2\}) \tag{3}$$

where I_c is the number of trials for class c and X_i is the $M \times N$ matrices containing the i^{th} windowed segment of trial i; N is the window length and M is the number of EEG channels – when CSP is used in conjunction with NTSPP, $M=P$ according to (1). The two covariance matrices, Σ_1 and Σ_2, are simultaneously diagonalized such that the Eigenvalues sum to 1. This is achieved by calculating the generalised eigenvectors W:

$$\sum_1 W = (\sum_1 + \sum_2) WD \tag{4}$$

where the diagonal matrix D contains the Eigenvalues of Σ_1 and the column vectors of W are the filters for the CSP projections. With this projection matrix the decomposition mapping of the windowed trials X is given as

$$E = WX. \tag{5}$$

To generalize CSP to 3 or more classes (multiclass paradigm), spatial filters are produced for each class vs. the remaining classes (one vs. rest approach). If q is the number of filters used then there are $q \times C$ surrogate channels from which to extract features. To illustrate how CSP enhances separability among 4 classes the hypothetical relative variance level of the data in each of the 4 classes are shown in Fig. 3.

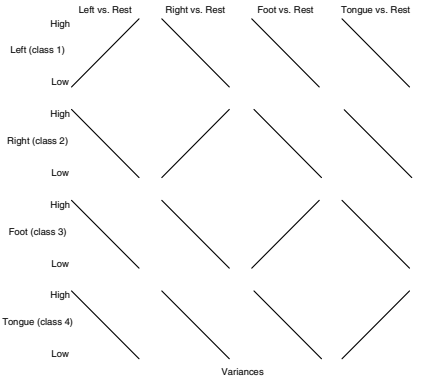
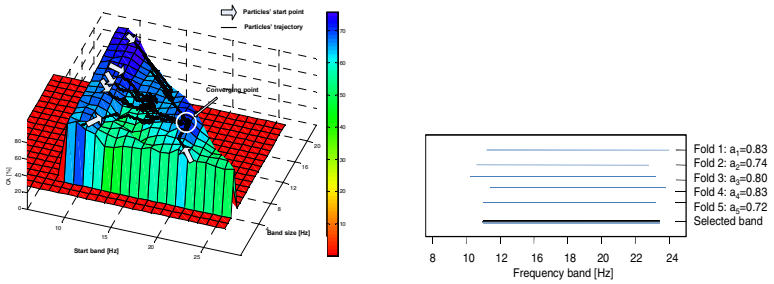


Fig. 3 Hypothetical relative variance level of the CSP transformed surrogate data



(a) (b)

Fig. 4 (a) Frequency band selection using PSO. The graphical representation of particles' motion over progressive generations where classification accuracy(CA) is assessed at teach generation ; (b) Frequency band selection over 5-folds (where a_i is the probability of correct classification in fold *I* and the selected band is the weighted average (weighted by a_i) of the band selected for each fold).

3.1.3 Spectral Filtering (SF)

Prior to the calculation of the spatial filters, *X* can be preprocessed with NTSPP and/or spectrally filtered in specific frequency bands. Optimal frequency bands are selected autonomously in the offline training stage using particle swarm atomization (PSO) [16][40][41] to band pass filter the data before CSP is applied. The search space is every possible band size in the 8 - 28Hz range as shown in Fig. 4(a). These bands encompass the μ and β bands which are altered during sensorimotor processing [42][43] and can be modulated via motor imagery.

3.2 Feature Extraction and Classification

Features are derived from the log-variance of preprocessed/surrogate signals within a two second sliding window:

$$\bar{\omega} = \log(\mathrm{var}(E)) \tag{6}$$

The dimensionality of $\bar{\omega}$ depends on the number of surrogate signals used from *E*. The common practice is to use several (*q* between 2 and 4) eigenvectors from both ends of the eigenvector spectrum, i.e., the columns of *W*. Using NTSPP the dimensionality of *X* can increase significantly. CSP, can be used to reduce the dimensionality therefore the benefits of combining NTSPP with CSP are twofold; 1) increasing separability and 2) maintaining a tractable dimensionality [22].

Linear Discriminant Analysis (LDA) is used to classify the features at the rate of the sampling interval. Linear classifiers are most commonly used for classifying motor imagery in BCI applications. Before describing how the parameters associat ed with these stages of signal processing are optimized in section 3.3, the following section describes the main novelty of this chapter for enhancement of this framework where the embedding parameters are selected using PMIS.

3.3 Partial Mutual Information

The selection of an optimal embedding dimension and its corresponding time lags is often referred to as the input variable selection (IVS) problem. The IVS problem is defined as the task of appropriately selecting a subset of k variables, from the initial candidate set C which comprise the set of all potential inputs to the model (i.e., candidates) [26]. Mutual information has been found to be a suitable measure of dependence among variables for IVS and quantifies the average amount of common information contained in Δ measurements of a time series. Information quantities can be used for selecting the optimal τ and Δ, to optimize prediction accuracy. Evaluation of mutual information and redundancy-based statistics as functions of τ and Δ can further improve insight into dynamics of a system under study.

In essence, two successive measurements of a random variable have no mutual information (in the case of more than two variables mutual information is commonly replaced by the term redundancy) but data based on an underlying rule may have some association; mutual information is proportional to the strength of that association. Utilising only one or two observations of a time series, x, may not provide enough information about a future value of x to make a reliable prediction. Generally, for periodic, quasi-periodic and even chaotic data redundancy tends to rise as each additional measurement of x (i.e., Δ is increased) is involved in the redundancy calculation, at a fixed lag. Mutual information is an arbitrary measure and makes no assumption of the structures of dependence among variables, be they linear or non-linear. It has also been shown to be robust to noise and data transformations.

Although mutual information is a strong candidate for IVS there are number of issues associated with applying the algorithm such as the ability of the selection algorithm to consider the inter-dependencies among variables (redundancy handling) and the lack of appropriate analytical methods to determine when the optimal set has been selected. One method involves the estimation of marginal redundancy, ς, which quantifies the average amount of information contained in the variables $x_{t+(\Delta-1)\tau},...x_{t+\tau}$ about the variable x_t and the quantity is the difference between two successive R calculations ($\varsigma = R_{\Delta+1} - R_{\Delta}$). Depending on the complexity of the data, usually ς increases as Δ is increased. Eventually further increases in Δ provide a lesser increase in ς. Finally, ς becomes approximately constant or begins to decrease. A constant ς indicates that further increases in Δ does not improve the ability of a sequence of measurements to predict the last measurement in the sequence at that value of τ (i.e., there is no advantage in increasing Δ). Another method of estimating the optimum value of Δ can be realised by plotting ς as a function of τ with Δ as a third variable. The relationship between plots of ς versus τ becomes closer as Δ is increased. The optimum Δ is chosen as the smallest Δ for which the plotted relations become relatively close to each other. For more information see [43][44][46][47][48]. In some cases the results from this type of redundancy analysis can be subjective and may not be fully conclusive whilst the calculation can also be time consuming.

Sharma [25] proposed an alternative algorithm and one that overcomes the difficulties in terms of determining the optimal sets of variables with mutual information by using the concept of partial mutual information (PMI). The approach was further assessed and developed by May et al. [26]

3.4 Estimation of Partial Mutual Information

The mutual information calculation stems from Shannon's information theory [49] formulated in (7)

$$I_{Y;X} = \iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{7}$$

where x and y are observations of random variables X and Y, respectively i.e., $y \in Y$ and $x \in X$. Considering Y is an output variable for which there is uncertainty around its observation and is dependent upon the random input variable x then the mutual observation of (x,y) reduces this uncertainty, since knowledge of x allows inferences of the values of y and vice versa. Within a practical context the true functional forms of the pdfs in (7) are typically unknown. In such cases the estimates of the densities are used instead. Substitution of the density estimates into a numerical approximation of the integral in (7) gives

$$I_{Y;X} = \sum_{i=1}^{N_s}\sum_{j=1}^{N_s} p(x_i,y_j)\log_2\frac{p(x_i,y_j)}{p(x_i)p(y_j)} \tag{8}$$

where N_s is the number of bins used for calculating the probability $p(x_i)$ of signal measurement x, occurring in bin x_i and the probability $p(y_j)$ of signal measurement y occurring in bin y_j. $p(x_i,y_j)$ is the joint probability of occurrence of both measurements of the signal. Equation (2) can be generalised to calculate redundancies among variables in a time series, as shown in (3)

$$\begin{aligned} &R(x_t, x_{t+\tau},...,x_{t+(\Delta-1)\tau}) \\ &= \sum^{N_r} p(x_t, x_{t+\tau},...,x_{t+(\Delta-1)\tau})\log_2\frac{p(x_t, x_{t+\tau},...,x_{t+(\Delta-1)\tau})}{p(x_t)p(x_{t+\tau})...p(x_{t+(\Delta-1)\tau})} \end{aligned} \tag{9}$$

where x_t is the measurement of the signal sampled at time t and N_r is the number of phase space routes (i.e., the number of combinations). Equations (8) and (9) can be derived in the probability form or entropy form (H):-

$$R(.) = H(x_t) + ... + H(x_{t+(\Delta-1)\tau}) - H(x_t,...,x_{t+(\Delta-1)\tau}) \tag{10}$$

Full derivations can be found in [45][46][47]. Depending on the number of measurements of the signal and the number of bins, the joint probability, $p(x_t, x_{t+\tau},...,x_{t+(\Delta-1)\tau})$, can encompass a very large number of sequence probabilities. For example, if Δ =5 and N_s=20 then the number of sequence probabilities to

be estimated is $N_r=N_s^D$ =3.2 x 10^6, increasing exponentially as Δ is increased. Estimating redundancies for Δ>*5* can be significantly time consuming.

Mutual information estimation is therefore largely dependent on the technique employed to estimate the marginal and joint pdfs. Non-parametric techniques such as kernel density estimation (KDE) are considered suitably robust and accurate although somewhat computationally intensive compared to alternative approaches such as the histogram approach. Substitution of the density estimates into a numerical approximation of the integral in (7) and (8) gives

$$I_{Y;X} \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i, y_j)\log_2 \frac{f(x_i, y_j)}{f(x_i)f(y_j)} \tag{11}$$

where *f* denotes the estimated density based on a sample of *n* observations of (*x, y*). The Parzen window approach is a simple KDE in which the estimator for *f* is given by

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) \tag{12}$$

where $\hat{f}(x)$ denotes the estimate of the pdf at *x*, $x_i\{\mathrm{i} = 1,\ldots,n\}$ denotes the samples observations of *X*, and K_h is the kernel function where *h* denotes the kernel bandwidth (or, smoothing parameter). A common choice for K_h is the Gaussian kernel

$$K_h = \frac{1}{(\sqrt{2\pi h})^d \sqrt{|\Sigma|}} \exp\left(\frac{-\|x - x_i\|}{2h^2}\right) \tag{13}$$

where *d* denotes the number of dimensions of *X*, $\Sigma = \{\sigma_{ij}\}$ is the sample covariance matrix and $\|x - x_i\|$ is the Mahalanobis distance metric given by

$$\|x - x_i\| = (x - x_i)^T \Sigma^{-1} (x - x_i). \tag{14}$$

The kernel expression in (12) is used with (13) and (14) to produce the kernel estimator as defined below

$$\hat{f}(x) = \frac{1}{n(\sqrt{2\pi h})^d \sqrt{|\Sigma|}} \sum_{i=1}^{n} \exp\left(\frac{-\|x - x_i\|}{2h^2}\right) \tag{15}$$

The performance of the kernel estimator, in terms of accuracy, is dependent more on the choice of bandwidth as opposed to choice of kernel itself [26][50]. The optimal choice of bandwidth depends on the distribution of the data samples. In [25][26][51] the Gaussian reference bandwidth, h_G, for MI estimation is adopted as an efficient choice. The Gaussian reference bandwidth is determined using the following rule proposed by Silvermann [52]

$$h_G = \left(\frac{1}{d+2}\right)^{1/(d+4)} \sigma n^{-1/d+4} \tag{16}$$

where σ is the standard deviation of the data samples. The MI calculation can be easily extended to the multivariate case where the response/output variable *Y* is dependent on multiple input variables. For example, two input variables *X* and *Z*. Given *X* the uncertainty is reduced by a certain amount and the partial mutual information is defined as the further reduction in the uncertainty surrounding *Y* that is gained by the additional mutual observation of *Z*. Partial MI (PMI) is analogous to the partial correlation coefficient, $R`_{ZY.X}$, which quantifies the linear dependence of *Y* on variable *Z* that is not accounted for by the input variable *X*. This is normally calculated by filtering *Y* and *Z* via regression on *X* to obtain some residuals, *u* and *v*, respectively [26]. Pearson's correlation can be used to estimate *X*. PMI can be applied in a similar way to estimate the arbitrary dependence between variables. Using the KDE approach an estimator for the regression of *Y* on *X* is written as

$$\hat{m}_Y(x) = E[y \mid X = x] = \frac{1}{n} \frac{\sum_{i=1}^{n} y_i K_h(x - x_i)}{\sum_{i=1}^{n} K_h(x - x_i)} \tag{17}$$

where $\hat{m}_Y(x)$ is the regression estimator; *n* is the number of observed values (y_i; x_i); K_h is given as in (12) and $E[y \mid X = x]$ denotes the conditional expectation of *y* given an observed *x*. An estimator $\hat{m}_Z(x)$ can be similarly constructed, and the residuals *u* and *v* estimated using the expressions

$$y = Y - \hat{m}_Y(X) \tag{18}$$

and

$$u = Z - \hat{m}_Z(X) \tag{19}$$

Using these residuals the PMI can then be calculated as

$$I'_{ZY.X} = I(v;u) \tag{20}$$

where the subscript notation $I'_{ZY.X}$ or $I(Z;Y \mid X)$ can be used. PMI allows for the evaluation of variables taking into account any information already provided by a given variable *X*.

Given a candidate set *C*, and output variable, *Y*, the PMI based input variable selection (PMIS) algorithm proceeds at each iteration by finding the candidate, C_s, that maximises the PMI with respect to the output variable, conditional on the inputs that have been previously selected. The statistical significance of the PMI estimated for C_s can be assessed based on the confidence bounds drawn from the distribution generated by a bootstrap loop. If the input is significant, C_s, is added to *S* and the selection continues; otherwise there are no more significant candidates remaining and the algorithm is terminated [25][26].

The size of the bootstrap, B, is important in the implementation of PMIS since it can influence both the accuracy and overall computational efficiency of the algorithm. May et al [26] discuss the implications for selecting the bootstrap size in terms of the accuracy – computational efficiency trade-off and present a new approach which does not rely on a bootstrap or direct comparison with the critical value of MI (as is necessary with some of the other approaches compared, such as the tabulated critical values approach [26]). In [26] the use of the Hampel test criterion is suggested as a termination criterion.

3.5 Hampel Test Criterion

Outlier detection methods are robust statistical methods for determining whether a given value, x, is significantly different from another within a set of values X. In the case of PMIS, having identified the most relevant candidate, the outlier, it is necessary to determine whether this candidate is statistically significantly greater than the others and to keep this candidate if it is. The Z-test is a commonly adopted approach for outlier detection where the deviation of a single observation is compared with the sample mean of all observations. Based on the 3σ rule for Gaussian distributions, outliers lie greater than three standard deviations from the population mean and therefore an observed value with a Z-score greater than 3 is generally considered to be an outlier. The Z-test can be particularly sensitive when a population contains multiple outliers. One very distant outlier could disrupt the distribution of the population (mean and variance) resulting in other outliers not being identified i.e., hiding and masking outliers. The sensitivity of outlier detection methods to masking is determined based on the proportion of outliers that must be present to significantly alter the data distribution, referred to as the breakdown point, which is $1/n$ for the Z-test since only one sufficiently large outlier will cause the test to breakdown [26].

Since the candidate set of variables in the PMIS method is likely to contain more than one relevant variable (analogous to outliers in the aforementioned outlier test) a modified Z-score is necessary to improve the robustness of the test. The Hampel distance test proposed in [53] and compared in [26] is based on the population median. Because the Hampel distance test breakdown point is 2/n it is considered to be one of the most robust outlier tests when the data contains multiple outliers. To calculate the Hampel distance, the absolute deviation from the median for all candidates is calculated as follows

$$d_j = \left| I_{C_j Y \cdot S} - I^{(50)}_{C_j Y \cdot S} \right| \tag{21}$$

where d_j denotes the absolute deviation and $I^{(50)}_{C_j Y \cdot S}$ is the medium PMI for candidate set C. Taking $d_j^{(50)}$ as the median absolute deviation (MAD), the Hampel distance (modified Z-score) for candidate C_j is

$$Z_j = \frac{d_j}{1.4826 d_j^{(50)}}. \tag{22}$$

The factor of 1.4826 scales the distance such that the rule Z>3 can be applied, as in the case of the conventional Z-test [26]. The value Z_s is determined for candidate C_s and if $Z_s > 3$, the candidate is selected and added to S; otherwise the forward selection algorithm is terminated as described in the following subsection.

PMIS algorithm using Hampel distance criterion

1: Let $S \rightarrow \phi$ (Initialisation)
2: While $S \neq \phi$ (Forward Selection)
3: Construct kernel regression estimator $\hat{m}_Y(S)$
4: Calculate residual output $u = Y - \hat{m}_Y(S)$
5: For each $C_j \in C$
6: Construct kernel regression estimator $\hat{m}_{C_j}(S)$
7: Calculate residual output $u = C_j - \hat{m}_{C_j}(S)$
8: Estimate $I(v;u)$
9: Find candidate C_s (and v_s) that maximises $I(v;u)$
10: Estimate Z_s for C_s
11: If $Z_s > 3$ (Selection/Termination)
12: Move C_s to S
13: Else
14: Break
15: Return

Using PMIS the optimal selection of time delayed EEG signal samples which minimise the uncertainty about a future output (prediction) can be estimated.

3.6 Using PMIS to Optimize the NTSPP Framework

For every channel, in the case of the BCI presented in sections 2 and 3.1, there is assumed to be an optimal selection of input variables (EEG time series embedding), that will enable accurate prediction for the channel and accurate specialisation of a neural network for that channel. The optimal selection however is likely to differ depending on the class of data (motor imagery) being assessed. So, for a 3 channel system with 2 classes of data there is assumed to be at least 6 optimal embedding configurations, one for each channel per class. When applying a BCI that involves time embedding the EEG for prediction, as is the case for the NTSPP framework, the optimal embedding for both classes cannot be applied simultaneously in the online BCI as the class is unknown a priori and a decision has to be made given data from 3 channels. It is therefore necessary to decide which embedding should be applied, not necessarily to maximise the prediction accuracy for both classes, but to maximise the specialisation for the networks in such a way that the difference between signals predicted for both classes is maximal i.e., separability is maximised. In the case where 3 channels and 2 classes are available,

Table 2 Different combinations of embedding dimension for three channels. For any configuration each can use the embedding parameters that are optimal for either class '1' or '2' but not both.

Configuration	*C3*	*C4*	C_z
A	1	1	1
B	1	1	2
C	1	2	1
D	1	2	2
E	2	1	1
F	2	1	2
G	2	2	1
H	2	2	2

$C \in \{1,2\}$, there are 2^3 possible configurations for deciding which of the channel's assumed optimal embedding configuration should be used as shown in Table 2.

From Table 2, if configuration A is selected then the embedding values selected for class 1 on all channels would be used whereas if configuration D is selected the embedding parameters chosen for class 1 would be used for channel C3 and those chosen for class 2 would be used for channels C4 and Cz. As this chapter presents the first assessment of this approach, a heuristic based approach was adopted to determine the best configuration. The following section describes the complete BCI setup and parameters optimization procedure.

3.7 Parameters Optimization and BCI Setup

In motor imagery BCIs, the parameter search space and the available data can be extensive, particularly when there are multiple stages of signal processing, therefore a phased approach to parameter selection is conducted. In the proposed BCI setup it is necessary to find of the optimal combination of lagged input variables (embedding parameters) to train the predictor networks. An inner-outer cross-validation (CV) is performed, where all other BCI parameters are optimized for each of the embedding configurations including the optimal subject-specific frequency bands (shown in Fig. 4). In the outer fold, NTSPP is trained on up to 10 trials randomly selected from each class (2 seconds of event related data from each trial resulting in 2500 samples for each channel/class) using standard time series embedding parameters: embedding (Δ=6) and time lag (τ=1). The trained networks then predict all the data from the training folds to produce a surrogate set of trials containing only EEG predictions. No parameter tuning is necessary at this stage as the SOFNN adapts autonomously to the signals [23]. The 4 training folds from the outer splits are then split into 5 folds on which an inner 5-fold cross validation is performed. Firstly, the time point of maximum separability is found for the inner data and, (if necessary, channel selection can be performed), both using the R^2 correlation analysis with a standard 8-26 Hz band [40]. Using the information regarding the optimal time point, a 2 second window of data around the time point

of maximum separability is taken from 10 randomly chosen trials from each class and PMIS is applied to find the optimal embedding for each channel and each class. Using one of the combinations shown in table 1 the NTSPP framework is retrained with these embedding parameters, again using 10 randomly chosen trials from the outer training folds but in this case using the 2 second window of data around the time point of maximum separability (using the most separable data segments increases network specialization). A new surrogate dataset for the outer training fold data is obtained and the 4 training folds from the outer splits are then split into 5 folds on which another inner 5-fold cross validation is performed. Firstly, the time point of maximum separability is found for the inner data using the R^2 correlation analysis with a standard 8-26 Hz band [40]. Using the best time point and best channels from the correlation analysis, a PSO based search is conducted to identify the optimal frequency bands where CSP, feature extraction and classification is performed to determine classification accuracy levels on each of the folds for each of the bands selected by PSO and tested using 4 CSP surrogate channels [40]. After each frequency band is tested on the test fold, PSO swarm particles communicate the accuracy levels to one another and the algorithm converges, identifying the optimal band for that test fold much quicker than searching the complete space of all the possible bands (cf. Fig. 4(a) for a graphical representation of a PSO search). After optimal bands for each of the inner folds have been identified the finally selected band is the average classification accuracy (CA) of the 5 bands weighted by the CA of the test fold as illustrated in Fig. 4(b). NTSPP-SF-CSP is then applied on the outer fold training set, where a feature set is extracted and an LDA classifier is trained at every time point across the trials and tested for that point on the outer test folds. The average across the five-folds is used to identify the optimal number of CSPs (between 1-3 from each side of W) and the final time point of maximum separation for the corresponding combination of PMIS selected lagged input variables.

There are eight combinations of selected lag variable combinations as shown in Table 2 therefore the above process is conducted for each of the combinations. At the end of this process the embedding configuration which provides the best mean accuracy is known however it is then necessary to select which channel-specific embedding works best i.e., from the outer cross validation PMIS is applied for each of the 5 training folds and each time the exact embedding parameters may differ for each channel. To obtain the best setup for cross session tests (to ensure generalization to the unseen testing data) the complete system is retrained a further 5 times on all the training data using the chosen parameters for each fold and tested on the training data 5 times. In this case the lag combination for each of the 5 folds is tested with other BCI parameters selected in the cross-validation. The embedding parameters configuration and parameter combination which provides the highest mean accuracy across the 5 tests on the complete training data set is used for cross session tests. In the case where two tests produce the same results on one of the training data tests the parameter setup that achieves the average best accuracy across 8 lag configurations for a particular training test, corresponding to the best average lag configuration across the 5 training tests, is used to determine

the best setup. The system is finally tested on the unseen test/evaluation data (as given for the BCI competition outlined in section 2). All parameter optimization is expedited using the Matlab® Parallel Processing Toolbox and a high performance computing (HPC) cluster with 384 cores. Subject analysis and 5-fold cross validations were run in parallel along with parallelization of parts of the kernel regression estimation and PMI calculation for selecting input variables and multiple cross validation tests.

4 Results

The objective of this research was to improve the NTSPP framework, which is a predictive framework involving training a mixture of experts on EEG data produced for two classes of motor imagery recorded from three channels. The hypothesis is that specializing the networks for a particular motor imagery (class) leads to improvement in the separability of the predicted output i.e., when the mixture of networks produce predictions for an unknown class of motor imagery, networks trained on that particular class of motor imagery should predict the data sufficiently accurately and differently compared to the other networks which are trained on the other class of motor imagery. Maximizing the difference in the prediction for each class of motor imagery ultimately should lead to better classification accuracy (BCI performance) when features are extracted from the predicted signals and classified i.e., when all other components are merged with the predictive framework. The hypothesis is based on the observation that dynamics of the EEG differs across channels and between motor imageries. The 2nd hypothesis is thus that each channel will have different and optimal embedding parameters which will optimize prediction performance and enable network specialization. Therefore, selecting these optimal embedding parameters will result in improved NTSPP performance and thus improved BCI performance. For example, if PMIS selected x(t-1), x(t-3), and x(t-5) as the best predictors for channel C3 for class 1 and x(t-1), x(t-2), x(t-3), and x(t-10) for the same channel but for class 2, training the networks for class 1 and class 2 on only one of these embedding combinations for this channel, for example, the embedding parameters for class 1, then the class 2 network would not be able to specialize/train on the same channel using class 2 data as accurately, as the optimal embedding parameters for that channel are not been utilised i.e., for each channel only of the two networks trained on the channel is specialized to predict the channel whilst the other is not.

To test both hypotheses the overall BCI performance and the selected embedding parameters for each channel and class that produce the optimal BCI performance are assessed. BCI performance with NTSPP and PMIS is compared with the case were no NTSPP is performed (only CSP and SF) and where NTSPP is performed with a standard embedding/time lag setup (NTSPP6).

Table 3 The optimal time series embedding parameters for each channel and for each class for all subjects. The optimal configuration for NTSPP is shown in column 2 (corresponding to table 2) and in bold in wide columns 3 and 4.

S	NTSPP Conf.	Class 1 (Left Hand Motor Imagery) C3	C4	Cz	Class 2 (Right Hand Motor Imagery) C3	C4	Cz
1	C 1,2,1	**x(t-1),x(t-10)**	x(t-1), x(t-9)	**x(t-1), x(t-3)**	x(t-1), x(t-5)	**x(t-1)**	x(t-1),x(t-10)
2	D 1,2,2	**x(t-1), x(t-2)**	x(t-1),x(t-2)	x(t-1),x(t-2) x(t-5),x(t-6)	x(t-1)	**x(t-1),x(t-2)**	**x(t-1)**
3	A 1,1,1	**x(t-1),x(t-2), x(t-4),x(t-5), x(t-6),x(t-7), x(t-10)**	**x(t-1),x(t-2), x(t-3),x(t-4), x(t-5),x(t-6)**	**x(t-1),x(t-2), x(t-3),x(t-5)**	x(t-1),x(t-2), x(t-3),x(t-4), x(t-5),x(t-6)	x(t-1),x(t-2), x(t-3),x(t-4), x(t-5),x(t-6)	x(t-1),x(t-2), x(t-3),x(t-10)
4	B 1,1,2	**x(t-1),x(t-10)**	**x(t-1)**	x(t-1),x(t-10)	x(t-1), x(t-7), x(t-8)	x(t-1)	**x(t-1),x(t-10)**
5	B 1,1,2	**x(t-1)**	**x(t-1),x(t-2), x(t-3),x(t-4)**	x(t-1), x(t-2)	x(t-1),x(t-2), x(t-3),x(t-4), x(t-5),x(t-7)	x(t-1), x(t-2)	**x(t-1),x(t-2), x(t-7),x(t-8)**
6	E 2,1,1	x(t-1),x(t-2), x(t-6),x(t-8), x(t-9)	**x(t-1), x(t-2), x(t-3),x(t-4), x(t-5)**	**x(t-1),x(t-2), x(t-3)**	**x(t-1), x(t-2), x(t-3), x(t-5)**	x(t-1), x(t-2), x(t-3), x(t-4), x(t-5),x(t-10)	x(t-1),x(t-2), x(t-3),x(t-7), x(t-8)
7	A 1,1,1	**x(t-1)**	**x(t-1)**	**x(t-5),x(t-6)**	x(t-1)	x(t-1)	x(t-6)
8	B 1,1,2	**x(t-1)**	**x(t-1), x(t-8)**	x(t-1)	x(t-1)	x(t-1),x(t-10)	**x(t-10)**
9	A 1,1,1	**x(t-1)**	**x(t-1)**	**x(t-1)**	x(t-1)	x(t-1)	x(t-2), x(t-3)
10	C 1,2,1	**x(t-1)**	x(t-1), x(t-2)	**x(t-1), x(t-2), x(t-3)**	x(t-1)	**x(t-1)**	x(t-1),x(t-2), x(t-3)
11	D 1,2,2	**x(t-1),x(t-2), x(t-3),x(t-4), x(t-5),x(t-6), x(t-8),x(t-9)**	x(t-1)	x(t-1),x(t-2), x(t-3),x(t-4), x(t-5),x(t-6)	x(t-1),x(t-2)	**x(t-1),x(t-2), x(t-5)**	**x(t-1),x(t-2), x(t-7),x(t-10)**
12	C 1,2,1	**x(t-1)**	x(t-1)	**x(t-1)**	x(t-1),x(t-10)	**x(t-1),x(t-10)**	x(t-1),x(t-8), x(t-9),x(t-10)
13	H 2,2,2	x(t-1)	x(t-1),x(t-8)	x(t-1),x(t-2), x(t-8)	**x(t-1),x(t-7), x(t-8)**	**x(t-1)**	**x(t-1),x(t-7), x(t-8)**
14*	G 2,2,1	x(t-1),x(t-2), x(t-9)	x(t-1),x(t-2)	**x(t-1),x(t-7), x(t-8),x(t-10)**	**x(t-1),x(t-2), x(t-3),x(t-4), x(t-5)**	**x(t-1),x(t-2), x(t-3),x(t-4), x(t-5)**	x(t-1),x(t-4)
15	A 1,1,1	**x(t-1),x(t-2),**	**x(t-1),x(t-2)**	**x(t-1),x(t-2)**	x(t-1)	x(t-1)	x(t-1),x(t-2),
16	H 2,2,2	x(t-1)	x(t-1), x(t-2), x(t-4),x(t-10)	x(t-1), x(t-2), x(t-3),x(t-10)	**x(t-1)**	**x(t-1), x(t-2), x(t-10)**	**x(t-1)**
17	H 2,2,2	x(t-1)	x(t-1)	x(t-1)	**x(t-1)**	**x(t-1)**	**x(t-1),x(t-10)**
18	G 2,2,1	x(t-9)	x(t-1)	**x(t-1)**	**x(t-1), x(t-9)**	**x(t-1)**	x(t-1)

4.1 PMIS Selected Embedding Parameters

Table 3 shows the PMIS selected embedding/lag parameters for each of the 3 channels for each class. The best configuration in terms of which embedding/lag parameters were used for each channel are shown in bold (as outlined only one embedding/lag setup can be used for each channel – the best training accuracy during system optimization as outlined above is used to determine which setup is used, either class 1 or class 2). As can be seen, across all subjects there are different embedding/lag parameters chosen using PMIS. Within subjects there are different embedding/lag parameters for each channel and for each class. Using the selection method outlined above, the best configuration of embedding/lag parameters differ significantly across subjects. Colum 2 shows the lag configurations corresponding to Table 2. For subjects 3, 7, 15, 16 and 17 the best configuration is

based on the embedding/lag combinations for one class across all signals. The optimal setup could be derived based on one class only for a number of other subjects, as the embedding/lag combination is identical for some channels regardless of class e.g., for subject channel Cz parameters could have been used for both classes and likewise for subject 18 (subject 4 could have used all class 1 parameters and subject 18 could have used all class 2 parameters). This selection process could have chosen either in these cases. For the remaining subjects however a mixture of parameters have been selected between the two classes, emphasizing the need to assess all embedding parameters combination on a channel-, class- and subject-specific basis. The following subsection outlines how these subjects performed in terms of overall classification accuracy using these embedding configurations in the NTSPP framework.

4.2 BCI Performance

The cross validation (CV) performances for all subjects are presented in Figure 6(a) whilst the cross session (x-Session) single-trial performances are presented in Figure 6(b). The average performances across subjects are presented in Figure 6(c). In the majority of cases NTSPP-PMIS provides the best cross-validation performance (within session) for all subjects (Fig. 6(a)). In some cases NTSPP6 outperforms NTSPP-PMIS. Subject 10 is the only case where there is a significant drop in CV performance given by NTSPP-PMIS compared to No-NTSPP. There is a slight drop in CV performance for subjects 17 and 18 but overall the CV results indicate a slight improvement in the average within-session performance for NTSPP6 and a greater improvement for NTSPP-PMIS across all subjects compared to No-NTSPP. The average across subjects is shown in Fig. 6(c). The averages are compared across all subjects as well as across the two competition groupings as both datasets have different attributes which influence performance[2]. In terms of the CV average there are slight improvements given by NTSPP-PMIS which are statistically significant ($p<0.05$) as shown in Table 4 where the results of two statistical tests are presented. The parametric statistical test repeated measures ANOVA (which is akin to a *t*-test (related) for two groups) [54] and the Wilcoxon signed rank test [55], a non-parametric statistical test, are used for clarity (the results indicate from both tests are similar and correlated). The improvement given by NTSPP6 over No-NTSPP is not significant in the CV test but is more significant (not statistically) in the cross-session tests on all subjects. The cross session performance difference between No-NTSPP and NTSPP-PMIS is statistically significant however NTSPP–PMIS is not shown to be statistically better than NTSPP6 in the cross session tests. The performance of the BEST NTSPP-PMIS cross session results are presented here to show what is theoretically possible with

[2] For dataset 2B (subjects 1-9), the data used for training are from a 3rd feedback session after 2 sessions without feedback and the testing data is from two further feedback sessions whereas the dataset for 2A (subjects 10-18) are trained and tested on 2 sessions with no feedback and within these sessions subjects performed another 2 motor imageries (4 class data acquisition; there is also a significant difference in the number or trials performed for both groups (cf. section 2 for further details).

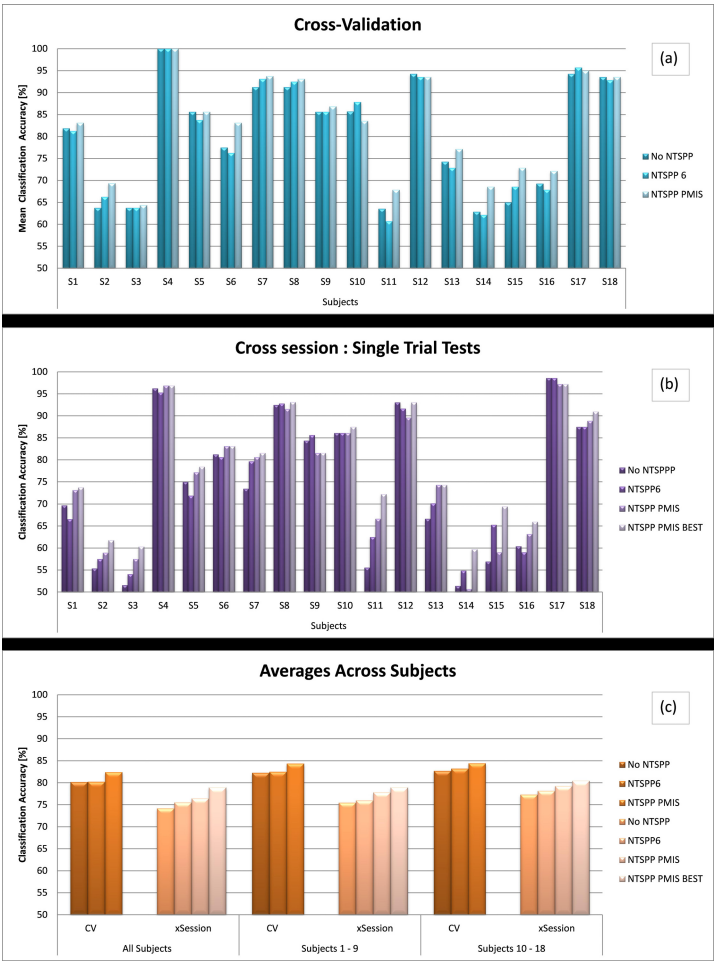


Fig. 5 BCI performance results for all subjects and approaches: (a) average cross-validation classification accuracy; (b) cross-session (x-Session) classification accuracy (c) mean accuracies across all subjects and two groups of subjects (groups based on datasets 2A and 2B). Results are presented for 3 methods: No NTSPP, NTSPP6 (standard embedding dimension 6 and time lag 1) and NTSPP with embedding selected using PMIS and configured according to Table 3. In figures (a) and (b) the absolute best performing NTSPP PMIS embedding setup is shown for information only (this best setup is the accuracy that could have been obtained if the absolute best embedding setup was determined from the training data i.e., in some cases the setup chosen did not provide the best generalization performance on the test data.

the proposed NTSPP-PMIS approach if the parameters can be selected appropriately from the training data (i.e., these results were generated by applying all possible NTSPP-PMIS configurations across the sessions and viewing the best results). The results show that the possible best performances are statistically

better than those produced using the proposed heuristic configuration optimization method for NTSPP-PMIS. The significance of the results is discussed in the following section.

Table 4 Results of statistical tests comparing the average performance across subjects (All, S1-S9 and S10-S19) for each of the methods/approaches (p-values), where the average performances are shown in Figure 6(c). $p<0.05$ indicates a statistical difference in the performance produced by the methods compared. Results of two statistical tests (parametric and non-parametric) are shown for comparison and verification.

Methods	**CV**	**Cross Session**		
Repeated Measures ANOVA	**All**	**All**	**S1-9**	**S10-18**
No NTSPP vs NTSPP6	0.8714	0.1033	0.6149	0.1094
No NTSPP vs NTSPP-PMIS	**0.0046**	**0.0264**	0.0566	0.1953
NNTSPP6 vs NTSPP-PMIS	**0.0093**	0.2823	0.1271	1
NTSPP PMIS vs BEST	-	**0.0052**	**0.0177**	**0.0151**
Wilcoxon Signed Rank Test				
No NTSPP vs NTSPP6	0.7925	0.1640	0.7969	0.1250
No NTSPP vs NTSPP-PMIS	**0.0019**	**0.0331**	0.0742	0.2656
NTSPP6 vs NTSPP-PMIS	**0.0083**	0.2366	0.1289	0.9688
NTSPP PMIS vs BEST	-	**0.0002**	**0.0313**	**0.0156**

5 Discussion

This study presents for the first time the use of partial mutual information input variable selection (PMIS) for selecting channel-, class- and subject specific embedding parameters from EEG time-series. The results presented in section 4.1 show that, depending on the particular brain state (class), the channel-specific embedding varies and is subject-specific. Past studies have investigated the optimal subject-specific embedding parameters for BCI [15][21] but focused on using one set for all channels with the same embedding parameters being selected for both classes i.e., the embedding was optimized based on overall classification performance without first selecting embedding parameters for particular channels. The results presented here show the variability in the brain and intra- and inter-subject differences in EEG dynamics. It is therefore recommended to optimize the channel specific embedding parameters when attempting to make predictions about a future brain state, be it one step or multiple steps ahead.

Although, in this study, the aim is not to exploit the use of advanced prediction of future brain states to reduce system latency (cf. next section for a discussion on how this may be potentially beneficial in BCI), the NTSPP approach is based on EEG time-series prediction and the results clearly demonstrate that there are improvements given by the NTSPP framework when channel- and class-specific embedding configurations are deployed for each subject. The results show that within session cross-validation differences between NTSPP-PMIS and NTSPP6 or No NTSPP are statistically significant and that the cross session performance

difference between NTSPP-PMIS and No-NTSPP is statistically significant. In previous work it was shown that in some cases, depending on the number of channels and classes being investigated, NTSPP6 provides significant improvement over No NTSPP whereas in this work the cross session differences are not shown to be statistically different, although an improvement is observable. The approach presented in [22] involved only testing four frequency bands between 8-24Hz for the EEG spectral filter whereas in this work the frequency bands are subject-specifically tuned in the setup with a fine resolution using PSO and then applied along with NTSPP6[3] i.e., the results presented here suggest that NTSPP6 is less effective when the subject-specific bands are tuned. However, as shown here, when we deploy PMIS embedding selection there are improvements even with optimized frequency bands in the spectral filter. With the proposed parameter configuration using PMIS and a heuristic as well as computationally intelligent search methods (PSO and SOFNN) for other parameter combinations/settings, NTSPP-PMIS can generalize reasonably well across sessions. This therefore is a positive indication that the use of NTSPP can indeed improve performance of the BCI. The BEST results (as outlined these are identified after viewing the testing performance across all embedding configurations) show there can be even greater gains provided by the NTSPP-PMIS framework if the parameter optimization approach is further improved to ensure better generalization. The following subsection outlines why the approach used is suboptimal and other limitations of this study.

5.1 Limitations

In terms of PMIS and NTSPP, only 10 trials randomly selected from the available trials from each class are used to, firstly, identify the optimal embedding for each channel and then to train the SOFNNs in the NTSPP framework. Using more trials in the PMIS setup, only using trials which are highly separable i.e., omitting trials which are less separable, may improve the specificity and accuracy of the PMIS algorithm. Likewise, for NTSPP and the SOFNN training, using more trials and only those that are most separable, along with PMIS selected embedding as described may enhance the specialization of the networks leading to increased difference in the prediction for both classes and enhanced separability. The SOFNN is deployed in this framework using standard hyper parameters, identified based on a study of a small number of subjects [37]. It is highly probable that fine tuning the SOFNN parameters to suit the channel-specific embedding will also lead to greater specialization. In addition, the data segments within the event-related portion of trials on which PMIS and the SOFNN are deployed could be fine-tuned and assessed more closely using smaller or larger segments around the time point of maximum separation in trials (in this work a 2s window around the max separation point was used). Using more trials from which to select data may also improve specialization. In this study only 2 seconds of the data was used from 10 randomly selected trials resulting in 2500 samples for PMI selection and SOFNN

[3] Even though the same subjects are analyzed the data splits in [22] are also different (based on the feedback and non-feedback sessions) and therefore results presented here are not directly comparable to those presented in [22].

training. This is a low number of trials relative to the amount of data available however results in a significant number of samples on which to train multiple networks, multiple times in a cross validation. Simply selecting data more instinctively could have a major impact on PMIS and the SOFNNs in the NTSPP framework without any additional parameter optimization.

Improvements to other elements of the BCI are possible and ongoing however it is desirable to keep parameter tuning in the BCI setup to a minimum or indeed in any application therefore there is always the aim to ensure the system can be setup quickly using an auto-calibrating approach, hence the reason for using computational intelligence based approaches such as the SOFNN which can adapt and tune its weights and structure automatically during the learning process, and the use of PSO to select optimal frequency bands quickly and efficiently. CSP is not only used here to improve separability but to help identify redundancy in the signals. The use of linear classifiers for easy training and adaptation is necessary but classifier performance can be improved to account for inter-session variability and sensorimotor learning as the subject endeavors to improve BCI performance (research is ongoing in this area [56][57]). Future work will involve investigating the parameters that are providing the best cross-session performances and developing an optimized and efficient framework where optimal performance and cross session generalization is guaranteed. For example, subject 14 in this work was poorly performing regardless of the method deployed but NTSPP-PMIS failed completely (~50% classification accuracy) given the parameters selected on the training data whereas the BEST performance shows that NTSPP-PMIS parameters could have been much better for this subject had the training data being more carefully used to setup the system. These issues are currently being investigated along with other potential benefits of the NTSPP framework as outlined in the following subsection.

6 Conclusions and Future Work

This chapter has shown for the first time that partial mutual information input variable selection (PMIS) can be used to select embedding parameters for EEG time series prediction and by selecting channel-, class- and subject-specific embedding parameters predictive performance and over all classification of EEG data can be improved for a two class EEG-based BCI using the NTSPP framework. The PMIS approach can be improved by using more data and further assessment of the criteria for considering whether a particular embedded sample of the time series provides information about the predicted input. This may be improved using the bootstrapping or Akaike Information Criterion as compared by May et al [26] however the approach used here, involving the Hampel distance criterion, is efficient. By exploring better parameters for the PMIS approach, the NTSPP setup and the complete BCI it is expected that the BCI presented here can be improved significantly. This work provides evidence of this potential. The PMIS approach will also aid in the investigation of other BCI configurations involving the NTSPP framework, for example, multiclass systems and multiple channel EEG montages. Previous work has already shown that the performance gains provided by the

NTSPP framework are greater when multiple channels and multiple classes are used [22][55]. Channel- and class-specific embedding is likely to further increase that improvement. NTSPP has also been shown to have the capacity to reduce the latency involved in motor imagery BCIs involving continuous classification; producing higher signal separability faster (i.e., earlier in the trial) by predicting the EEG times series multiple steps ahead [59]. This has the potential to reduce the time required (latency) for a subject to exceed a threshold with the continuous classifier output, as the NTSPP predicts multiple steps ahead in time characteristics in the data which are more separable. Features can then be extracted from the predicted separable segments of the data before that separability actually is produced by the sensorimotor activity. A preliminary study of this is presented in [59]. Again, that preliminary study used standard embedding parameters. For multiple-step-ahead prediction the prediction error increases as the prediction horizon increases and therefore PMIS embedding parameter selection will be even more pertinent and can be exploited in such a multi-step-ahead NTSPP framework. Further work will be carried out to verify if combining CSP and SF with the multiple-step-ahead prediction NTSPP framework and PMIS has potential for improved accuracy and information transfer rate in BCI. It may also be possible that PMIS selected EEG embedding parameters can be used as class predictors i.e., the optimal selected embedding parameters can be selected on a trial-by-trial basis using PMIS and used as signal features. The investigation would involve determining if such features provided sufficient inter class variability and intra class correlation to enable reliable discrimination of brain states.

In summary, this work shows how a range of traditional signal processing tools can be combined with multiple computational intelligence based approaches to develop a more autonomous parameter optimization setup and ultimately a more accurate BCI. Finally, the novel developments in signal processing and embedding selection using PMIS will be integrated with our real-time BCI, when sufficiently validated, for application in assistive technologies and entertainment for the physically impaired [4][5][6][8] and rehabilitation [7].

References

[1] Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtscheller, G., Vaughan, T.M.: Brain-computer interfaces for communication and control. J. Clinical Neurophysiology 113, 767–791 (2002)

[2] Kubler, A., Kotchoubey, B., Kaiser, J., Wolpaw, J.R., Birbaumer, N.: Brain-Computer communication: unlocking the locked-in. Psychological Bulletin 127(3), 358–375 (2001)

[3] Pfurtscheller, G., Guger, C., Muller, G., Krausz, G., Neuper, C.: Brain oscillations control hand orthosis in a tetraplegic. Neurosci. Lett. 292, 211–214 (2000)

[4] Coyle, D., Satti, A., Stow, J., McCreadie, K., Carroll, A., McElligott, J.: Operating a Brain Computer Interface: Able Bodied vs. Physically Impaired Performance. In: Proc. of the Recent Advances in Assistive Technology & Engineering Conference (2011)

[5] Stow, J., Coyle, D., Carroll, A., Satti, A., McElligott, J.: Achievable Brain Computer Communication through Short Intensive Motor Imagery Training despite Long Term Spinal Cord Injury. In: Proc. of the Annual IICN Registrar's Prize in Neuroscience (2011)
[6] Coyle, D., Carroll, A., Stow, J., McCann, A., Ally, A., McElligott, J.: Enabling Control in the Minimally Conscious State in a Single Session with a Three Channel BCI. In: Proc. of the 1st International DECODER Workshop (2012)
[7] Prasad, G., Herman, P., Coyle, D., McDonough, S., Crosbie, J.: Applying a brain-computer interface to support motor imagery practice in people with stroke for upper limb recovery: a feasibility study. J. Neuroeng. Rehab. 7(60), 1–17 (2011)
[8] Coyle, D., Garcia, J., Satti, A., McGinnity, T.M.: EEG-based Continuous Control of a Game using a 3 Channel Motor Imagery BCI. In: IEEE Symposium Series on Computa-tional Intelligence, pp. 88–93 (2011)
[9] Enzinger, C., Ropele, S., Fazekas, F., Loitfelder, M., Gorani, F., Seifert, T., Reiter, G., Neuper, C., Pfurtscheller, G., Muller-Putz, G.: Brain motor system function in a patient with complete spinal cord injury following extensive brain–computer interface training. Exp. Brain Res. 190, 215–223 (2008)
[10] Chatrian, G.E., Peterson, M., Lazarte, J.A.: The blocking of the rolandic wicket rhythm and some central changes related to movement. Electroencephalogr. Clin. Neurophysiol. 11, 497–510 (1959)
[11] Pfurtscheller, G., Neuper, C., Flotzinger, D., Pregenzer, M.: EEG-based discrimination between imagination of right and left hand movement. Electroencephalography and Clinical Neurophysiology 113(6), 642–651 (1997)
[12] Felzer, T., Freisleben, B.: Analyzing EEG signals using the probability estimated guarded neural classifier. IEEE Trans. on Neural Sys. and Rehab. Eng. 11(2), 361–371 (2003)
[13] Anderson, C., Sijercic, Z.: Classification of EEG signals from four subjects during five mental tasks. In: Proc of the Conference on Eng. Applications in Neural Networks (EANN 1996), pp. 407–414 (1996)
[14] Muller, K.-R., Anderson, C.W., Birch, G.E.: Linear and nonlinear methods for brain-computer interfaces. IEEE Trans. on Neural Systems and Rehab. Eng. 11(2), 165–169 (2003)
[15] Schlogl, A., Flotzinger, D., Pfurtscheller, G.: Adaptive autoregressive modelling used for single-trial EEG classification. Biomedizinische Technik, Band 42, 162–167 (1997)
[16] Forney, E., Anderson, C.W.: Classification of EEG during Imagined Mental Tasks by Forecasting with Elman Recurrent Neural Networks. In: Proceedings of the International Joint Conference on Neural Networks, pp. 2749–2755 (2011)
[17] Pfurtscheller, G., Neuper, C., Schlogl, A., Lugger, K.: Separability of EEG signals recorded during right and left motor imagery using adaptive autoregressive parameters. IEEE Transactions on Rehabilitation Engineering 6(3), 316–324 (1998)
[18] Schloegl, A.: The electroencephalogram and the adaptive autoregressive model: theory and applications. Shaker Verlag, Aachen (2000)
[19] Kohlmorgen, J., Müller, K.-R., Rittweger, J., Pawelzik, K.: Identification of non-stationary dynamics in physiological recordings. Biological Cybernetics 83(1), 73–84 (2000)
[20] Haselsteiner, E., Pfurtscheller, G.: Using Time-Dependent NNs for EEG classification. IEEE Trans. on Rehab. Eng. 8(4), 457–462 (2000)

[21] Coyle, D., Prasad, G., McGinnity, T.M.: A time-series prediction approach for feature extraction in a brain-computer interface. IEEE Transactions on Neural Systems and Rehabilitation Engineering 13(4), 461–467 (2005)
[22] Coyle, D.: Neural network based auto association and time-series prediction for biosignal processing in brain-computer interfaces. IEEE Computational Intelligence Magazine 4(4), 47–59 (2009)
[23] Coyle, D., Prasad, G., McGinnity, T.M.: Faster self-organizing fuzzy neural network training and a hyperparameter analysis for a brain-computer interface. IEEE Transactions on Systems, Man and Cybernetics (Part B) 39(6), 1458–1471 (2009)
[24] Coyle, D., Prasad, G., McGinnity, T.M.: Improving the separability of multiple feature types for a brain-computer interface by neural time-series prediction preprocessing. Biomedical Signal Processing and Control, 196–204 (2010)
[25] Sharma, A.: Seasonal to inter annual rainfall probabilistic forecasts for improved water supply management: part 1 – a strategy for system predictor identification. Journal of Hydrology 239, 232–239 (2000)
[26] May, R.J., Maier, H.R., Dandy, G.C., Gayani Fernando, T.M.K.: Non-linear variable selection for artificial neural networks using partial mutual information. Environmental Modelling and Software 23, 1312–1326 (2008)
[27] Blankertz, et al.: BCI Competition III and IV (2005), http://www.bbci.de/competition/
[28] Blankertz, et al.: The BCI competition. III: Validating alternative approaches to actual BCI problems. IEEE Trans. Neural. Syst. Rehabil. Eng. 14, 153–159 (2006)
[29] Schlogl, A., Lee, F., Birschof, H., Pfurtscheller, G.: Characterization of four-class motor imagery EEG data for the BCI-competition 2005. J. of Neural Engineering 2, L.14–L.22 (2005)
[30] Schlogl, et al.: BCI-Competition IV (Dataset 2A and 2B) (2008), http://www.bbci.de/competition/iv/desc_2b.pdf, http://www.bbci.de/competition/iv/desc_2a.pdf
[31] Leeb, R., Lee, F., Keinrath, C., Scherer, R., Bischof, H., Pfurtscheller, G.: Brain-computer communication: motivation, aim, and impact of exploring a virtual apartment. IEEE Transactions on Neural Systems and Rehabilitation Engineering 15, 473–482 (2007)
[32] Schlogl, A., Keinrath, C., Zimmermann, D., Scherer, R., Leeb, R., Pfurtscheller, G.: A fully automated correction method for EOG artifacts in EEG recordings. Clin. Neuro-Phys. 118(1), 98–104 (2007)
[33] Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Networks 2, 359–366 (1989)
[34] Jang, J.S.R.: Neuro-Fuzzy & Soft Computing. Prentice-Hall (1997)
[35] Leng, G.: Algorithmic Developments for Self-Organising Fuzzy Neural Networks. PhD Dissertation, University of Ulster (2003)
[36] Prasad, G., McGinnity, T.M., Leng, G., Coyle, D.: On-line identification of self-organizing fuzzy neural networks for modelling time-varying complex systems. In: Plamen, et al. (eds.) Evolving Intelligent Systems, pp. 302–324. John Wiley, NY (2010)
[37] Coyle, D., Prasad, G., McGinnity, T.M.: Faster Self-organising Fuzzy Neural Network Training and Improved Autonomy with Time-Delayed Synapses for Locally Recurrent Learning. In: Temel (ed.) System and Circuit Design for Biologically-Inspired Learning, pp. 156–183. IGI-Global (2010)

[38] Ramouser, H., Muller-Gerking, J., Pfurtscheller, G.: Optimal spatial filtering of single trial EEG during imagined hand movement. IEEE Trans. on Rehab. Eng. 8(4), 441–446 (2000)
[39] Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., Müller, K.R.: Optimizing spatial filters for robust EEG Analysis. IEEE Signal Processing Magazine, 41–56 (2008)
[40] Satti, A., Coyle, D., Prasad, G.: Spatio-spectral & temporal parameter searching using class correlation analysis and particle swarm optimization for a brain computer interface. In: Proc. of the 2009 IEEE Systems, Man and Cybernetics Conference, pp. 1731–1735 (2009)
[41] Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings IEEE International Conference on Neural Networks, vol. 1, pp. 1942–1948 (1995)
[42] Herman, P., Prasad, G., McGinnity, T.M., Coyle, D.: Comparative analysis of spectral approaches to feature extraction for EEG-based motor imagery classification. IEEE Transactions on Neural Systems and Rehabilitation Engineering 16(4), 317–326 (2008)
[43] Coyle, D., Prasad, G., McGinnity, T.M.: A time-frequency approach to feature extraction for a brain-computer interface with a comparative analysis of performance measures. EURASIP JASP, Trends in Brain-Computer Interfaces (special issue) 19, 3141–3151 (2005)
[44] Coyle, D., Prasad, G., McGinnity, T.M., Herman, P.: Estimating the predictability of EEG recorded over the motor cortex using information theoretic functionals. In: Proceedings of the 2nd International Brain-Computer Interface Workshop and Training Course, Biomedizinische Technik, pp. 43–44 (2004)
[45] Fraser, A.M.: Information and Entropy in Strange Attractors. IEEE Trans. on Info. Theory. 35(2), 245–262 (1989)
[46] Palus, M., Pecen, L., Pivka, D.: Estimating predictability: The redundancy and surrogate data method. Neural Network World 5(4), 537–552 (1995)
[47] Palus, M.: Testing for nonlinearity using redundancies: Quantitative and qualitative aspects. Physica D, 186–205 (1995)
[48] Williams, G.P.: Chaos Theory Tamed. Taylor and Francis, London (1997)
[49] Shannon, C.E., Weaver, W.: The mathematical theory of communication. University of Illinois Press (1963)
[50] Scott, D.W.: Multivariate Density Estimation: Theory, Practice and Visualisation. John Wiley and Sons, New York (1992)
[51] Chow, T.W.S., Huang, D.: Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information. IEEE Transactions on Neural Networks 16(1), 213–224 (2005)
[52] Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Chapman and Hall, London (1986)
[53] Davies, L., Gather, U.: The identification of multiple outliers. Journal of the American Statistical Association 88(423), 782–792 (1993)
[54] Zar, J.H.: Biostatistical Analysis, 4th edn., pp. 255–259. Upper Saddle River, New-Jersey (1999)
[55] Greene, J., D'Oliveira, M.: Learning to use statistical tests in psychology. Open University Press (1982)

[56] Satti, A., Guan, C., Coyle, D., Prasad, G.: A covariate shift minimisation method to alleviate non-stationarity effects for an adaptive brain-computer interface. In: 20th International Conference Pattern Recognition, August 23-26, pp. 105–108 (2010)
[57] Krusienski, D.J., Grosse-Wentrup, M., Galan, F., Coyle, D., Miller, K.J., Forney, E., Anderson, C.W.: Critical Issues in Brain Computer Interface Research. Journal of Neural Engineering 8, 025002 (8pp) (2011)
[58] Coyle, D., McGinnity, T.M., Prasad, G.: A multi-class brain-computer interface with SOFNN-based prediction preprocessing. In: IEEE World Congress on Computational Intelligence, pp. 3695–3702 (2008)
[59] Coyle, D., Prasad, G., McGinnity, T.M.: Improving information transfer rates of a brain-computer interface by self-organising fuzzy neural network-based multi-step-ahead time-series prediction. In: Proceedings of the 3rd IEEE Systems, Man and Cybernetics (UK&RI Chapter) Conference, pp. 230–235 (2004)