Medical & Biological Engineering & Computing (2021) 59:1643–1657
https://doi.org/10.1007/s11517-021-02386-y

ORIGINAL ARTICLE

Detection of microsleep states from the EEG: a comparison of feature reduction methods

Sudhanshu S. D. P. Ayyagari[1,2,3] · Richard D. Jones[1,2,4] · Stephen J. Weddell[1,2,3]

Received: 2 April 2020 / Accepted: 17 May 2021 / Published online: 17 July 2021

Abstract

Microsleeps are brief lapses in consciousness with complete suspension of performance. They are the cause of fatal accidents in many transport sectors requiring sustained attention, especially driving. A microsleep-warning device, using wireless EEG electrodes, could be used to rouse a user from an imminent microsleep. High-dimensional datasets, especially in EEG-based classification, present challenges as there are often a large number of potentially useful features for detecting the phenomenon of interest. Thus, it is often important to reduce the dimension of the original data prior to training the classifier. In this study, linear dimensionality reduction methods—principal component analysis (PCA) and probabilistic PCA (PPCA)—were compared with eight non-linear dimensionality reduction methods (kernel PCA, classical multi-dimensional scaling, isometric mapping, nearest neighbour estimation, stochastic neighbourhood embedding, autoencoder, stochastic proximity embedding, and Laplacian eigenmaps) on previously collected behavioural and EEG data from eight healthy non-sleep-deprived volunteers performing a 1D-visuomotor tracking task for 1 h. The effectiveness of the feature reduction algorithms was evaluated by visual inspection of class separation on 3D scatterplots, by trustworthiness scores, and by microsleep detection performance on a stacked-generalisation-based linear discriminant analysis (LDA) system estimating the microsleep/responsive state at 1 Hz based on the reduced features. On trustworthiness, PPCA outperformed PCA, but PCA outperformed all of the non-linear techniques. The trustworthiness score for each feature reduction method also correlated strongly with microsleep-state detection performance, providing strong validation of the ability of trustworthiness to estimate the relative effectiveness of feature reduction approaches, in terms of predicting performance, and ability to do so independently of the gold standard.

Keywords Microsleeps · Detection · Feature reduction · EEG · Classification

1 Introduction

Microsleeps are brief ($\lesssim$ 15 s) involuntary sleep–related lapses in consciousness, during which a person falls asleep momentarily and has a brief suspension of performance [1]. They are more likely when one is drowsy but can also occur when not sleep-deprived [1–3], and, for the most part, occur with no prior warning [4]. Occurrence of microsleeps when driving, can result in fatal accidents [5, 6]. Therefore, the detection of microsleeps, especially in subjects working in high-risk occupations, is very important to workplace safety.

A real-time microsleep-warning device, using wireless dry EEG scalp electrodes, in which an individual's state of responsiveness is monitored continuously, could be used to trigger an alert to rouse a user from an imminent microsleep, potentially avoiding a (multi-)fatal accident. Such a microsleep-prevention system, incorporating advanced signal processing and optimal machine learning classification algorithms for detecting/predicting microsleeps from the underlying EEG, would be of considerable importance in many high-risk occupations, such as commercial truck drivers, pilots, air-traffic controllers, and medical staff.

✉ Richard D. Jones
richard.jones@nzbri.org

[1] Department of Electrical and Computer Engineering, University of Canterbury, Christchurch, New Zealand

[2] Christchurch Neurotechnology Research Programme, Christchurch, New Zealand

[3] Computational Design and Adaptation, University of Canterbury, Christchurch, New Zealand

[4] New Zealand Brain Research Institute, Christchurch 8011, New Zealand

For classification/detection, multiple feature sets are extracted from the raw data and input to a classifier. Ideally, these features contain orthogonally-useful (additional) information for distinguishing between classes, while minimising irrelevant information. An EEG feature is defined as an arbitrary time-series extracted from a single EEG referential or bipolar derivation using a given signal processing algorithm [7]. A feature vector is a vector of all features for a particular epoch. Due to the high dimensionality of EEG datasets and the intrinsic tendency of most classifiers to overfit to the training data (especially non-linear classifiers), forming a high-performance classifier model becomes challenging. The exploration of new potentially useful features can produce very large feature vectors which are impractical for learning, as the space of the classifier model becomes too large to search [8].

In fields such as image processing, speech processing, and biomedical engineering, measured data vectors are usually high-dimensional [9]. Hence, there is often a need to substantially reduce the number of features so as to simplify the complexity of the problem and to pass only features containing at least partially orthogonal information to a classifier—i.e., the processes of feature selection and/or feature reduction.

Feature selection methods, such as filter and wrapper methods [10, 11], can be used to efficiently discard large numbers of irrelevant features but can have substantial drawbacks [12]. In contrast, feature/dimensionality reduction aims to transform high-dimensional data into a meaningful representation of reduced dimensions by extracting essential information from a dataset [13]. Ideally, the reduced representations should contain a minimum number of parameters needed to account for the observed properties of the data [14]—i.e., the intrinsic dimensionality of the data. The importance of feature reduction lies in its ability to mitigate the *curse of dimensionality* of high-dimensional datasets [9]. In most cases, high-dimensional datasets (e.g., multi-channel EEG) present many challenges, as not all of the features are necessary for optimal classification of the phenomena of interest. While certain non-linear (and computationally expensive) methods can construct predictive models with high accuracy from high-dimensional data, it is important in most applications to reduce the dimension of the original data prior to modelling the data [15].

Feature reduction schemes can be extremely useful in tasks such as classification, visualisation, and compression of high-dimensional data. Traditionally, feature reduction has been performed using standard linear techniques such as principal component analysis (PCA) [16, 17] and factor analysis [18]. PCA has proven an extremely useful tool for approximation of datasets with inherent low dimensionality but, when it comes to higher dimensional datasets, non-linear feature reduction techniques can yield better performance as they are better able to handle complex non-linear data [19, 20]. Consequently, in the past few decades, there has been a substantial increase in the number of non-linear techniques for feature reduction [20, 21].

Most real-world tasks involve datasets which are non-linear in nature. Therefore, using non-linear feature reduction techniques would appear likely to offer distinct advantages. This is supported by numerous studies which have shown that non-linear feature reduction techniques outperform their linear counterparts [22–24]. Motivated by the success of such non-linear feature reduction methods, our aim was to compare several advanced non-linear-dimensionality reduction techniques against classical PCA and probabilistic linear PCA (PPCA), which have been widely used as benchmarks to provide comparisons with the newer non-linear methods [25, 26]. We hypothesised that PCA and PPCA would not perform as well as the non-linear feature reduction methods. We chose eight non-linear feature reduction methods largely on the basis of their popularity in the literature [22–26] and what we considered to be their computational efficiency for real-time microsleep detection.

2 Methods

2.1 Feature reduction—linear techniques

2.1.1 Linear principal component analysis

In a mean square error sense, linear PCA is the ideal method as it provides a linear-dimension-reduction solution and is based on the covariance matrix of the variables. PCA has been used in a wide range of research areas as a non-parametric method for extracting relevant information from complex and often ill-defined datasets and is a well-established unsupervised dimensionality reduction technique [27].

PCA seeks to reduce the dimension of the data by finding orthogonal linear combinations—principal components (PCs)—of the original variables which accommodate the largest variance in an unsupervised manner. PCA has also been used as a baseline for comparison with non-linear-feature reduction methods due (i) its widespread usage and (ii) its use in previous microsleep detection studies [7, 28, 29].

Despite having proven an extremely useful tool, PCA has its limitations: (i) The covariance matrix is difficult to accurately evaluate, (ii) even the simplest invariance cannot be captured by PCA unless the training data explicitly provides this information, and (iii) directions in data maximising the variance do not always maximise information [20, 23, 30].

2.1.2 Linear probabilistic principal component analysis

Linear probabilistic PCA (PPCA) obtains a probabilistic formulation of PCA from a Gaussian latent variable model, which is closely related to statistical factor analysis; this allows a likelihood measure which in turn enables comparison with other probabilistic techniques, while facilitating statistical testing [31]. PPCA has been particularly used as a method to estimate the principal axes when any data vector has one or more missing values [23]. PPCA has been applied in many signal processing applications and, interestingly, is not based on a probability model but rather is determined through maximum-likelihood estimation of the parameters in a latent model closely related to factor analysis [31, 32].

PPCA is based on an isotropic error model. It seeks to relate a p-dimensional observation vector $\mathbf{y}$ to a corresponding k-dimensional vector of latent (or unobserved) variable $\mathbf{x}$, which is normal with zero-mean and covariance I(D). The relationship is established as follows:

$$\mathbf{y}^{\mathrm{T}} = \mathbf{W} \times \mathbf{x}^{\mathrm{T}} + \boldsymbol{\mu} + \varepsilon, \tag{1}$$

where $\mathbf{y}$ represents the column vector of observed variable, $\mathbf{x}$ represents the row vector of latent variables, $\boldsymbol{\mu}$ represents column vector and ε is defined as an isotropic error term which is Gaussian with zero-mean and covariance of $\upsilon \times \mathbf{I}(\mathbf{D})$, where υ is the residual variance. In the case of PPCA, k needs to be smaller for the rank of the residual variance to be greater than zero. Linear PCA, where the residual variance is zero, is the limiting case of PPCA. $\mathbf{W}$ is a matrix of observations relating to the latent and observed values.

Here, the values of $\mathbf{y}$ are conditionally independent and identically distributed, given the values of $\mathbf{x}$. Therefore, it is possible that the values of $\mathbf{x}$ can explain the correlations between values of $\mathbf{y}$ and error ε, and can also explain the variability unique to a particular element of $\mathbf{y}$. The marginal probability of the observation given model parameters can be expressed as follows:

$$\mathbf{y} = N\left(\boldsymbol{\mu}, \mathbf{W} \times \mathbf{W}^{\mathrm{T}} + \upsilon \times \mathbf{I}(\mathbf{D})\right). \tag{2}$$

Tipping and Bishop [31] suggested that there is no closed-form solution for both $\mathbf{W}$ and υ, and, therefore, estimates are determined by iterative maximisation of log-likelihood using an expectation maximisation algorithm. As such, PPCA has been proposed as a powerful alternative in several image processing, time-series prediction, and pattern recognition tasks.

2.2 Feature reduction—non-linear techniques

2.2.1 Kernel Principal component analysis

Kernel PCA (KPCA) is a reformulation of traditional linear PCA in a high-dimensional space that is constructed using a kernel function. KPCA achieves non-linear feature reduction through the use of kernel functions. KPCA with a linear kernel is the same as traditional PCA. Since KPCA is a kernel-based feature reduction method, the mapping function from the KPCA is highly reliant on the type of the kernel function being used. There are multiple variants for a non-linear kernel, such as Gaussian and polynomial [33].

The kernel matrix $\mathbf{K}$ is computed for the data points x_i and y_i, and the entries are defined as follows:

$$\mathbf{k}_{ij} = \mathrm{k}\left(x_i, x_j\right), \tag{3}$$

where κ represents the kernel function and κ can be any function that generates a positive-semi-definite kernel $\mathbf{K}$ [30]. Consequently, κ is centred with the following modification to the entries as follows:

$$\mathbf{k}_{ij} = \mathbf{k}_{ij} - \frac{1}{n}\sum_{\boldsymbol{il}} \mathbf{k}_{il} - \frac{1}{n}\sum_{\boldsymbol{jl}} \boldsymbol{k_{jl}} + \frac{1}{n^2}\sum_{\boldsymbol{lm}} \boldsymbol{k_{lm}}. \tag{4}$$

In KPCA, the centring operation corresponds to subtracting the mean of the features in traditional PCA. The kernel centring operation determines that the features in higher dimensional space are defined by the kernel function and contain a zero mean. Finally, the principal eigenvectors of the centred kernel matrix are computed by the following:

$$\boldsymbol{\alpha}_i = \frac{\mathbf{1}}{\sqrt{\boldsymbol{\lambda_i}}} \mathbf{v}_i, \tag{5}$$

where $\boldsymbol{\alpha}_i$ represents all the eigenvectors of the covariance matrix in the higher dimensional space and $\mathbf{v}_i$ represents the scaled versions of the eigenvectors of the kernel matrix. From this relationship, it is evident that the eigenvectors of the covariance matrix can be scaled versions of the eigenvectors of the kernel matrix.

To obtain low-dimensional data representation $\mathbf{Y}$ via, KPCA, the data is typically projected onto the eigenvectors of the covariance matrix $\boldsymbol{\alpha}_i$. The resulting $\mathbf{Y}$ is denoted as follows:

$$\mathbf{Y} = \sum_j \boldsymbol{\alpha_1} \mathbf{K}\left(x_{\boldsymbol{j}}, x\right), \sum_j \boldsymbol{\alpha_2} \mathbf{K}\left(x_{\boldsymbol{j}}, x\right) \ldots\ldots\ldots\ldots \sum_j \boldsymbol{\alpha_d} \mathbf{K}\left(x_{\mathbf{j}}, x\right). \tag{6}$$

2.2.2 Classical multi-dimensional scaling

Multi-dimensional scaling (MDS) is a widely accepted technique for data visualisation of EEG, fMRI, and other biomedical and biomolecular based analyses [34, 35]. MDS was proposed as a classical approach to the problem of finding underlying attributes or dimensions via a visual representation of the pattern of proximities (i.e., similarities or distances) among a set of objects [36]. The MDS algorithm represents a collection of non-linear techniques which can map high-dimensional data to a much lower dimensional representation whilst retaining the pairwise distances between the observable datapoints as much as possible [30]. MDS is often considered as an extension of the Sammon mapping problem [37].

MDS uses a stress function φ to express the quality of mapping, which is a measure of the error between the pairwise distances in both low- and high-dimensional representations (**Y** and **X**, respectively) of the observable data. The stress function is defined as follows:

$$\varphi(\mathbf{Y}) = \mathbf{X}\left(k\mathbf{x}_{\boldsymbol{i}}-\mathbf{x}_{j}k-k\mathbf{y}_{\boldsymbol{i}}-\mathbf{y}_{j}k\right)^{2}, \tag{7}$$

where $k\mathbf{x}_{\boldsymbol{i}}-\mathbf{x}_{j}k$ represents the Euclidean distance between the high-dimensional observations $\mathbf{x}_i$ and $\mathbf{x}_{j,}$ and $k\mathbf{y}_{\boldsymbol{i}}-\mathbf{y}_{j}k$ represents the Euclidean distance between the low-dimensional observations $\mathbf{y}_i$ and $\mathbf{y}_j$, and k represents the nearest neighbours.

2.2.3 Isometric mapping

Isometric mapping (Isomap) is one of the earliest approaches for manifold learning and is widely accepted as an extension for MDS or KPCA methods [38]. Isomap seeks a low-dimensional embedding which maintains geodesic distances between all points. The geodesic or curvilinear distance is the distance between two points measured over the manifold. Manifold learning algorithms are based on the idea that the dimensionality of many data sets is only artificially high). Isomap uses the same principles as the MDS algorithm. The key steps in Isomap are as follows:

1. Obtain a matrix of proximities (distances between points in a dataset)
2. Calculate the distance matrix via inner products
3. Carry out an Eigen-decomposition of the distance matrix to derive lower dimensional embedding

A substantial difference between Isomap and MDS is the way in which the distance matrix is constructed. In Isomap, the geodesic distances between observations $\mathbf{x}_i$ are computed by constructing a neighbourhood graph G, in which every datapoint $\mathbf{x}_j$ is connected with its k nearest neighbours $\mathbf{x}_{ij}$ in the dataset **X.**

The shortest path between two points in the graph forms a good estimate—i.e., an over-estimate of the geodesic distance between these two points, which can be computed via Dijkstra's shortest path algorithm [39]. In the Isomap algorithm, distances between points are considered as the weight of the shortest path in a point-graph. The pairwise geodesic distance matrix is formed from the geodesic distances between all observations in **X**. Finally, the low-dimensional representations $\mathbf{y}_i$ of the observations $\mathbf{x}_i$ in the low-dimensional space **Y** are computed by applying the MDS framework to obtain a distance matrix.

2.2.4 Nearest neighbour estimation

The nearest neighbour estimation (NNE) algorithm is based on the estimation of the number of neighbouring observations covered by a hypersphere of radius r. The NNE algorithm does not explicitly count the number of observations inside the hypersphere but computes the minimum radius r of the hypersphere necessary to cover k nearest neighbours [30]. The nearest neighbour estimator computes the following:

$$C(k)=\frac{1}{n}\sum T_k(x_i), \qquad \text{where C}=\begin{cases}1, & \text{if } \left|\left|x_i-x_j\right|\right| < r\\ 0, & \text{if } \left|\left|x_i-x_j\right|\right| > r\end{cases} \tag{8}$$

where T_k (x_i) signifies the radius of the smallest hypersphere with centre (x_i) which covers k neighbouring observations. The dimensionality of a dataset $\widehat{d}$ is estimated by the following:

$$\widehat{d} = \frac{\log\left(C\left(k_2-C\left(k_1\right)\right)\right)}{\log(k_2-k_1)}. \tag{9}$$

2.2.5 Stochastic neighbourhood embedding

Stochastic neighbourhood embedding (SNE) is a probabilistic approach in which features described by high-dimensional vectors or by pairwise dissimilarities are placed in a low-dimensional space with their neighbour identities preserved. In essence, SNE is only slightly different to MDS in terms of the distance measure that it uses as a minimalist cost function. In SNE, a Gaussian is centred on each object in the high-dimensional space and a few dissimilarities under this Gaussian are used to define a probability distribution over all of the potential neighbours of the features [40].

In SNE, the matrix **P** denotes the distribution of all the individual probabilities p_{ij} for all the observations $\mathbf{x}_i$ and $\mathbf{x}_j$ generated by the same Gaussian. SNE models the similarity of datapoint $\mathbf{x}_i$ to datapoint $\mathbf{x}_j$ as the conditional probability p_{ij}, that $\mathbf{x}_i$ would pick $\mathbf{x}_j$ as its neighbour if neighbours were picked in proportion to their probability density under a Gaussian centred at $\mathbf{x}_{i.}$

The probabilities p_{ij} are calculated using the Gaussian kernel function:

$$\mathrm{w}_{ij} = e^{-\frac{\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2}{2\sigma^2}}, \tag{10}$$

where w_{ij} represents the individual weights of the observations and σ represents the variance of the Gaussian. The algorithm then sets the coordinates of the low-dimensional representations $\mathbf{y}_i$.

Following this, the probabilities for the low-dimensional counterparts $\mathbf{y}_i$ and $\mathbf{y}_j$ of the high-dimensional datapoints $\mathbf{x}_i$ and $\mathbf{x}_j$ and a similar conditional probability q_{ij} can be generated by the same Gaussian computed and stored in the matrix **Q** which uses the same Gaussian kernel.

The SNE algorithm aims to minimise the difference between the probability distributions **P** and **Q**. Hinton and Roweis [40] stated that in a perfect low-dimensional representation of the data, the matrices **P** and **Q** are identical. In SNE, a natural cost function is the sum of the Kullback-Leibler divergences [41], which is the natural distance measure to measure the difference between two probability distributions, given by the following:

$$\varphi(\mathbf{Y})=\sum_{ij}\mathbf{P}_{ij}\log\frac{\mathbf{P}_{ij}}{\mathbf{Q}_{ij}}. \tag{11}$$

SNE aims to minimise the sum of Kullback-Leibler divergences by applying methods such as the gradient descent technique [40].

2.2.6 Autoencoder

Autoencoders are multi-layered and are a type of feed-forward neural network possessing an odd number of hidden layers [42, 43]. An autoencoder uses an unsupervised learning rule that applies backpropagation, setting the target values to be equal to the inputs $y_i = x_i$. Hence, autoencoder networks are trained to minimise the mean square error (MSE) between the input and the output of the network.

When a set of data is passed through an autoencoder, the network compresses (encodes) an input vector to fit a smaller representation and then tries to reconstruct (decode) the information back. The training algorithm aims to find the most efficient encoding representation for an input sequence. Autoencoders also try to convert a vector of n-dimensional space to an m-dimension, while retaining all of the necessary information and, at the same time, removing noise.

In order to train an autoencoder to model a non-linear mapping between the high-dimensional and low-dimensional data representation, sigmoid activation functions are typically used. Autoencoders with linear activation functions resemble PCA [32].

2.2.7 Stochastic proximity embedding

Stochastic proximity embedding (SPE) is considered an extension to multi-dimensional scaling. SPE runs an iterative algorithm to minimise the raw stress function ϕ of MDS:

$$\varphi(\mathbf{Y})=\sum_{ij}\left(d_{ij}-r_{ij}\right)^2, \tag{12}$$

where r_{ij} is the proximity between the high-dimensional data points $\mathbf{x}_i$ and $\mathbf{x}_i$, and d_{ij} is the Euclidean distance between their lower dimensional counterparts $\mathbf{y}_i$ and $\mathbf{y}_j$ in the current approximation of the embedded space [30].

The SPE algorithm updates the current estimate of the low-dimensional data representation. SPE also has a behaviour comparable to that of the isomap in that SPE can be readily applied to retain only distances in a neighbourhood graph G defined on the data by setting d_{ij} and r_{ij} to 0 if $(i, j) \notin$ G.

The updating in the SPE algorithm is performed using updated rules:

$$\mathbf{y}_i \leftarrow \mathbf{y}_i+\lambda\frac{r_{ij}-d_{ij}}{2d_{ij}+\varepsilon}\left(\mathbf{y}_i-\mathbf{y}_j\right), \tag{13}$$

$$\mathbf{y}_j \leftarrow \mathbf{y}_j+\lambda\frac{r_{ij}-d_{ij}}{2d_{ij}+\varepsilon}\left(\mathbf{y}_j-\mathbf{y}_i\right), \tag{14}$$

where λ is defined as a learning parameter and λ decreases with the number of iterations, and ε is a regularisation parameter.

2.2.8 Laplacian Eigenmaps

Lastly, the Laplacian Eigenmaps (LE) algorithm bears a resemblance to the Isomap in that it constructs a graph representation of all observations. The LE algorithm is based on the pairwise distance between the neighbours. LE computes a low-dimensional representation of the data in which the distances between a datapoint and its k nearest neighbours are minimised [30].

The LE algorithm begins by constructing a neighbourhood graph G in which every observation x_i is connected to its k nearest neighbours. For all the points $\mathbf{x}_i$ and $\mathbf{x}_j$ in the graph G that are connected by an edge, the weight of the edge is computed using a Gaussian kernel function leading to a sparse adjacency matrix **W** [44]. The cost function in the low-dimensional representations is minimised and represented as follows:

$$\varphi(\mathbf{Y}) = \sum_{ij}\left(\mathbf{y}_i-\mathbf{y}_j\right)^2 w_{ij}. \tag{15}$$

A degree matrix **M** and a graph Laplacian **L** are computed for a graph G which allows for formulating the minimisation problem as the Eigen problem [45]. The degree matrix **M** of **W** is a diagonal matrix, whose entries are the row sums of $\mathbf{w} = (m_{ii} = \sum_j w_{ij})$.

The graph Laplacian **L** is computed by **L** = **M** – **W**. It can be shown that the following holds

$$\varphi(\mathbf{Y}) = \sum_{ij}\left(\mathbf{y}_i-\mathbf{y}_j\right)^2 w_{ij} = 2\,\mathbf{Y}^{\mathrm{T}}\mathbf{L}\mathbf{Y}. \tag{16}$$

Therefore, minimising $\phi(\mathbf{Y})$ is proportional to minimising $\mathbf{Y}^{\mathbf{T}}\mathbf{L}\mathbf{Y}$. The low-dimensional data representation **Y** can thus be found by solving the generalised eigenvector problem as follows:

$$\mathbf{L}\upsilon=\boldsymbol{\lambda}\mathbf{M}\upsilon, \tag{17}$$

where ν is a vector that minimises the objective function.

2.3 Data

Subjects This study used previously collected behavioural and EEG data from 8 healthy male non-sleep-deprived volunteers, aged 18–36 years (mean = 26.5), performing a 1D-visuomotor tracking task for 2 sessions (at least 1 week apart) of 1-h duration [1, 7]. None of the 8 subjects had a current or previous neurological or sleep disorder and all had visual acuities of 6/9 (= 20/30) or better in each eye. All subjects considered that they had slept normally the previous night (mean =7.8 h, SD = 1.2 h, min = 5.1 h) and, hence, were considered non-sleep-deprived.

Continuous tracking task Subjects used a steering wheel (395 mm diameter, wheel-to-screen gain = 1.075 mm/deg) to control an arrow-shaped cursor located near the bottom of the screen in a 1D visuomotor tracking task. The eye-to-screen distance was 136 cm. A pseudo-random target (bandwidth 0.164 Hz, period 128 s), with an 8-s preview, scrolled down the screen at 21.8 mm/s. The target signal was generated by summing 21 sinusoids evenly spaced at 0.00781-Hz intervals and with random phases [7]. Subjects were instructed to keep the point of the arrow (which could only move horizontally) as close as possible to the target waveform. Subjects performed the task continuously for 1 h and undertook two sessions, 1 week apart.

Neurophysiological and behavioural measures During this task, EEG was recorded from electrodes at 16 scalp locations, band-pass filtered (0.5–100 Hz), and digitised at 256 Hz with a 16-bit A-D converter. Eye-blink artefacts were removed via ICA followed by notch filtering at 50 Hz to remove mains activity. The mean and standard deviation of the first 2 min (baseline) of the signal were calculated and the signal was transformed into z-scores relative to the baseline of the signal. Epochs of 2.0 s containing samples with an absolute z-score > 3.0 were rejected as artefacts and excluded from analysis in the signal processing algorithms. Bipolar derivations were used to calculate power spectra: Fp1–F3, Fp1–F7, Fp2–F4, Fp2–F8, F3–C3, F4–C4, F7–T3, F8–T4, T3–T5, C3–P3, P3–O1, T5-O1, C4–P4, T4–T6, P4–O2, and T6–O2 [7]. Bipolar derivations were chosen over referential due to minimisation of common-mode artefacts.

Facial video was also recorded during the tracking sessions. Video-based microsleeps were primarily identified by prolonged eye-lid closure. Video-based transitions were rated at 1/s [7].

Generation of the gold standard Validation of training and testing data required binary-labelled microsleep (‘1’) and responsiveness (‘0’) states. This behavioural gold standard was created by human experts from tracking and video measures and was used to estimate feature reduction performance.

Lapses in tracking performance are most obvious when the response cursor simply stops moving for an extended period while the target is moving (i.e., ‘flat spots’) or when the tracking response is non-coherent with the target. Only flat spots were included in an intentionally conservative analysis, as lapses in the second category are difficult to identify with confidence. Flat spots occurring when the target velocity was approximately zero (at turning points) were not counted, as at these times the subject could track adequately without moving the response cursor.

Spectral features A 2.0-s window with a 1.0-s overlap (50%) between successive windows was used for all signal processing algorithms. The sliding process generated feature samples at a rate of 1 Hz, resulting in 3600-element-long feature vectors for a 1-h recording. The 2.0-s window was chosen to obtain a reasonable degree of spectral resolution (where appropriate) and the overlap of 1.0-s was chosen to ensure reasonable temporal resolution (an estimate every second) for the features. This was important since a key requirement of the desired microsleep detection system was its ability to detect short microsleeps (~1 s).

Data in each 2.0-s epoch were detrended to remove any DC shifts and the spectrum was estimated using a 40th-order Burg model [46]. Thirty-four spectral features, comprising 13 spectral power (SP), 12 normalised spectral power (NSP), and 9 power ratio (PR) features (Table 1), were calculated for each of the 16 derivations, giving a total of 544 spectral features.

Table 1 Spectral features calculated from each EEG derivation

Feature	Frequency band
Mean spectral power [a]	
Delta (δ)	1.0–4.5 Hz
Theta (θ)	4.5–8.0 Hz
Alpha 1 (α1)	8.0–10.5 Hz
Alpha 2 (α2)	10.5–12.5 Hz
Alpha (α)	8.0–12.5 Hz
Beta 1 (β1)	12.5–15.0 Hz
Beta 2 (β2)	15.0–25.0 Hz
Beta (β)	12.5–25.0 Hz
Gamma 1 (γ1)	25.0–35.0 Hz
Gamma 2 (γ2)	35.0–45.0 Hz
Gamma (γ)	25.0–45.0 Hz
High	>45.0 Hz
Overall	0.1–100 Hz
Spectral power ratios [b]	
θ/β, θ/α, α/β, δ/θ, α/δ, β/δ, β2/α, β1/β2	–

[a] Absolute values and normalised values

[b] Absolute values only

Classification of the gold standard and performance analysis A stacked-generalisation-based linear discriminant analysis (LDA) model [7, 47] was used to form the microsleep detection system capable of classifying the microsleep and responsiveness states from the generated gold standard data. Stacking determines how best to combine base models via an additional meta-learner algorithm [7, 47]. The performance of this microsleep detection system was calculated in terms of ability to detect the microsleep state in consecutive 1-s epochs. Classification performance was determined by leave-one-subject-out cross-validation corresponding to the 8 subjects on session 1. Pearson correlation (phi) was considered the primary performance metric because of it being largely independent of the substantial class imbalance ratios (number of microsleep states vs. number of responsive states, ranging 1:813–1:2.26) and due to it primarily representing a combination of both sensitivity and precision.

2.4 Comparison of feature reduction methods

2.4.1 Visual inspection of class separation

Data visualisation techniques can guide a feature reduction algorithm in identifying meaningful coordinate projections for datasets with high dimensionality [48]. Visual inspection as an information retrieval task is one of the core ingredients of exploratory data analysis as it provides a detailed understanding of the underlying mathematical principles of the machine learning algorithms. To this end, different exploratory techniques have been proposed to interpret the essential structural characteristics of a dataset.

Some of the classical approaches used for visual inspection include interpretation of data using tree maps, graph-based visualisation, scatterplots, parallel coordinate maps, and outlier maps [49]. In this research, 3-D scatterplot analysis was performed on each of the ten feature reduction algorithms to visualise the degree of class separation achieved. Further justification behind the implementation of this scatterplot analysis was to determine the efficacy of the underlying feature reduction algorithm and to select inputs for a linear discriminant analysis classifier.

2.4.2 Trustworthiness

The trustworthiness metric was specially proposed for low-dimensional embedding. Venna and Kaski [22] proposed trustworthiness as a measure of feature reduction performance. Trustworthiness measures the proportion of points that are too close together in low-dimensional space [20, 22]. The trustworthiness $T(k)$ measure is defined as follows:

$$\mathrm{T}_{\mathrm{trust}}(k) = 1 - A(k)\sum_{i=1}^{N} \sum_{j \in \mathrm{U}_k(i)} \left(r(i,j) - k\right), \tag{18}$$

where i, j are low-dimensional datapoints, and $r(i, j)$ represents the rank of a low-dimensional datapoint j to the pairwise distances between the low-dimensional datapoints. $U_{(k)}(i)$ indicates the set of points that are amongst the k nearest neighbours in the lower dimensional datapoint i.

The term $A(k)$ normalises the measure to between 0 and 1. The error gets its maximum value when the ranks in the input and output space are reversed. The scaling term can then be found by considering the maximum error in each data point's neighbourhood as the sum of the k last ranks (minus the neighbourhood size k). Thus, the scaling term for N data samples becomes

$$A(k) = \begin{cases} \dfrac{2}{Nk(2N-3k-1)}, \text{if } k < \dfrac{N}{2} \\ \dfrac{2}{N(N-k)(N-k-1)}, \text{if } k \geq \dfrac{N}{2} \end{cases} \tag{19}$$

Essentially, the trustworthiness measure is closely related to the precision measure for the case where the objects are ranked based on their relevance [20]. The neighbourhood size k in the trustworthiness measure defines the number of items retrieved. There is a distinct advantage over measuring the reconstruction errors when evaluating the trustworthiness and generalisation errors, because a high reconstruction error does not necessarily imply that the dimensionality reduction technique performed poorly [22, 30].

2.4.3 Validation via classification performance

The reduced features from each of the feature reduction methods were separately used to train an LDA classifier running a level-1 stacking framework, as also used by Peiris et al. [7]. The LDA classifier was implemented on each of the 8 subjects, on session 1 only, to:

(1) Detect the occurrence of microsleep states in 1-s epochs
(2) Compare leave-one-subject-out cross-validation classification performances for each of the 8 subjects on each of the 10 feature reduction algorithms in terms of phi coefficient and AUC-ROC
(3) Compare the mean phi performances with the mean trustworthiness scores for each of the 10 feature reduction methods in terms of Pearson correlation

The mean phi correlation (φ) and area under the receiver-operator characteristic curve (AUC-ROC) were used as the primary performance metrics because of their reasonable independence from class distributions, in addition to being, from our experience, the best integrated measures of the other performance metrics for unbalanced data.

2.5 Statistics

The non-parametric Wilcoxon signed-rank test [50] was used for all paired comparisons between trustworthiness metrics and between LDA classifier phi correlations for the 10 feature reduction methods. Because of the small number of comparisons, no correction was made for multiple comparisons. All comparisons were 2-sided.

3 Results

3.1 Spectral features

From feature reduction via PCA in particular, we could identify the more important features making up the low-dimensional meta-features. The first principal components (PCs), PC1–PC7, indicated highest weights for high-frequency gamma features (particularly >45 Hz). In contrast, the alpha bands (alpha, alpha 1 and 2) were the highest-weighted features in PCs 10–20, and PCs 8 and 9 had a mixture of high and mid-ranged frequency features.

3.2 Visual inspection of class separation

Ten 3-D scatterplots for a typical subject (subject 1) are shown in Fig. 1 (linear) and Fig. 2 (non-linear).

3.2.1 Principal component analysis

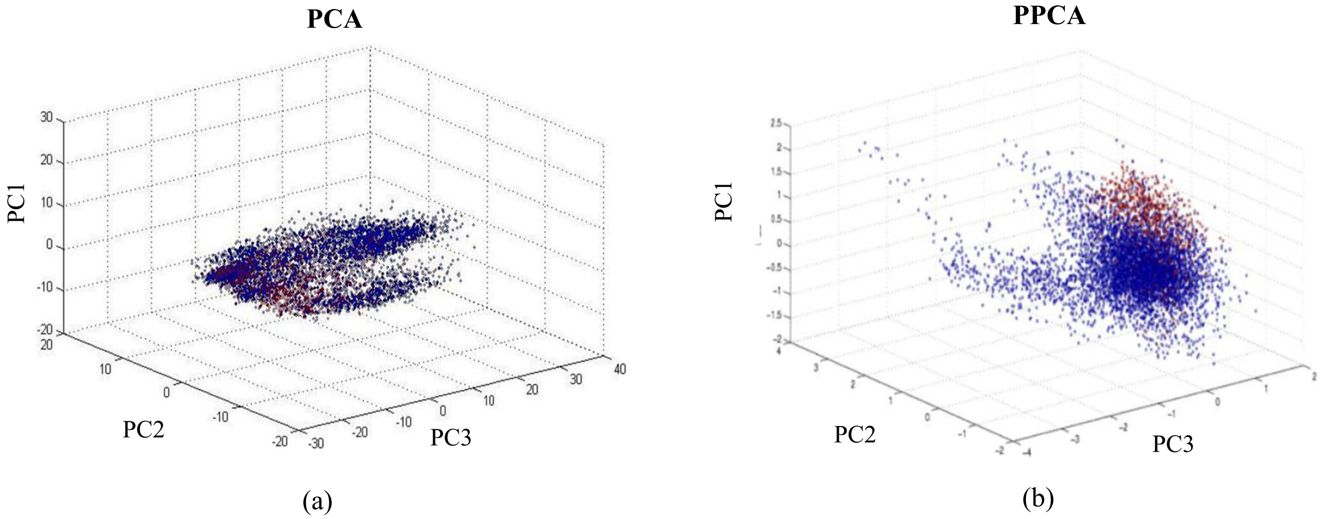
Figure 1(a) depicts the 3-D scatter pattern of the meta-features generated by PCA from the 544 spectral features for subject 1. The spectral features containing events are identified with red circles (microsleep states) and the features without an event are identified with blue circles (responsive state). This figure not only depicts how the meta-features are interpreted from a processed dataset using PCA but also depicts the innate intricacies of the current microsleep classification problem on the whole. As evident from the plot, the meta-features corresponding to both classes (microsleep and responsive) are clustered on top of each other, emphasising the challenge of this classification problem. The corresponding trustworthiness score reported was $T_{N=8} = 0.49$.

3.2.2 Probabilistic principal component analysis

Figure 1(b) depicts the application of the PPCA algorithm to subject 1. It is clear that PPCA is superior to PCA and KPCA. Furthermore, PPCA reported the best trustworthiness scores with the highest mean $T_{N=8} = 0.54$. There is also compelling evidence that PPCA-based approaches work well on time-series and pattern recognition problems [31].

3.2.3 Kernel principal component analysis

The 3-D scatter plot in Fig. 2(a) demonstrates the application of the non-linear KPCA technique using the Gaussian kernel on subject 1. The three most significant meta-features (PCs) are plotted against the *x, y*, and *z*-axes. Although KPCA is considered to give superior feature reduction to traditional PCA (due to its kernel functions), visual comparison of the scatter plots in Fig. 1 do not reveal any obvious new insights or improvement in separation between the two classes. Both scatter plots indicate that any classifier, be it linear or non-linear, would struggle to achieve ideal separation between the two classes. Notwithstanding, the scatter pattern in



(a) (b)

Fig. 1 Visualising the class distributions and separations of the linear feature reduction algorithms investigated on Subject 1. The three axes represent the top 3 PCs and/or meta-features 1–3. Blue circles represent the class 'responsive state' and red circles the 'microsleep state'. Subfigures **a** and **b** depict the first 3 PCs of the 50 reduced meta-features from the PCA and PPCA schemes

KPCA is a little less disarranged to that of PCA in Fig. 1(a), together with a moderate trustworthiness score $T_{N = 8} = 0.40$.

3.2.4 Classical multi-dimensional scaling

Figure 2(b) depicts the 3D scatter plot of the MDS approach on the microsleep detection problem for subject 1. Based on the visualisation of the MDS algorithm, it is inconclusive to determine whether or not this approach could offer a better separation between the classes. Although MDS works by preserving distances between points in the data, the individual clusters of each class appear to be close to one another.

A major disadvantage of this approach is that it is based on Euclidean distances and does not take into account the distribution of the neighbouring observations. This was also demonstrated on the current dataset with a low mean trustworthiness score of $T_{N=8} = 0.15$.

3.2.5 Isometric mapping

Figure 2(c) depicts the isomap application of the microsleep detection problem on subject 1. From the 3D scatter plot, it can be seen that performance of the Isomap algorithm is very poor. A possible reason behind this is that the Isomap algorithm failed to attain a topological stability due to being riddled with several erroneous connections while computing the neighbourhood graph G. The corresponding trustworthiness score was $T_{N = 8} = 0.10$, the lowest mean trustworthiness score of all 10 of the algorithms investigated.

3.2.6 Nearest neighbour estimation

Figure 2(d) depicts NNE applied to subject 1. From the scatterplot it can be seen that the NNE algorithm was able to reasonably distinguish and separate both classes, albeit with some errors, and do so less than PCA (Fig. 1(a)). The trustworthiness score for the NNE-based algorithm was $T_{N=8} =$ 0.40.

3.2.7 Stochastic neighbourhood embedding

Figure 2(e) represents the application of SNE to subject 1. The 3D scatter plot indicates that, due to the probabilities, distributions of the SNE majority of the class distributions are spread around in the corner of the image. This led to a reasonably high trustworthiness $T_{N=8} = 0.38$.

3.2.8 Autoencoder

Figure 2(f) depicts the 3D cluster plot from the application of the autoencoder schemes to subject 1. It is evident from the class distribution in the cluster pattern that the autoencoder-based-neural network would not perform well, which may be due to overfitting within the data. This is supported by a low mean trustworthiness score of $T_{N=8} = 0.12$.

3.2.9 Stochastic proximity embedding

Figure 2(g) provides an illustration of the 3D scatter plot from the application of SPE to subject 1. There was a reasonable mean trustworthiness score of $T_{N = 8} = 0.38$. SPE is computationally inexpensive and can accommodate a large number of iterations for updating its embedded coordinates.

3.2.10 Laplacian Eigenmaps

Figure 2(h) represents the 3D cluster plot of the LE algorithm on subject 1. The visualisation of the scatter plot depicts results which are inconclusive as the class separation is unevenly spread in the plane. The mean trustworthiness score of $T_{N = 8} = 0.15$ is low.

3.3 Trustworthiness

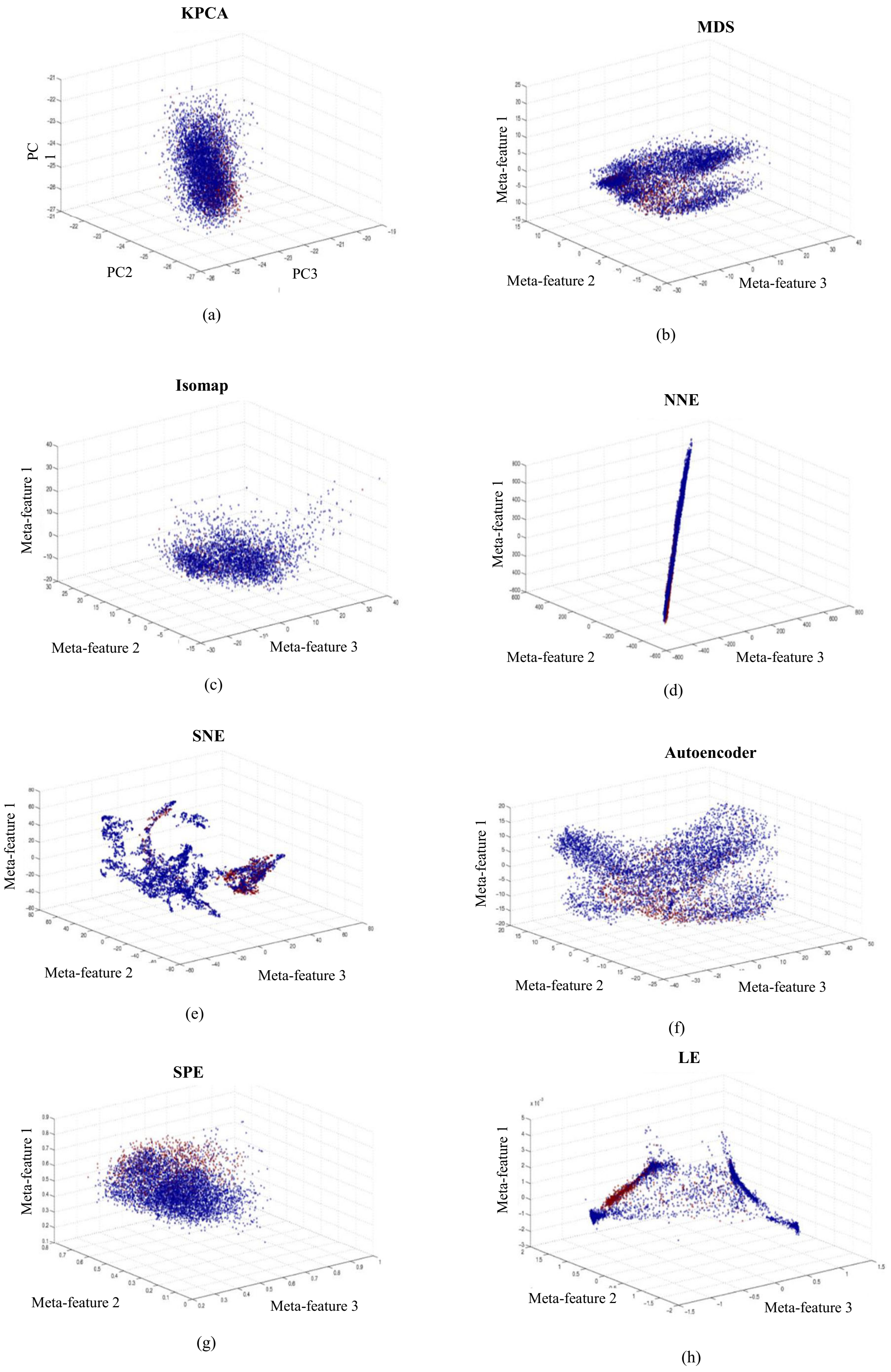
The trustworthiness scores of the 10 feature reduction approaches for each of the 8 subjects are shown in Table 2. The 3 PCA-based methods had superior mean T scores with PPCA having the highest, followed by standard PCA. Other methods with reasonable mean T values were NNE, SNE, and SPE. Despite being neighbourhood-based approaches, Isomap, and autoencoder performed poorly, both subjectively by way of their 3D scatter patterns and quantitatively in their trustworthiness scores.

It is also notable that the both the linear feature reduction methods PPCA and PCA outperformed all the other non-linear methods ($p < 0.012$). PPCA was the best feature reduction method, outperforming both PCA ($p = 0.021$) and all of the non-linear methods. However, PCA also outperformed all of the other non-linear techniques ($p < 0.012$) on the microsleep dataset, despite the ability of the other non-linear techniques to learn the structure of complex non-linear manifolds.

3.4 Validation of the feature reduction methods using a linear discriminants-based classifier

Table 3 provides a summary of leave-one-subject-out system performance for an LDA-based microsleep detector using each of the 10 feature reduction schemes.

Linear detectors based on PPCA and PCA provided the best generalisation performances with $\varphi = 0.42$ and $\varphi =$ 0.40, respectively, followed by the detectors based on NNE ($\varphi = 0.37$), KPCA ($\varphi = 0.36$), SNE and SPE ($\varphi = 0.34$). The high performances observed in terms of φ were also confirmed by observing largest mean AUC-ROC values as shown in Fig. 3. The highest mean AUC-ROC scores were seen on the PCA-based techniques, with PPCA meta-features at 0.91,


KPCA
PC 1
PC2
PC3
MDS
Meta-feature 1
Meta-feature 2
Meta-feature 3
Isomap
Meta-feature 1
Meta-feature 2
Meta-feature 3
NNE
Meta-feature 1
Meta-feature 2
Meta-feature 3
SNE
Meta-feature 1
Meta-feature 2
Meta-feature 3
Autoencoder
Meta-feature 1
Meta-feature 2
Meta-feature 3
SPE
Meta-feature 1
Meta-feature 2
Meta-feature 3
LE
Meta-feature 1
Meta-feature 2
Meta-feature 3

(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

◀ **Fig. 2** Visualising the class distributions and separations of the non-linear feature reduction algorithms investigated on Subject 1. The three axes represent the top 3 PCs or meta-features 1–3. Blue circles represent the class responsive state and red circles the microsleep state. Subfigure **a** depicts the first 3 PCs of the 50 reduced meta-features from KPCA. Subfigures **b**, **c**, **d**, and **g** depict the top 3 meta-features from the 50 reduced features of the MDS, Isomap, NNE and SPE schemes. Subfigures **e**, **f**, and **h** depict the top 3 meta-features from the 10, 40 and 60 reduced features of the SNE, Autoencoder, and LE schemes, respectively

followed by PCA meta-features at 0.88 and KPCA meta-features at 0.84. The lowest mean AUC-ROC scores were seen on the autoencoder meta-features with AUC-ROC = 0.70 and Isomap with AUC-ROC = 0.71, respectively. The neighbourhood-based methods of NNE, SNE, and SPE had AUC-ROC scores marginally higher than the non-linear parametric Isomap and autoencoder and Laplacian Eigenmaps schemes with AUC-ROC scores at 0.81, 0.79, and 0.78, respectively.

PPCA was superior to PCA in terms of trustworthiness ($p = 0.021$) but not in terms of phi ($p = 0.11$). Also, all of the 8 non-linear techniques were worse than PCA in terms of trustworthiness ($p < 0.017$).

Furthermore, the mean trustworthiness score for each of the feature reduction methods correlated strongly ($r = 0.992$) with the mean phi for microsleep-state detection over the 8 subjects. This provides strong validation of the ability of trustworthiness to (i) estimate the relative effectiveness of feature reduction approaches in terms of ability to predict performance of an LDA classifier and (ii) do so independent of the gold standard. Figure 4 shows the linear correlation between the matched pairs mean trustworthiness scores and mean phi correlation.

4 Discussion

We have determined and compared the effectiveness of two linear and eight non-linear techniques on their ability to optimally reduce EEG-based features and separate microsleep versus responsive states for input to a classifier for EEG-based microsleep detection. We have shown that, for EEG-based microsleep detection, the linear probabilistic PCA scheme is superior for feature reduction over all other methods. We have also shown that the two linear feature reduction techniques, PCA and probabilistic PCA, were able to outperform all of the eight non-linear methods, despite the ability of the latter to learn the structure of complex non-linear manifolds. Furthermore, despite being neighbourhood-based approaches, Isomap and autoencoder performed particularly poorly, both subjectively by way of their 3D scatter patterns and quantitatively in their trustworthiness scores. In terms of trustworthiness, probabilistic PCA outperformed the other linear method, PCA, but PCA, in turn, outperformed all of the other non-linear techniques. It was also demonstrated that the linear detectors based on PPCA and PCA feature sets provided the best generalisation performances. We also demonstrated a very high correlation between mean trustworthiness scores for each of the feature reduction methods and their microsleep-state detection performance. This provides strong validation of the ability of trustworthiness to (i) estimate the relative effectiveness of feature reduction approaches in terms of ability to predict performance of an LDA classifier and (ii) do so independent of the gold standard.

We have provided a comparative study of traditional linear PCA-based and non-linear techniques for feature reduction used in microsleep detection. Our results indicate that the non-linear techniques for dimensionality reduction are, despite their large variance of the parameters, not yet capable of outperforming traditional PCA-based methods (PCA and probabilistic PCA). A possible reason for the relatively low detector performances and trustworthiness scores of the non-linear feature reduction methods may be attributed to the overfitting of the training data with random noise. In general, our results indicate that overfitting was more likely associated with the non-linear and non-parametric models which had more flexibility when learning the target function [51].

Table 2 Trustworthiness scores of the feature reduction algorithms evaluated

Subject	PCA	PPCA	KPCA	MDS	Isomap	NNE	SNE	Autoencoder	SPE	LE
1	0.74	0.89	0.61	0.29	0.13	0.63	0.52	0.10	0.56	0.29
2	0.62	0.66	0.57	0.17	0.07	0.44	0.46	0.19	0.40	0.17
3	0.08	0.11	0.03	0.01	0.00	0.02	0.03	0.00	0.04	0.01
4	0.67	0.74	0.49	0.11	0.21	0.54	0.41	0.11	0.51	0.11
5	0.32	0.33	0.23	0.09	0.01	0.30	0.26	0.09	0.21	0.09
6	0.51	0.50	0.44	0.14	0.1	0.47	0.34	0.11	0.33	0.14
7	0.33	0.37	0.27	0.17	0.04	0.27	0.11	0.17	0.35	0.17
8	0.72	0.74	0.60	0.22	0.22	0.55	0.62	0.21	0.68	0.22
Mean	0.49	0.54	0.40	0.15	0.10	0.40	0.34	0.12	0.38	0.15

Table 3 Leave-one-subject-out performance (phi coefficient) of LDA-based classifier on the feature reduction algorithms

Subject	PCA	PPCA	KPCA	MDS	Isomap	NNE	SNE	Autoencoder	SPE	LE
1	0.78	0.83	0.66	0.57	0.50	0.73	0.64	0.59	0.63	0.48
2	0.57	0.60	0.50	0.50	0.33	0.54	0.53	0.39	0.51	0.47
3	0.09	0.09	0.04	0.04	0.02	0.03	0.03	0.04	0.03	0.02
4	0.13	0.14	0.10	0.07	0.07	0.09	0.10	0.09	0.41	0.04
5	0.56	0.53	0.41	0.34	0.29	0.48	0.51	0.29	0.26	0.39
6	0.49	0.53	0.43	0.29	0.34	0.47	0.40	0.24	0.34	0.34
7	0.20	0.24	0.35	0.13	0.15	0.27	0.18	0.17	0.11	0.13
8	0.41	0.40	0.38	0.22	0.24	0.35	0.36	0.29	0.62	0.38
Mean	0.40	0.42	0.36	0.27	0.24	0.37	0.34	0.26	0.34	0.28

Furthermore, since the behavioural gold standard was human-rated, errors/noise in the gold standard rating process are more than likely to have reduced the classifier performances and increased likelihood of overfitting of the non-linear models. Other limitations in the current study include the small dataset size (in terms of number of subjects), the variable quality of EEG features between subjects, and the features being specific to microsleeps. Thus, while our conclusions on the relative performance of several feature reduction techniques are compelling, we cannot, and are not, contending that they would necessarily apply to larger datasets and/or other transient events in the EEG.

Notwithstanding, the value of this research has been well demonstrated in the quantification of characteristics of microsleeps and EEG-based detection of microsleeps, with high temporal resolution [7, 28, 29, 52]. Despite the findings presented in this research, considerable future work in the field of EEG-based microsleep detection remains. As such, future work will look into development of linear and non-linear techniques for dimensionality reduction that (i) do not suffer from trivial optimal solutions and (ii) do not rely on neighbourhood graphs to model the (local) structure of the data manifold. We are also exploring reservoir-computing approaches to microsleep detection [53] and deep convolutional neural networks [54] due to the latter's intrinsic advantage in not needing explicit dimensionality reduction; of course, this advantage does not necessarily confer superior classification performance [55, 56].

Future studies could also explore enhancements on the signal processing front, for example analysis on the higher order spectra, which might make a substantial contribution towards improved lapse detection system. Higher order spectra have been shown to be of value in other EEG-based classification models, such as in the detection of epileptic activity in the EEG [57].

An interesting observation is that the first principal components from PCA indicated high weights for high-frequency gamma features. As EEG gamma and EMG spectra have a substantial overlap, and as frontal EMG tends to reduce on going from wake to sleep [58, 59], this raises the possibility

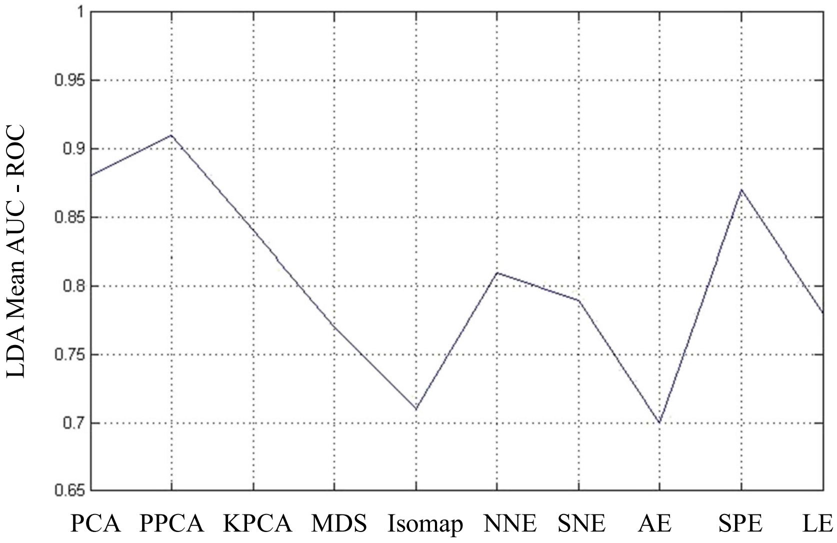


Fig. 3 LDA-based classifier performance of the feature reduction methods in terms of mean AUC-ROC

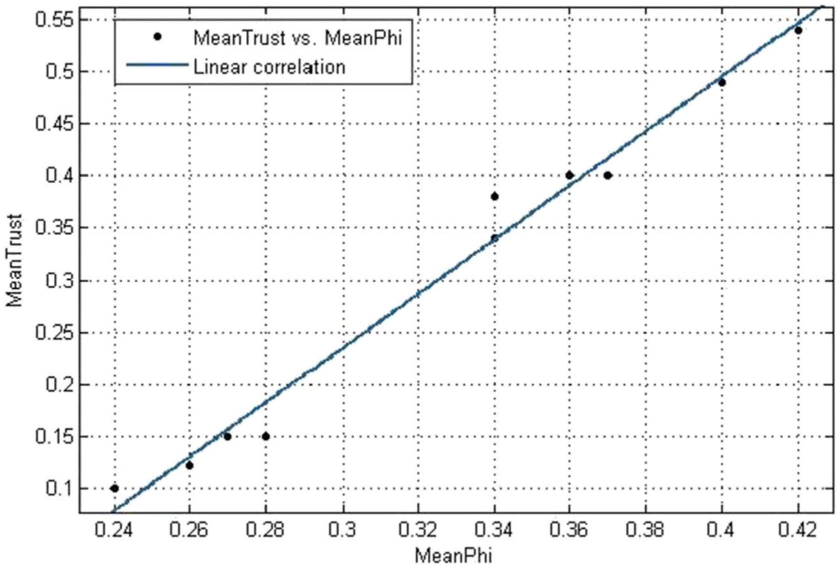


Fig. 4 Plot of mean trustworthiness of each of the 10 feature reduction methods versus mean phi of microsleep-state detection by an LDA classifier across the 8 subjects

that changes in frontal EMG power this raises the possibility that changes in frontal EMG power, particularly from forehead (frontalis) and eye closure (orbicularis oculi) muscles, may contribute to EEG-based detection of microsleep states.

5 Conclusion

Towards the development of an EEG-based system for detection and warning of microsleeps—the cause of many fatal accidents, especially in transport sectors—we have explored ten feature reduction techniques in recognition of the importance of the feature reduction step prior to training a detection classifier. From visual inspection of 3D scatterplots, trustworthiness scores, and microsleep detection performance, we have demonstrated that PPCA was not only superior to PCA but also that PCA was superior to all eight non-linear feature reduction techniques; thus, negating our hypothesis on the expected superiority of the non-linear feature reduction approaches. Furthermore, we demonstrated that unsupervised trustworthiness scores strongly correlate with gold standard–supervised microsleep-state detection performance, hence validating the ability of trustworthiness to estimate the relative effectiveness of feature reduction approaches and to do so independent of the gold standard.

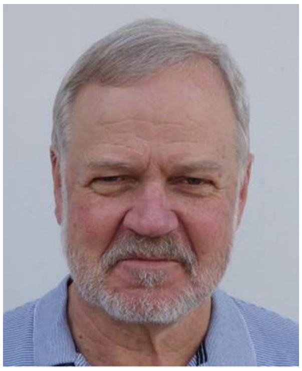
References

1. Peiris MTR, Jones RD, Davidson PR, Carroll GJ, Bones PJ (2006) Frequent lapses of responsiveness during an extended visuomotor tracking task in non-sleep-deprived subjects. J Sleep Res 15:291–300
2. Innes CRH, Poudel GR, Jones RD (2013) Efficient and regular patterns of nighttime sleep are related to increased vulnerability to microsleeps following a single night of sleep restriction. Chronobiol Int 30:1187–1196
3. Poudel GR, Innes CRH, Bones PJ, Watts R, Jones RD (2014) Losing the struggle to stay awake: divergent thalamic and cortical activity during microsleeps. Hum Brain Mapp 35:257–269
4. Herrmann US, Hess CW, Guggisberg AG, Roth C, Gugger M, Mathis J (2010) Sleepiness is not always perceived before falling asleep in healthy, sleep-deprived subjects. Sleep Med 11:747–751
5. Akerstedt T (2000) Consensus statement: fatigue and accidents in transport operations. J Sleep Res 9:395
6. Dingus TA, Klauer SG, Neale VL, Petersen A, Lee SE, Sudweeks J, Perez MA, Hankey J, Ramsey D, Gupta S, Bucher C, Doerzaph ZR, Jermeland J, and Knipling RR (2006) The 100-car naturalistic driving study. Phase II - results of the 100-car field experiment.: National Highway Traffic Safety Administration (NHTSA), US Department of Transportation
7. Peiris MTR, Davidson PR, Bones PJ, Jones RD (2011) Detection of lapses in responsiveness from the EEG. J Neural Eng 8(016003):1–15
8. Garrett D, Peterson DA, Anderson CW, Thaut MH (2003) Comparison of linear, nonlinear, and feature selection methods for EEG signal classification. IEEE Trans Neural Syst Rehab Eng 11:141–144
9. Jimenez LO, Landgrebe DA (1998) Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data. IEEE Trans Syst Man Cybern C Appl Rev 28:39–54
10. Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell Med 97:273–324
11. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
12. Domingos P (2012) A few useful things to know about machine learning. Commun ACM 55:78–87
13. Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55:119–139
14. Fukunaga K (1990) Introduction to statistical pattern recognition, 2nd edn. Academic Press, New York
15. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Elements 1:337–387

16. Pearson K (1901) LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine Series 6 2: 559–572
17. Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Edu Psychol 24:417–441
18. Spearman C (1904) "General intelligence", objectively determined and measured. Am J Psychol 15:201–292
19. Burges CJC (2005) Geometric methods for feature extraction and dimensional reduction. In: Maimon O, Rokach L (eds) Data Mining and Knowledge Discovery Handbook, 2005. Springer, Boston, MA, pp 59–91
20. Venna J (2007) Dimensionality reduction for visual exploration of similarity structures. PhD Thesis Computer Science and Engineering, Aalto University
21. Saul L, Weinberger KQ, Ham JH, Sha F (2006) Spectral methods for dimensionality reduction. In: O. Chapelle, B. Scholkopf, and A. Zien, (Eds) Semi-supervised learning. MIT Press Scholarship Online 2006:293–306
22. Venna J, Kaski S (2006) Local multidimensional scaling. Neural Netw 19:889–899
23. Roweis ST (1998) EM algorithms for PCA and SPCA. Proc Conf Adv Neural Inf Process Syst 10:626–632
24. van der Maaten LJP, Postma EO, and van den Herik HJ (2008) Dimensionality reduction: a comparative review: Tilburg University Technical Report. p. 1-35
25. Orsenigo C, Vercellis C (2013) Linear versus nonlinear dimensionality reduction for banks' credit rating prediction. Knowl-Based Syst 47:14–22
26. Sumithra VS, Surendran S (2015) A review of various linear and non linear dimensionality reduction techniques. Int J Comput Sci Inf Technol 6:2354–2360
27. Jolliffe IT (2002) Principal component analysis. J Am Stat Assoc 98:487
28. Davidson PR, Jones RD, Peiris MTR (2007) EEG-based lapse detection with high temporal resolution. IEEE Trans Biomed Eng 54: 832–839
29. Ayyagari SSDP, Jones RD, Weddell S (2015) Optimized echo state networks with leaky integrator neurons for EEG-based microsleep detection. Conf Proc IEEE Eng Med Biol Soc 37:3775–3778
30. van der Maaten LJP (2009) Feature extraction from visual data. PhD Thesis, Tilburg University
31. Tipping ME, Bishop CM (1999) Probabilistic principal component analysis. J R Stat Soc Ser B Stat Methodol 61:611–622
32. Kung SY, Diamantaras KI, Taur JS (1994) Adaptive principal component extraction (APEX) and applications. IEEE Trans Sig Process 42:1202–1217
33. Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge University Press, New York, NY, USA
34. Tagaris GA, Richter W, Kim S, Pellizer G, Andersen P, Ugurbil K, Georgopoulos AP (1998) Functional magnetic resonance imaging of mental rotation and memory scanning: a multidimensional scaling analysis of brain activation patterns. Brain Res Rev 26:106–112
35. Venkatarajan M, Braun W (2001) New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties. J Mol Model 7:445–453
36. Wickelmaier F (2003) An introduction to MDS, reports from the sound quality research unit. Aalborg University, Denmark, Denmark, pp 1–26
37. Li JX (2004) Visualization of high-dimensional data with relational perspective map. Inf Vis 3:49–59
38. Balasubramanian M, Schwartz EL (2002) The isomap algorithm and topological stability. Science 295:7–7
39. Dijkstra EW (1959) A note on two problems in connexion with graphs. Numer Math (Heidelb) 1:269–271
40. Hinton GE, Roweis ST (2003) Stochastic neighbor embedding. In: Becker S, Thrun S, Obermayer K (eds) Advances in neural information processing systems, 2003. MIT Press, Cambridge, MA, USA, pp 833–840
41. Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22:79–86
42. DeMers D, Cottrell GW (1993) Non-linear dimensionality reduction. In: Hanson SJ, Cowan JD, Giles CL (eds) Advances in neural information processing systems, vol 1993. Morgan-Kaufmann, pp 580–587
43. Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313:504–507
44. Belkin M, Niyogi P (2002) Laplacian Eigenmaps and spectral techniques for embedding and clustering. In: Dietterich TG, Becker S, Ghahramani Z (eds) Advances in neural information processing systems, vol 2002. MIT Press, pp 585–591
45. Anderson WN, Morley TD (1985) Eigenvalues of the Laplacian of a graph. Linear Multilinear Algebra 18:141–145
46. Naidu PS (1995) Modern spectrum analysis of time series. CRC Press, Boca Raton, Florida
47. Wolpert D (1992) Stacked generalization. Neural Netw 5:241–259
48. Yang HH, Moody J (1999) Data visualization and feature selection: new algorithms for nongaussian data. Proc Adv Neural Info Proc Syst 12:687–693
49. Gisbrecht A, Schulz A, Hammer B (2015) Parametric nonlinear dimensionality reduction using kernel t-SNE. Neurocomputing 147:71–82
50. Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics 1:80–83
51. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer Series in Statistics, Springer, New York
52. Buriro AB, Shoorangiz R, Weddell SJ, Jones RD (2018) Predicting microsleep states using EEG inter-channel dependencies. IEEE Trans Neural Syst Rehab Eng 26:2260–2269
53. Weddell SJ, Ayyagari SSDP, Jones RD (2021) Reservoir-computing approaches to microsleep detection. J Neural Eng 18(046021):1–11
54. Krishnamoorthy V, Shoorangiz R, Weddell SJ, Beckert L, Jones RD (2019) Deep learning with convolutional neural network for detecting microsleep states from EEG: a comparison between the oversampling technique and cost-based learning. Conf Proc IEEE Eng Med Biol Soc 41:4152–4155
55. Zhang P, Wang X, Zhang W, Chen J (2019) Learning spatial-spectral-temporal EEG features with recurrent 3D convolutional neural networks for cross-task mental workload assessment. IEEE Trans Neural Syst Rehab Eng 27:31–42
56. Sors A, Bonnet S, Mirek S, Vercueil L, Payen J-F (2018) A convolutional neural network for sleep stage scoring from raw single-channel EEG. Biomed Signal Process Control 42:107–114
57. Chua KC, Chandran V, Aeharya R (2007) Higher order spectral (HOS) analysis of epileptic EEG signals. Conf Proc IEEE Eng Med Biol Soc 37:6495–6498
58. Levendowski DJ, St Louis EK, Strambi LF, Galbiati A, Westbrook P, Berka C (2018) Comparison of EMG power during sleep from the submental and frontalis muscles. Nat Sci Sleep 10:431–437
59. Good R (1975) Frontalis muscle tension and sleep latency. Psychophysiol 12:465–467

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

Sudhanshu Ayyagari received his BTech in Electrical and Communication Engineering, Indian Institute of Technology Guwahati, India, and PhD degree in Electrical & Computer Engineering from University of Canterbury, New Zealand. He is an Australian and New Zealand patent attorney and Australian trade marks attorney. His practice focuses on high-tech patent work in the electrical and computer engineering domains, including machine learning, artificial intelligence, biomedical engineering, human computer interaction, and wireless power electronics. He is a fellow of the New Zealand Institute of Patent Attorneys.



Richard Jones received his ME degree in Electrical & Electronic Engineering from University of Canterbury and PhD in Medicine from University of Otago. He is Director of the Christchurch Neurotechnology Research Programme of the New Zealand Brain Research Institute, a Professor in Electrical & Computer Engineering and Psychology at University of Canterbury, and a Research Professor in Medicine at University of Otago. His research interests fall largely within neural engineering and the neurosciences. He is a Fellow of ACPSEM, EngNZ, AIMBE, InstP, and IEEE.

Stephen Weddell received his MAppSc degree in Electrical & Computer Engineering from Curtin University, Australia, and PhD degree in Electrical & Electronic Engineering from University of Canterbury, New Zealand. He is a Associate Professor, Leader of the Computational Design and Adaptation group, and past Director of Computer Engineering Studies at University of Canterbury. His research interests are in image and digital signal processing, machine learning, neuroengineering, adaptive optics, brain computer interfaces, systolic computer architectures, and high-performance computing. He is a Senior Member of IEEE.