Computers in Biology and Medicine 139 (2021) 104969

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Classification of alcoholic EEG signals using wavelet scattering transform-based features

Abdul Baseer Buriro [a,*], Bilal Ahmed [a], Gulsher Baloch [a], Junaid Ahmed [a], Reza Shoorangiz [b,c,d,e], Stephen J. Weddell [b], Richard D. Jones [b,c,d,e]

[a] Department of Electrical Engineering, Sukkur IBA University, Sukkur, 65200, Pakistan
[b] Department of Electrical and Computer Engineering, University of Canterbury, Christchurch, 8041, New Zealand
[c] New Zealand Brain Research Institute, Christchurch, 8011, New Zealand
[d] School of Psychology, Speech and Hearing, University of Canterbury, Christchurch, 8041, New Zealand
[e] Department of Medicine, University of Otago, Christchurch, 8011, New Zealand

ARTICLE INFO

Keywords:
Wavelet scattering transform (WST)
Feature extraction
Machine learning
Convolutional neural network (CNN)
Support vector machine (SVM)
Alcoholism

ABSTRACT

Following the research question and the relevant dataset, feature extraction is the most important component of machine learning and data science pipelines. The wavelet scattering transform (WST) is a recently developed knowledge-based feature extraction technique and is structurally like a convolutional neural network (CNN). It preserves information in high-frequency, is insensitive to signal deformations, and generates low variance features of real-valued signals generally required in classification tasks. With data from a publicly-available UCI database, we investigated the ability of WST-based features extracted from multichannel electroencephalogram (EEG) signals to discriminate 1.0-s EEG records of 20 male subjects with alcoholism and 20 male healthy subjects. Using record-wise 10-fold cross-validation, we found that WST-based features, inputted to a support vector machine (SVM) classifier, were able to correctly classify all alcoholic and normal EEG records. Similar performances were achieved with 1D CNN. In contrast, the highest independent-subject-wise mean 10-fold cross-validation performance was achieved with WST-based features fed to a linear discriminant (LDA) classifier. The results achieved with two 10-fold cross-validation approaches suggest that the WST together with a conventional classifier is an alternative to CNN for classification of alcoholic and normal EEGs. WST-based features from occipital and parietal regions were the most informative at discriminating between alcoholic and normal EEG records.

1. Introduction

Feature extraction, after the research question followed by the relevant dataset, is an important component of machine learning and data science pipelines [1]. Feature extraction is a mapping from an input space to an output space and can involve a signal or statistical processing algorithm [2–4]. Feature extraction simultaneously compresses the data and retains the relevant information. Subsequently, it improves the generalization ability of classifiers and reduces the computational and storage requirements. A good feature is translation and deformation invariant and simultaneously has minimum intra- and maximum inter-class variability.

Raw values, such as direct measurements of amplitudes of acoustic or electroencephalogram (EEG) signals or images, can be considered feature extraction. Such values, however, are often not amenable to learning [1]. Derived values, such as variances, entropies, and spectra of EEG signals [5], mel-frequency cepstrum of acoustic signals [6], and statistical attributes of images [7] from raw measurements are all considered knowledge-based feature extraction. Such feature extraction generally involves domain-specific knowledge and, subsequently, manual selection of a signal or statistical processing algorithm. Automatic feature extraction involves a computing system such as an artificial neural network (ANN) or one of its variants, such as a convolutional neural network (CNN). A CNN inherently maps an input space to an output space by convolving the input data with a linear filter, adding a bias term, and applying a non-linear function [2,8–11].

The Fourier transform is one of the most-used and locally time-invariant signal processing techniques employed to extract features.

* Corresponding author.
E-mail address: abdul.baseer@iba-suk.edu.pk (A.B. Buriro).

https://doi.org/10.1016/j.compbiomed.2021.104969
Received 3 August 2021; Received in revised form 20 October 2021; Accepted 20 October 2021
Available online 22 October 2021

However, the Fourier transform is unstable to high-frequency deformations [12]. Wavelet transforms have been effective tools for analyzing and classifying non-stationary and nonlinear signals. Wavelet transforms are stable to such deformations but are not translation invariant if subsampling is involved [13]. The Fourier and wavelet transform are therefore not ideal feature extractors.

The wavelet scattering transform (WST) [12] is a recently-developed knowledge-based feature extraction technique. Structurally, it is like a convolutional network (such as CNN and deep belief), in that it involves cascaded decomposition and convolution of a signal with wavelets, followed by complex modulus, and local averaging. Consequently, WST-based features possess translation invariance, deformation stability, and high-frequency information [13]. The WST, therefore, becomes a very suitable feature extractor for non-linear and non-stationary signals and has been widely used in audio, music, and image classification [12,14,15].

The WST and CNN have common properties of multiscale contraction, linearization of hierarchical symmetries, and sparse representation. In contrast, the main difference between them is that the WST uses predefined filter banks, while CNN requires training of the filters [16]. The WST can, therefore, work accurately and efficiently for small data sets, whereas a CNN requires a large amount of training data.

Alcoholism is a common psychiatric disorder associated with considerable morbidity and mortality [17]. A diagnosis of alcoholism is often confirmed by assessing responses to criticism to cut down drinking, guilty feelings, and first drink at dawn, which, due to subjectivity, has low accuracy and can be misleading [18]. EEG, on the other hand, noninvasively measures the electrical activity of the human brain. Besides high temporal resolution, EEG acquisition is relatively inexpensive and convenient in real-time applications [19]. EEG recordings have been widely used in diagnostic, clinical, and sleep-related research settings [20–23]. EEG signals may therefore be a means to identify and monitor alcoholic patients. However, EEG signals are very low amplitude and have a low signal-to-noise ratio (SNR) [18]; that is, task-irrelevant sources affect EEG signals more than task-relevant sources. They are also inherently nonlinear and nonstationary [24,25], and vary among individuals due to their physiological differences and subject-specific cognitive styles [26]. Compared to audio signals, EEG signals are low-frequency and generally multivariate (i.e., multichannel). Therefore, the classification of EEG signals is fundamentally more difficult than classification of image and audio signals.

Many recent studies have been conducted to explore different feature extraction techniques aimed at classifying alcoholism from EEG [27–32]. Zhang et al. [33] have investigated multiple combinations of CNN architectures used as feature extractors to classify pre-disposition of alcoholism. Mukhtar et al. [34] explained dropout, batch normalization, and kernel regularization techniques for stepwise improvement of a baseline CNN model.

The current study aimed to empirically investigate the usefulness of WST-based features to accurately and automatically detect alcoholism based on EEG. As CNN has been the leading deep-learning architecture used in more than 40% of EEG-based studies [35,36] and is structurally similar to the WST, this study systematically analyzed and compared their respective classification performances. K-fold cross-validation was performed on concatenated EEG records (termed as record-wise cross--validation), for which the training and test datasets are not independent – i.e., records from the same subject can be in both training and test sets. In addition, 10-fold cross-validation was performed on individual subjects, followed by a concatenation of respective EEGs, and was termed subject-wise k-fold cross-validation.

2. Methods

2.1. Data

The EEG dataset used in this study was obtained from the University of California, Irvine Knowledge Discovery (UCI KDD) [37]. The dataset provided recordings from 64 electrodes placed on the subject's scalp with a sampling frequency of 256 Hz (3.9 ms intervals) for 1000 ms. The EEG signals were referenced to the Cz electrode. Two bipolar derivations were used to record the horizontal and vertical electrooculograms (EOG).

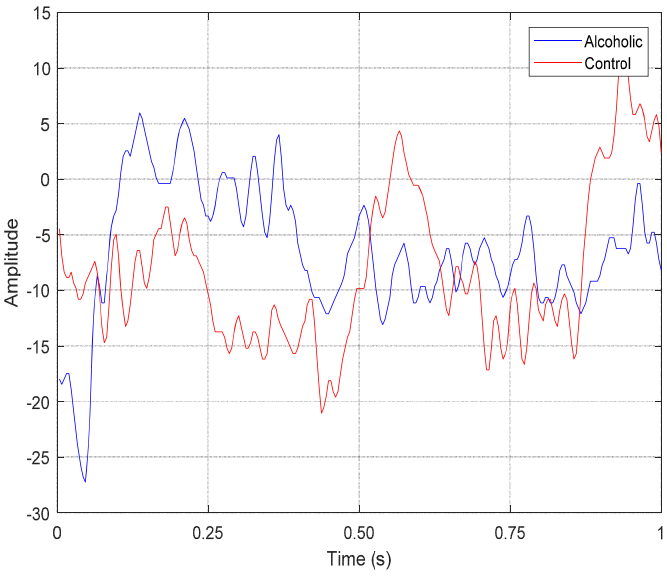

Fig. 1. Raw alcoholic and control EEG signal from FP1 channel.

There are three versions of the dataset: the small, large, and full dataset. Each set contains two groups of alcoholic and control subjects. The full dataset comprises 122 male subjects (mean age 35.8 ± 5.3 years): 77 diagnosed with alcoholism and 45 healthy controls. The subjects belonging to both the alcoholic and control subjects' groups were excluded from further analysis. All subjects completed 120 trials. Trials with error messages and excessive body, muscle, and eye movement were also rejected. 43 alcoholic and 22 control subjects remained after the above elimination process. The current study used 16 trials of each of the 20 alcoholics and 20 controls, selected from identification numbers starting with 'co2'. Typical 1.0-s alcoholic and normal EEG signals are shown in Fig. 1.

All subjects were exposed to single visual stimuli (S1) and two stimuli (S1 and S2). The stimuli were composed of 90 pictures of objects chosen from the Snodgrass and Vander picture set [38]. When two stimuli were shown, they were presented in either a matched condition where S1 was identical to S2 or in a non-matched condition where S1 differed from S2. The duration of each stimulus in each trial was 300 ms and the interval between trials was fixed at 3.2 s.

High-density EEG (i.e., channels >20) recordings require a longer time to train and optimize machine learning models and may cause overfitting. The Full 64-channel and 16-channel EEG sets have been shown to give similar average performances at classifying alcoholic and normal subjects [39]. Therefore, the original data with 64-channel was considered unnecessarily high-density, and 16 channels were chosen according to the International 10–20 system: FP1, FP2, F3, F4, F7, F8, C3, C4, P3, P4, T3, T5, T4, T6, O1, O2. To attain generalized classification performances, the models were evaluated using two 10-fold cross-validations performed respectively on EEG records and subjects.

2.2. Wavelet scattering transform

The WST is a cascaded decomposition and convolution of a signal with wavelets, followed by complex modulus, and local averaging. The first step to calculate the WST is to convolve the signal x with the dilated

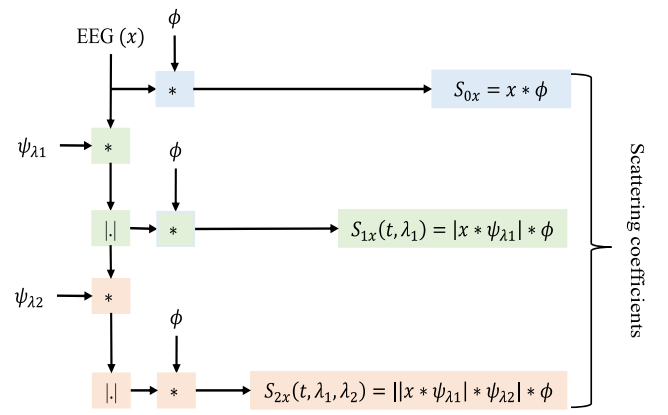

Fig. 2. Schematic of feature extraction from EEG channels using second-order WST. S_{0x}, S_{1x}, and S_{2x} respectively indicate 0^{th} (time-averaged or low pass filtered), 1st, and 2nd order scattering coefficients of WST. * and |.| are the convolution and modulus operators, respectively.

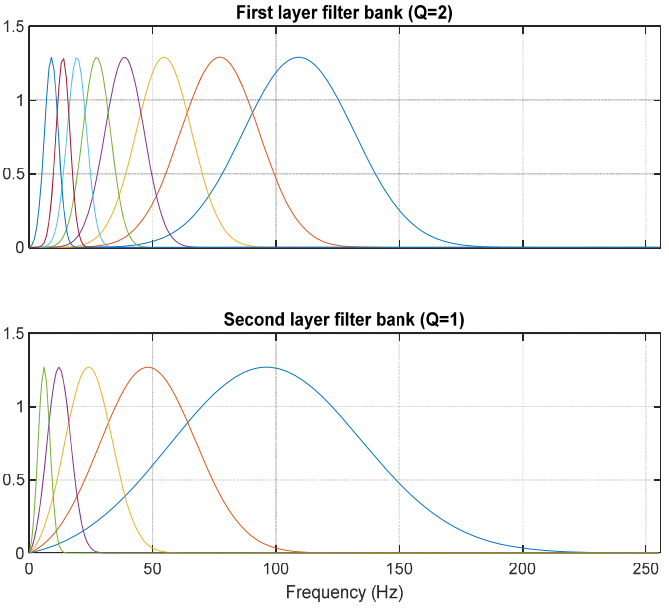

Fig. 3. Illustration of two-level filter banks (i.e., wavelets) for different center frequencies λ.

mother wavelet ψ with a center frequency of λ (i.e., $x*\psi_\lambda$). The convolved signal oscillates at a scale 2^j and averaging such signal results in zero. To remove such oscillations (i.e., complex phase), a nonlinear (modulus/rectifier) operator is therefore performed on the convolved signal (i.e., $|x*\psi_\lambda|$). This operation generally increases the frequency of a signal by a factor of 2 and can be used to compensate the loss of information due to down sampling. The final step is to apply a time-average/low-pass filter φ to the absolute convolved signal (i.e., $|x*\psi_\lambda|*\varphi$). The first-order scattering coefficients are therefore defined as the average absolute amplitudes of wavelet coefficients for any scale (i.e. $1 \le j \le J$), over a half-overlapping time window of size 2^j, and are obtained by

$$S_{1x}(t,\lambda_1) = |x * \psi_{\lambda_1}| * \varphi. \tag{1}$$

The second-order scattering coefficients are calculated by repeating the above steps applied to each of $|x*\psi_{\lambda_1}|$, i.e.,

$$S_{2x}(t,\lambda_1,\lambda_2) = ||x * \psi_{\lambda_1}| * \psi_{\lambda_2}| * \varphi. \tag{2.2}$$

The higher orders (i.e. $m \ge 2$) wavelet scattering coefficients can be calculated by iterating the above process as

$$S_{mx}(t,\lambda_1,\ldots,\lambda_m) = |||x * \psi_{\lambda_1}| * \psi_{\lambda_2}| \ldots * \psi_{\lambda_m}| * \varphi. \tag{2.3}$$

The zero-order scattering coefficients describe the local translation invariance of the signal and are obtained with a time-average $S_{0x}(t) = x*\varphi$. The averaging operation at each stage results in the loss of high-frequency contents of the convolved signal, which can be recovered by convolving it with the wavelet in the next stage.

The energy of scattering coefficients decreases with an increase in layers, and the first two layers contain 99% of the energy [15]. Furthermore, Ahmed et al. [23] used WST to extract features from EEG and concluded that an m of 2 was optimal. In the current study, the same value of m was used to extract features from 16-channel concatenated EEG signals (i.e. 16 trials $\times$ 256 samples/trial, 16 channels) from each subject. Fig. 2 illustrates the steps taken to compute WST coefficients at each level and, subsequently, aggregated coefficients were used as the features.

The mother wavelet used was the Morlet (Gabor) wavelet, which is closely related to the human visual cortex [40]. With dimensionless frequency ω_0 and time η, the Gabor wavelet is defined as $\psi = \pi^{0.25} e^{-i\omega_0\eta} e^{-0.5\eta^2}$. Q defines the number of wavelets per octave with center frequencies $\lambda = 2^{k/Q}$ for $k \in \mathbb{Z}$, and Q^{-1} corresponds to the bandwidth of ψ. These bandpass wavelets ψ cover the whole desired frequency spectrum and are used to discretize the scale J. Fig. 3 shows Morlet wavelets ψ_1 with $Q_1 = 2$, and ψ_2 with $Q_2 = 1$. There is, however, no rule-of-thumb for selecting an octave frequency resolution. In this study, four different combinations of Q, i.e., $Q(Q_1, Q_2) = \{(2,\ 1), (4,\ 1), (8,\ 1), (4,\ 2)\}$ were used and the one that gave the highest Fisher score was selected. Furthermore, the number of features (dimensionality) increases with an increase in Q. Similarly, using the Fisher score, the

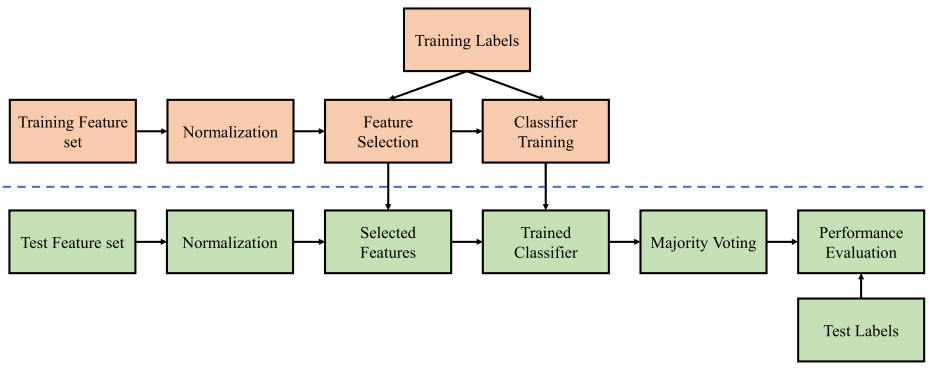

Fig. 4. Steps taken in this study for conventional machine learning pipeline. The feature sets were the coefficients of the WST (Fig. 2). The final label against an EEG record (being alcoholic or control) was achieved using a majority voting applied on the window (16) times decisions and posterior probabilities of the classifiers.

invariance scale (time scale of the low-pass filter) was chosen from a set of $S = \{0.25,\ 0.50,\ 1.0\}$ *s*. The smaller the invariance scale, the larger the bandwidth of the low-pass filter. For a fixed value of Q, a decrease in the invariance scale decreases the number d and increases the length n of coefficients/features, $F \in \mathbb{R}^{n\times d}$.

The output of the WST is a tensor, $F \in \mathbb{R}^{\text{coefficients }(c)\times \text{ window }(w)\times \text{EEG channels }(16)}$, which was converted into a matrix $F \in \mathbb{R}^{w\times 16c}$, suitable as an input for machine learning and data science pipelines. Since multiple scattering windows are obtained for each EEG record, the new labels were created by window times repeating the corresponding labels. The features from the training dataset were normalized to zero mean and unit variance. The features from the test dataset were normalized per the mean and variance of the features from the training dataset.

2.3. Classification with conventional classifiers

To assess consistency in classification performances, two classification pipelines were used: (1) linear discriminant analysis (LDA)-based feature selection and classification, and (2) support vector machine with radial basis function (SVM-RBF)-based embedded feature selection and classification.

The complexity of a model increases with the number of features and the accuracy of such a model on test data may not be generalized. Additionally, redundant and irrelevant features can substantially deteriorate classification performance [41]. Feature selection (as shown in Fig. 4) was employed on the training dataset to address the issue of a large number of features.

For the first classification pipeline, the overall feature selection comprised a Fisher-score-based ranking [42] followed by correlation-based filtering and linear LDA-based wrapping. All features were ranked as per their respective Fisher scores. The pairwise correlation was then iteratively performed among the ranked features. In each iteration, all low-ranked and correlated features (with $|r| > 0.90$) were discarded. Feature dependencies and combined discriminatory power in the ranking are generally ignored, and feature selection becomes suboptimal [43,44]. Features were finally selected by using the sequential forward selection (SFS) method and the objective function used was the area under the curve of the receiver operating characteristic AUC_{ROC} [45]. The mean AUC_{ROC} of the 5-fold cross-validation with the top-ranked feature was calculated and saved. The successive feature from the subset was then selected if their combined mean AUC_{ROC} was improved, otherwise, it was discarded. The process was iterated until a stopping criterion was met. Like [5,21] the stopping criterion in this study was a logical *OR* of a maximum number of selected features (70) and allowed iterations (70) in which no performance improvement was observed.

Finally, an LDA classifier was used on selected WST-based features to classify individual 1.0-s alcoholic and normal EEG records. LDA has a linear decision boundary and the density of each class is assumed to be multivariate normal with a common covariance matrix. Due to stable estimates, LDA works well even for deviated data distributions [46]. The generalized system was achieved by training the classifier on the training dataset and tested on the test dataset.

The second classification pipeline was an SVM-based feature selection and classification [47]. Such feature selection is termed as embedded technique, where an optimal subset of the features is achieved during the training process of a classifier without explicitly splitting the training dataset into training and testing sets. Embedded feature selection is classifier-dependent but computationally less expensive than wrapper methods [41,43]. The SVM classifier has been widely used in EEG-based classification studies [18,24,29,48,49]. A classifier using training data maximizes the margin between two classes, which leads to better generalization capability – i.e., classification performance on independent data [46,50]. The RBF kernel, defined by $K(x_i, x_j) = \exp\left(-\frac{x_i - x_j^2}{2\sigma^2}\right)$, $\sigma \neq 0$, maps the data to a higher dimensional space and was used to handle the nonlinearly separable data. A 10-fold cross-validation on the training dataset was performed to choose the regularization parameter from $C = \{10^{-3} : 10 : 10^2\}$ and the kernel width $\sigma = \{1,\ 0.1,\ 0.01\}$.

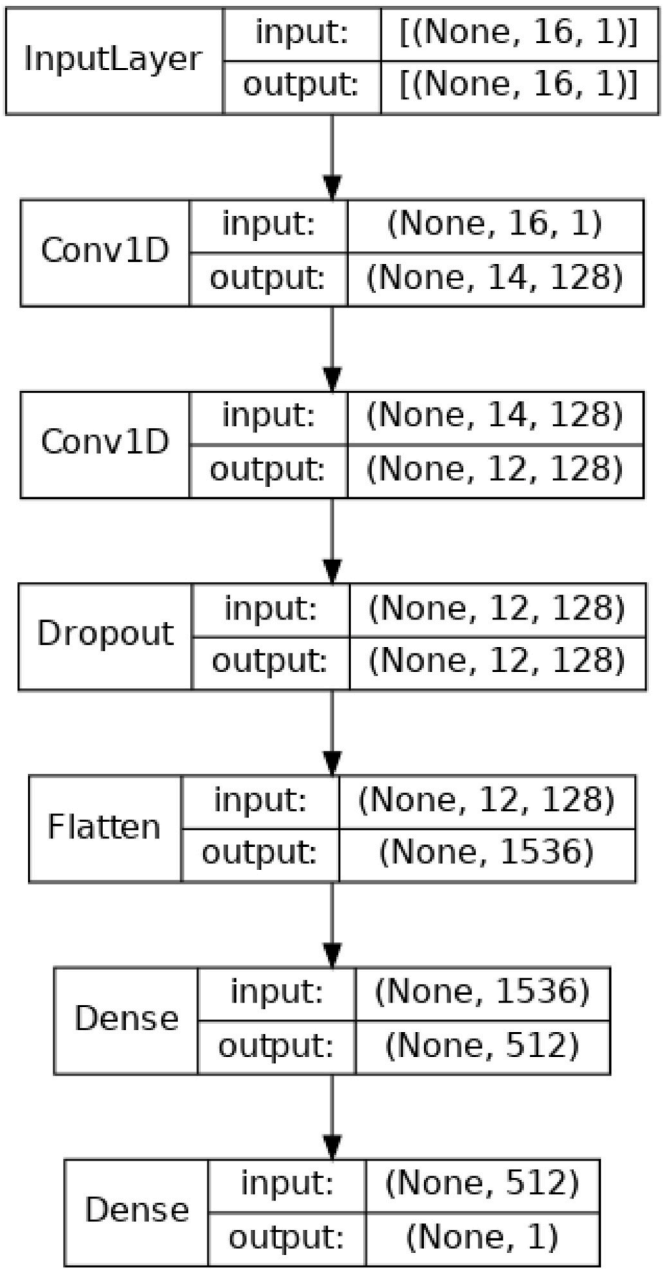

Fig. 5. Block diagram example of the untrained 1D-CNN architecture used in this study, where None represents batch size. The filter and kernel sizes in both Conv1D layers were selected using mean 10-fold cross-validation accuracy on the training dataset. Using majority voting, each of the 256 outputs were transformed to one output corresponding to a single 1.0-s EEG record.

For large and complex datasets, the training of, and predictions via SVM classifiers, particularly with nonlinear kernels, are computationally very expensive. ThunderSVM is a newly developed SVM library, which exploits the high-performance of graphic processing units (GPU) and multi-core central processing units (CPU), and is much faster than conventionally-used LibSVM [51]. This study used ThunderSVM in Google Colab to implement the SVM-RBF classifier.

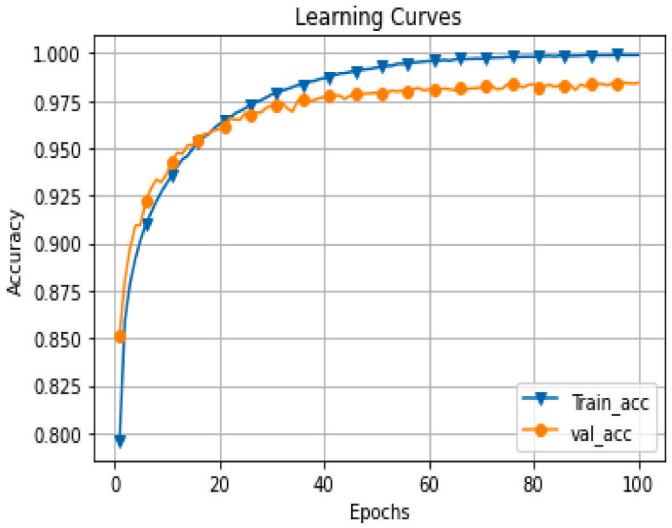

Fig. 6. An example of training and validation accuracies on the training dataset to select the number of epochs.

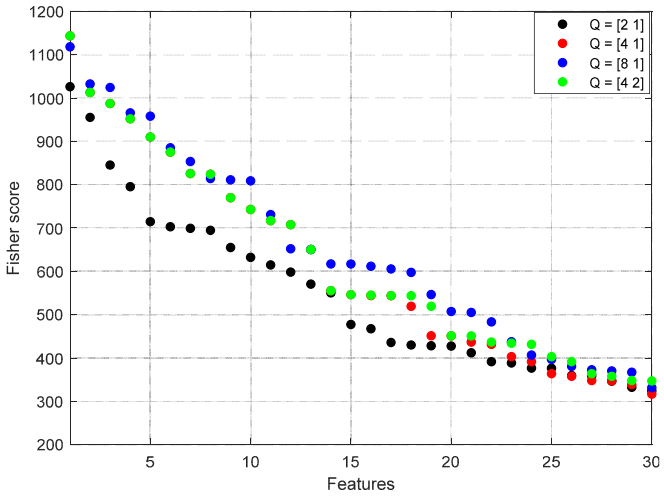

Fig. 7. Effect of Q on Fisher score (in descending order) with scale invariance of 0.25 s of top-ranked 30 features.

Class prediction (i.e., alcoholic, control) against each EEG record was achieved using voting performed on decisions and posterior probabilities of the classifier [52] of the corresponding scattering representations.

2.4. Classification using convolutional neural network

The one-dimensional CNN, as shown in Fig. 5, was used in Keras, a deep-learning application programming interface (API) with TensorFlow at its backend. As the WST contained two layers, the CNN model, for fair comparisons, was restricted to two convolutional layers. One dropout layer for regularization was used to reduce overfitting. A flatten layer was used to transform the data into one dimension to make it feasible for the dense layers. The final output layer had a size of one, corresponding to a binary classification problem.

A CNN requires a large dataset for training, and inherently extracts features, and reduces dimensionality, referred to as feature engineering [1]. CNNs are therefore generally fed with, and are capable of handling high-dimensional and large-size raw EEG signals [35,36]. The concatenated EEGs from the training dataset were therefore used to model the CNN and tune its hyperparameters. The filter length and the kernel size affect the extracted features and there is no rule-of-thumb to select them. In this study, four filter lengths $L = \{16, 32, 64, 128\}$ were used and the one that gave the highest mean 10-fold cross-validation accuracy was chosen. Similarly, the kernel size was chosen from a set of $K = \{1, 3, 5, 7\}$ using 10-fold cross-validation on the training dataset. New labels were created by repeating the corresponding label by 256 times. The number of epochs was chosen according to when the validation accuracy plateaued (refer to Fig. 6). Whereas, due to a complex network (in terms of filter size), the value of the drop-out was set to 0.5. The rectified logic unit (Relu) [53] has been the widely-used nonlinear activation function, employed in 70% of CNN architectures, and performs the best with the sigmoid function [36]. To learn more effective feature representations, Relu, and sigmoid functions were used to perform the nonlinear activations in the hidden and output layers, respectively. The weights of the model were optimized using the adam optimizer [54]. The binary cross-entropy loss was used to update the weights during training.

The final classification of each record was achieved via voting performed on the individual decisions and corresponding probabilities of the CNN against 256 samples. The model was implemented in a cloud-based platform of Google Colab.

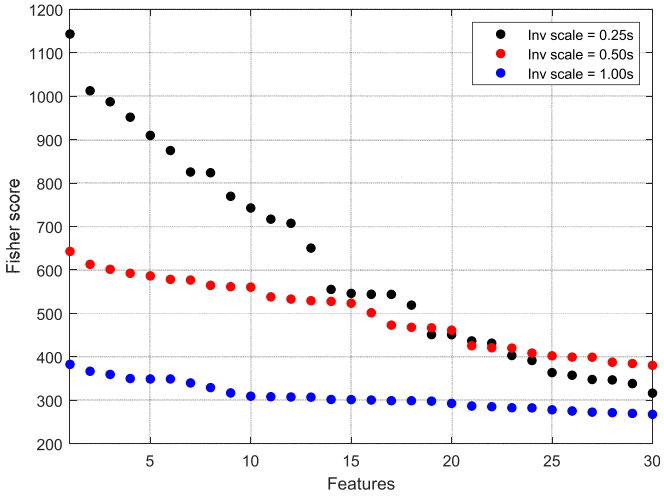

Fig. 8. Effect of invariance scale on Fisher score (in descending order) with $Q_1 = 4$ and $Q_2 = 1$ of top-ranked 30 features.

2.5. Performance evaluation

Classification performances were evaluated using two 10-fold cross-validation approaches – record-wise and subject-wise – and were based on 16 clean 1.0-s EEG records from each of the 20 alcoholic and 20 healthy subjects. In the record-wise approach, 10-fold cross-validation was performed on concatenated EEG records, where both the training and test datasets can have EEG records (or some samples of the record) from the same subject. In the subject-wise approach, 10-fold cross-validation was performed on all 40 subjects. For each fold, the respective EEG records from 36 training and 4 test subjects were concatenated. For both cases, mean values of sensitivity and specificity are reported.

However, these metrics alone can be biased and misleading [55]. As the data is balanced (i.e., the number of subjects and trials in both classes have been made intentionally equal), these performance metrics were also combined into the widely-used performance metric of accuracy. Furthermore, for binary balanced datasets, AUC_{ROC} is statistically more consistent and discriminating than accuracy at comparing learning algorithms [56] and is also reported.

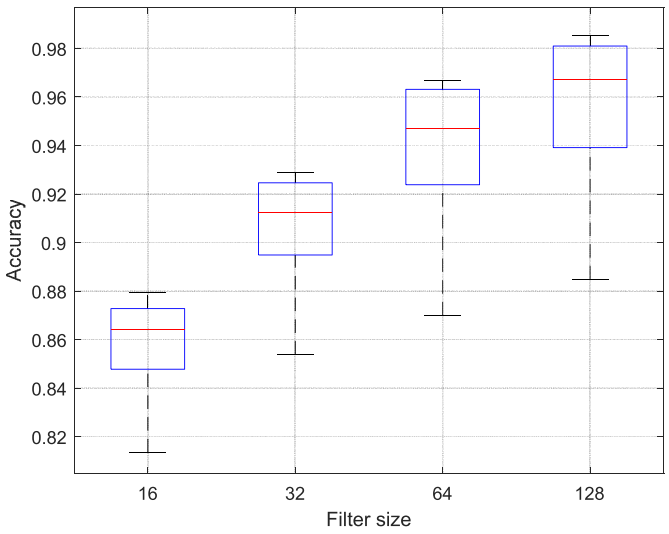

Fig. 9. Effect of filter length on 10-fold cross-validation accuracy of two-layer 1D-CNN with a kernel size of 3 and a fixed epoch length of 15.

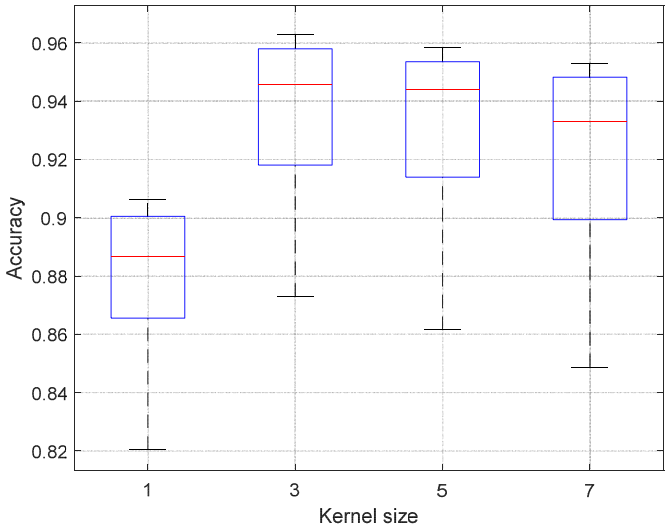

Fig. 10. Effect of kernel size on 10-fold cross-validation performance of two-layered 1D-CNN with a filter size of 64 and a fixed epoch length of 15.

3. Results

3.1. Analysis

For a fixed invariance scale of 0.25 s, the overall and one-to-one discriminatory power (in terms of Fisher score) difference among features with different values of Q was small. The one-to-one discriminatory powers of the top 17 features with $Q = (4,\ 1)$ and $Q = (4,\ 2)$ were the same. Whereas the one-to-one discriminatory powers of features ranked beyond 20 were insensitive to the values of Q as shown in Fig. 7. The lowest discriminatory power of the top 22 features was for $Q = (2,\ 1)$.

The highest Fisher score of 1150 was achieved for $Q = (4, 1)$ and an invariance scale of 0.25 s. Invariance scales of 0.5 and 1.0s resulted in their respective highest Fisher scores of 650 and 400. With an increasing invariance scale, the Fisher score decreased and the difference between the Fisher scores of consecutive features flattened as shown in Fig. 8. However, the one-to-one discriminatory power difference among features, ranked beyond 25, was small. With $Q = (4, 1)$ and an invariance scale of 0.25 s, the top five discriminatory features, each with Fisher scores >900, corresponded to O1, P4, P8, O2, and C4. A high Fisher score of a feature indicates high inter-class and low intra-class variabilities. The Fisher scores of WST-based features were found to be more sensitive to the invariance scale than to the value of Q.

The 10-fold cross-validation accuracy consistently increased with increasing filter length, as shown in Fig. 9. The highest accuracy of 98% was achieved with a filter size of 128 and a kernel size of 3. For a fixed filter length, the highest 10-fold cross-validation accuracy of 96% was achieved with a kernel length of 3. For higher kernel lengths, accuracy decreases as shown in Fig. 10.

3.2. Performances

In record-wise 10-fold cross-validation and across the folds, both SVM classifier (using WST-based features) and 1D-CNN correctly classified all 1.0-s EEG records of the selected subjects in the alcoholic and healthy groups. However, the LDA classifier with WST-based features resulted in lower mean performance metrics (sensitivity, specificity, accuracy, AUC_{ROC}) of (89%, 93%, 91%, 98%), shown in Fig. 11.

In contrast, using WST-based features, the subject-wise highest mean 10-fold cross-validation performance metrics (sensitivity, specificity, accuracy, AUC_{ROC}) of (66%, 67%, 66%, 77%) were achieved with the LDA classifier. Performances achieved with the SVM classifier were comparable (60%, 69%, 65%, 75%). Except for the having the highest specificity (albeit at the expense of sensitivity), the 1D CNN had lower performance metrics of (48%, 76%, 62%, 64%), shown in Fig. 12.

4. Discussion

Using WST features, the SVM-based machine-learning pipeline required the selection of four parameters: number of wavelets per octave in terms of Q, invariance scale, regularization, and kernel width. In contrast, the key pre-training parameters required by the 1D-CNN are filter size, kernel, epoch size, dropout rate, learning rate, and size and number of dense layers. Therefore, compared to a conventional machine learning pipeline with a nonlinear classifier (i.e., SVM), the implementation, training, and optimization of the CNN classifier, is more complex and computationally demanding.

In both the WST and CNN, filter sizes were directly related to the number of features. CNNs with large filters and appropriate dropout rates can generally achieve higher accuracy but large filter sizes are susceptible to overfitting and require large training data. In addition, a larger filter size leads to a higher variance in performance (see Fig. 9). Compared to filter size, scale invariance in WST results in more discriminative features.

The computations involved in WST and the resultant coefficients are directly related to the number of layers and the number of corresponding filters. Scattering coefficients, however, are restricted to order $m \leq 2$ because amplitudes beyond that are negligible [16]. Similarly, feature invariance and robustness in CNNs are directly related to the depth of the network [57,58]. However, unlike WST, increasing depth in CNN leads to a complex system with many parameters to tune, which eventually requires larger training datasets.

Feature engineering (i.e., feature extraction and selection) in CNN is automatic and depends empirically (e.g., k-fold cross-validations) on selected and trained filters. Besides computational ease, filters involved in computing WST-based features do not require training and, subsequently, the overall feature extraction step is independent of the length of the data. Therefore, using simple and linear classifiers (like LDA) with wavelet scattering features, higher classification performances can be achieved on small datasets. Furthermore, due to mathematical background, WST-based features can be more directly interpreted. As the EEG signals were time-locked with visual stimuli (refer to section 2.1), it is not surprising that the top five WST-based features discriminating EEG

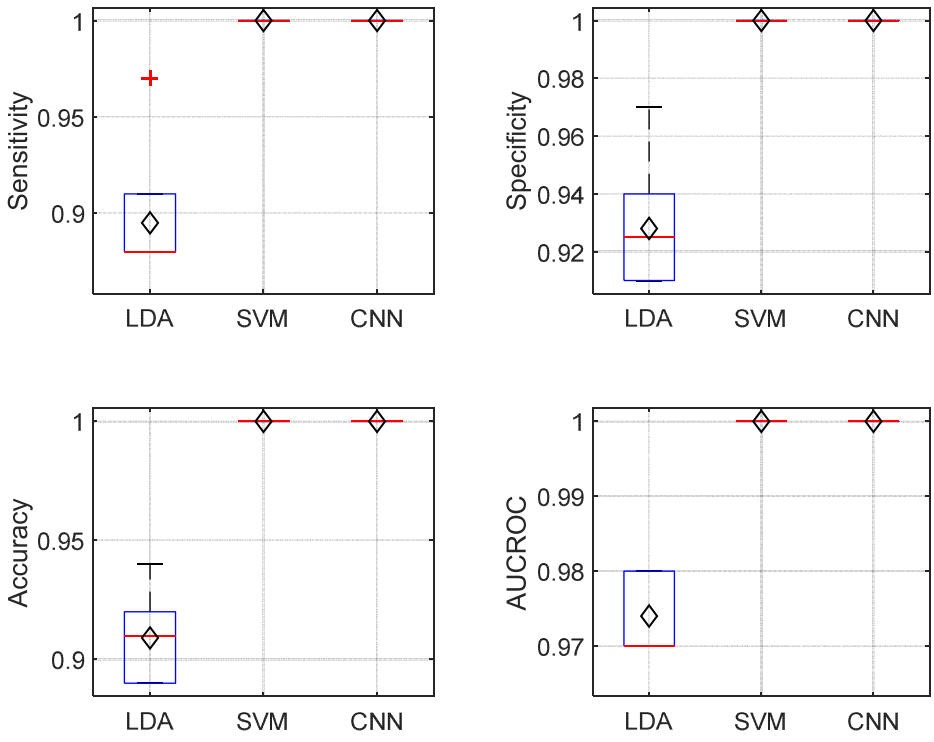

Fig. 11. Record-wise 10-fold cross-validation test performance metrics. Black diamonds indicate the respective mean values.

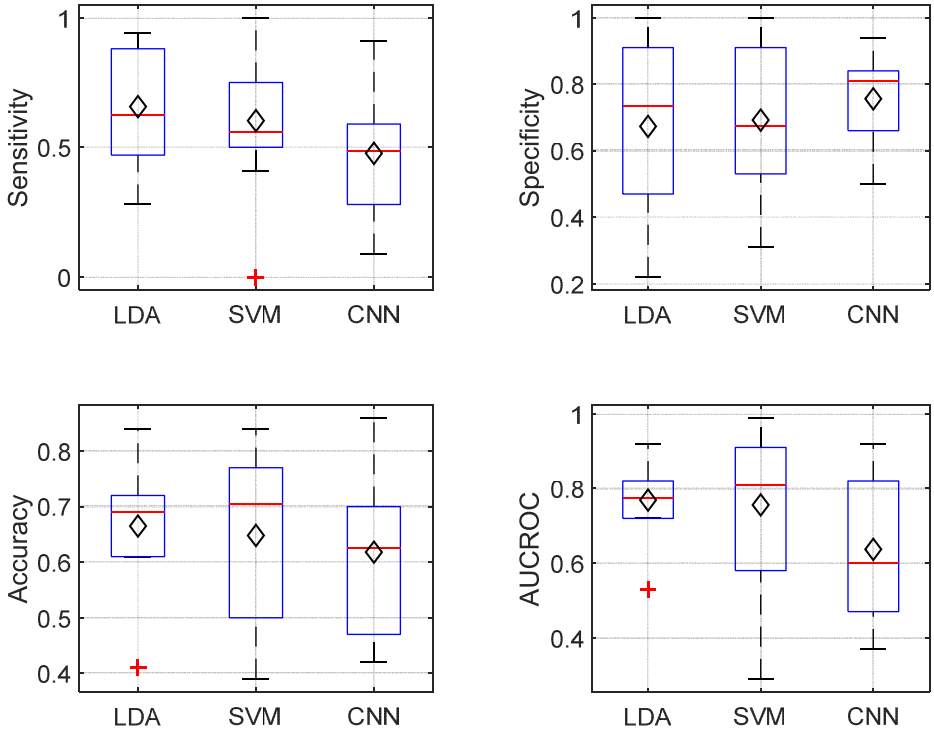

Fig. 12. Subject-wise 10-fold cross-validation test performance metrics. Black diamonds indicate the respective mean values.

Table 1
Record-wise mean performance comparisons with recent studies of UCI alcoholic dataset.

Author	Database	Features	Classifier	Evaluation method	Performances (%)
Anuragi et al. [32]	120 alcoholic and 120 control EEG records of 8-s duration	Entropies extracted from empirical wavelet transform	SVM RBF	Leave-one-sample-out	Accuracy = 98.8 Sensitivity = 98.3 Specificity = 99.1
Anuragi and Sisodia [30]		Statistical features from EEG sub-bands obtained from flexible analytical wavelet transform	Least square SVM with polynomial kernel, Naïve Bayes	10-fold cross-validation	Accuracy = 99.2 Sensitivity = 99.2 Specificity = 99.4
Sharma et al. [29]		L2 norm and log-energy entropies from EEG sub-bands obtained from dual-tree complex wavelet transform	Least square SVM, Sequential minimal optimization SVM		Accuracy = 97.9 AUCROC = 100.0 MCC = 95.8
Patidar et al. [18]		Correntropy based on tunable Q-factor wavelet transform	Least square SVM		Accuracy = 97.0
Acharya et al. [24]		Approximate and sample entropy, Lyapunov exponent, higher order spectra	SVM with polynomial and RBF kernels	3-fold cross-validation	Accuracy = 91.7 Sensitivity = 90.0 Specificity = 93.3
Siuly et al. [27]		Statistical features extracted from optimum allocation-based sampling	Decision table	10-fold cross-validation	Accuracy = 99.6 Sensitivity = 99.6 Precision = 99.6 AUCROC = 99.4
Mukhar et al. [34]	30 alcoholic and 30 control trials	Deep CNN			Accuracy = 98.4 Precision = 100.0 Sensitivity = 96.8 AUCROC = 98.4
Bae et al. [48]	24 alcoholic and 16 control subjects selected randomly from the full dataset	Clustering coefficients, assortativity, average neighbourhood degree, and node between centrality obtained from effective connectivity network	SVM with a polynomial of order 3	4-fold cross-validation	Accuracy = 90.0 Sensitivity = 95.3 Specificity = 82.4
Rodrigues et al. [31]	10 alcoholic and 10 control subjects, with 30 trials per subject	Statistical features obtained from coefficients of wavelet packet decomposition	Naive Bayes	Training: Test Split = 75%:25%	Accuracy = 99.9
Bavkar et al. [28]	40 alcoholic and 40 control subjects, with 10 trials per subject	Absolute gamma band power	Ensemble-space KNN	10-fold cross-validation	Accuracy = 95.1
Current study	20 alcoholic and 20 control subjects, with 16 trials per subject from the full dataset	WST coefficients	SVM RBF		Accuracy = 100.0 Sensitivity = 100.0 Specificity = 100.0 AUCROC = 100.0
			LDA		Accuracy = 91.0 Sensitivity = 89.0 Specificity = 93.0 AUCROC = 98.0
		1D-CNN			Accuracy = 100.0 Sensitivity = 100.0

(*continued on next page*)

Table 1 (*continued*)

Author	Database	Features	Classifier	Evaluation method	Performances (%)
					Specificity = 100.0 AUCROC = 100.0

records from alcoholic and healthy subjects were from occipital and parietal regions of the brain. Our findings are in accordance with a meta-analysis [59] concluding that alcohol affects multiple cognitive domains, including visuospatial abilities, that rely on occipital and parietal regions, and executive functions, that rely on prefrontal cortex.

In the record-wise 10-fold cross-validation approach (see Fig. 11), both the SVM classifier with WST-based features and the 1D-CNN classifier were able to correctly discriminate between all 1.0-s EEG records of subjects with alcoholism disorder and healthy subjects. Such comparable results indicate that the features extracted through both approaches are closely informative and have low intra-class and high inter-class variability. However, in the subject-wise 10-fold cross-validation approach (see Fig. 12), both LDA and SVM classifiers fed with WST-based features, and except for specificity, gave the higher mean performances than 1D-CNN. The higher performance achieved with WST-based features may be indicative of their relatively low intra- and inter-subject variability. Therefore, features extracted via WST seem to be highly suitable for multichannel, nonlinear, and dynamic EEG signals. The highest subject-wise mean 10-fold cross-validation performances achieved via the LDA classifier indicate its robustness against the covariate shift in data distributions.

Several recent studies on aspects of the EEG data of alcoholic subjects from the UCI repository are given in Table 1. Most of the studies [18,24,27,29,30,32], on the same UCI repository, used 240 EEG records of 8-s duration (see Table 1) and have not explicitly mentioned the number of subjects and the dataset they used. Therefore, the performances reported in this study can't be compared with their reported performances. On the other hand, Bavkar et al. [28], Rodrigues et al. [31], and Bae et al. [48] have respectively used 80, 20, 40 subjects. The highest reported accuracy of 99.9% was achieved with wavelet packets and the Naïve Bayes classifier [31]. However, due to a single train-test split, their promising performance may considerably vary and the model may perform worst. The performances reported by Bavkar et al. [28] may be misleading as the full UCI dataset contains ~45 subjects in both groups. Zhang et al. [33] achieved their highest accuracy of 95.3% with MobileNet and SVM.

EEG signals are non-stationary and vary among individuals due to their physiological differences and subject-specific cognitive styles. Such inherent intra- and inter-subject variabilities cause covariate shifts in data distribution. Consequently, the transferability of the model among subjects and sessions is impeded and, subsequently, the model can't be generalized. Furthermore, k-fold cross-validation performed on concatenated EEGs may result in training and test datasets containing EEGs from the same subject, referred as information leakage. Therefore, the promising mean k-fold cross-validation performances achieved by all recent studies mentioned in Table 1, are erroneous and may lead to wrong conclusions. To determine the true generalized performances, k-fold cross-validation needs to be performed on subjects followed by the concatenation of the respective EEGs.

5. Conclusion

We empirically investigated WST-based EEG features to classify EEG records from a subset of alcoholic and normal subjects from the full dataset. In record-wise 10-fold cross-validation, both WST-based features with SVM classifiers and 1D-CNN resulted in 100% mean test accuracies on selected 1.0-s EEG records from subsets of alcoholic and normal subjects from the full dataset. Whereas, in subject-wise 10-fold cross-validation, WST-based features with both the conventional classifiers (i.e., LDA and SVM) gave higher mean performances than those achieved with 1D-CNN. The results suggest that WST-based features together with a conventional machine learning algorithm is a compelling objective alternative to CNN for detecting alcoholic subjects based on their 1.0-s EEG records. The most discriminatory features were from the occipital and parietal regions of the brain.

Contributions

Abdul Baseer Buriro: Conceptualization, Methodology, Software, Visualization, Investigation, Formal Analysis, Writing-Original Draft. Bilal Ahmed & Gulsher Baloch: Data Curation. Junaid Ahmed & Reza Shoorangiz: Validation, Writing-Review. Stephen J. Weddell & Richard D. Jones: Writing-Reviewing and Editing, Supervision.

Declaration of competing interest

The authors declare that there is no conflict of interest.

Acknowledgement

We gratefully acknowledge the collection and availability of this data made by Henri Begleiter of the Neurodynamics Laboratory at State University of New York Health Center at Brooklyn.

References

[1] P. Domingos, A few useful things to know about machine learning, Commun. ACM 55 (2012) 78–87, https://doi.org/10.1145/2347736.2347755.
[2] T. Wiatowski, H. Bolcskei, A mathematical theory of deep convolutional neural networks for feature extraction, IEEE Trans. Inf. Theor. 64 (2018) 1845–1866, https://doi.org/10.1109/TIT.2017.2776228.
[3] A.K. Jain, Artificial neural networks for feature extraction and multivariate data projection, IEEE Trans. Neural Network. 6 (1995) 296–317, https://doi.org/10.1109/72.363467.
[4] M.T.R. Peiris, P.R. Davidson, P.J. Bones, R.D. Jones, Detection of lapses in responsiveness from the EEG, J. Neural. Eng. 8 (2011), https://doi.org/10.1088/1741-2560/8/1/016003.
[5] A.B. Buriro, Prediction of Microsleeps Using EEG Inter-channel Relationships, University of Canterbury, Christchurch, New Zealand, 2019.
[6] P. Dhanalakshmi, S. Palanivel, V. Ramalingam, Classification of audio signals using SVM and RBFNN, Expert Syst. Appl. 36 (2009) 6069–6075.
[7] J. Guo, L. Liu, W. Song, C. Du, X. Zhao, The study of image feature extraction and classification, Int. Conf. Prog. Informatics Comput. (2017) 174–178, https://doi.org/10.1109/PIC.2017.8359537.
[8] E. Fernandez-Blanco, D. Rivero, A. Pazos, EEG signal processing with separable convolutional neural network for automatic scoring of sleeping stage, Neurocomputing 410 (2020) 220–228, https://doi.org/10.1016/j.neucom.2020.05.085.
[9] V. Krishnamoorthy, R. Shoorangiz, S.J. Weddell, L. Beckert, R.D. Jones, Deep learning with convolutional neural network for detecting microsleep states from EEG: a comparison between the oversampling technique and cost-based learning, Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (2019) 4152–4155, https://doi.org/10.1109/EMBC.2019.8857588.
[10] Z. Liu, G. Yao, Q. Zhang, J. Zhang, X. Zeng, Wavelet scattering transform for ECG beat classification, Comput. Math. Methods Med. 2020 (2020), https://doi.org/10.1155/2020/3215681.
[11] M. Miao, W. Hu, H. Yin, K. Zhang, Spatial-frequency feature learning and classification of motor imagery EEG based on deep convolution neural network, Comput. Math. Methods Med. 2020 (2020), https://doi.org/10.1155/2020/1981728.
[12] J. Andén, S. Mallat, Deep scattering spectrum, IEEE Trans. Signal Process. 62 (2014) 4114–4128, https://doi.org/10.1109/TSP.2014.2326991.

[13] J. Bruna, S. Mallat, Invariant scattering convolution networks, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 1872–1886, https://doi.org/10.1109/TPAMI.2012.230.

[14] J. Andén, S. Mallat, Multiscale scattering for audio classification, Int. Soc. Music Inf. Retr. Conf. (2011) 657–662.

[15] J. Bruna, S. Mallat, Classification with scattering operators, in: IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., IEEE Computer Society, 2011, pp. 1561–1566, https://doi.org/10.1109/CVPR.2011.5995635.

[16] S. Mallat, Understanding deep convolutional networks, Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 374 (2016), https://doi.org/10.1098/rsta.2015.0203.

[17] M.-A. Enoch, D. Goldman, Problem drinking and alcoholism: diagnosis and treatment, Am. Fam. Physician 65 (2002) 441.

[18] S. Patidar, R.B. Pachori, A. Upadhyay, U. Rajendra Acharya, An integrated alcoholic index using tunable-Q wavelet transform based features extracted from EEG signals for diagnosis of alcoholism, Appl. Soft Comput. J. 50 (2017) 71–78, https://doi.org/10.1016/j.asoc.2016.11.002.

[19] J. Kayser, C.E. Tenke, Issues and considerations for using the scalp surface Laplacian in EEG/ERP research: a tutorial review, Int. J. Psychophysiol. 97 (2015) 189–209, https://doi.org/10.1016/j.ijpsycho.2015.04.012.

[20] G. Torres, M.P. Cinelli, A.T. Hynes, I.S. Kaplan, J.R. Leheste, Electroencephalogram mapping of brain states, J. Neurosci. Neuroengineering. 3 (2014) 73–77, https://doi.org/10.1166/jnsne.2014.1098.

[21] A.B. Buriro, R. Shoorangiz, S.J. Weddell, R.D. Jones, Predicting microsleep states using EEG inter-channel relationships, IEEE Trans. Neural Syst. Rehabil. Eng. 26 (2018) 2260–2269, https://doi.org/10.1109/TNSRE.2018.2878587.

[22] N.V.T. Shanbao Tong, Quantitative EEG Analysis Methods and Clinical Applications, Artech House, 2009.

[23] M.Z. Ahmad, A.M. Kamboh, S. Saleem, A.A. Khan, Mallat's scattering transform based anomaly sensing for detection of seizures in scalp EEG, IEEE Access 5 (2017) 16919–16929, https://doi.org/10.1109/ACCESS.2017.2736014.

[24] U.R. Acharya, S.V. Sree, S. Chattopadhyay, J.S. Suri, Automated diagnosis of normal and alcoholic EEG signals, Int. J. Neural Syst. 22 (2012) 1–11, https://doi.org/10.1142/S0129065712500116.

[25] A. Gramfort, D. Strohmeier, J. Haueisen, M.S. Hämäläinen, M. Kowalski, Time-frequency mixed-norm estimates: sparse M/EEG imaging with non-stationary source activations, Neuroimage 70 (2013) 410–422, https://doi.org/10.1016/j.neuroimage.2012.12.051.

[26] S. Saha, M. Baumert, Intra- and inter-subject variability in EEG-based sensorimotor brain computer interface: a review, Front. Comput. Neurosci. 13 (2020) 87, https://doi.org/10.3389/fncom.2019.00087.

[27] S. Siuly, V. Bajaj, A. Sengur, Y. Zhang, An advanced analysis system for identifying alcoholic brain state through EEG signals, Int. J. Autom. Comput. 16 (2019) 737–747, https://doi.org/10.1007/s11633-019-1178-7.

[28] S. Bavkar, B. Iyer, S. Deosarkar, Rapid screening of alcoholism: an EEG based optimal channel selection approach, IEEE Access 7 (2019) 99670–99682, https://doi.org/10.1109/ACCESS.2019.2927267.

[29] M. Sharma, P. Sharma, R.B. Pachori, U.R. Acharya, Dual-tree complex wavelet transform-based features for automated alcoholism identification, Int. J. Fuzzy Syst. 20 (2018) 1297–1308, https://doi.org/10.1007/s40815-018-0455-x.

[30] A. Anuragi, D.S. Sisodia, Alcohol use disorder detection using EEG Signal features and flexible analytical wavelet transform, Biomed. Signal Process Control 52 (2019) 384–393, https://doi.org/10.1016/j.bspc.2018.10.017.

[31] J. das, C. Rodrigues, P.P.R. Filho, E. Peixoto, A.K. N, V.H.C. de Albuquerque, Classification of EEG signals to detect alcoholism using machine learning techniques, Pattern Recogn. Lett. 125 (2019) 140–149, https://doi.org/10.1016/J.PATREC.2019.04.019.

[32] A. Anuragi, D.S. Sisodia, R.B. Pachori, Automated alcoholism detection using Fourier-Bessel series expansion based empirical wavelet transform, IEEE Sensor. J. 20 (2020) 4914–4924, https://doi.org/10.1109/JSEN.2020.2966766.

[33] H. Zhang, F.H.S. Silva, E.F. Ohata, A.M.G. Medeiros, P.P. Rebouças Filho, P.R. F. Pedro, Bi-dimensional approach based on transfer learning for alcoholism predisposition classification via EEG signals, Front. Hum. Neurosci. 14 (2020) 365, https://doi.org/10.3389/FNHUM.2020.00365.

[34] H. Mukhtar, S.M. Qaisar, A. Zaguia, Deep convolutional neural network regularization for alcoholism detection using EEG signals, Sensors 21 (2021) 5456, https://doi.org/10.3390/S21165456.

[35] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T.H. Falk, J. Faubert, Deep learning-based electroencephalography analysis: a systematic review, J. Neural. Eng. 16 (2019) 37pp.

[36] A. Craik, Y. He, J.L. Contreras-Vidal, Deep learning for electroencephalogram (EEG) classification tasks: a review, J. Neural. Eng. 16 (2019) 28, https://doi.org/10.1088/1741-2552/ab0ab5.

[37] UCI machine learning repository: EEG database data set, UCI KDD database (n.d.), https://archive.ics.uci.edu/ml/datasets/eeg+database. (Accessed 21 April 2021). accessed.

[38] J.G. S, M. V, A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity, J. Exp. Psychol. Hum. Learn. 6 (1980) 174–215, https://doi.org/10.1037//0278-7393.6.2.174.

[39] K.-M. Ong, K.-H. Thung, C.-Y. Wee, R. Paramesran, Selection of a subset of EEG channels using PCA to classify alcoholics and non-alcoholics, Annu. Int. Conf. IEEE Eng. Med. Biol. (2005) 4195–4198, https://doi.org/10.1109/IEMBS.2005.1615389.

[40] T.S. Lee, Image representation using 2D Gabor wavelets, IEEE Trans. Pattern Anal. Mach. Intell. 18 (1996) 959–971, https://doi.org/10.1109/34.541406.

[41] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Comput. Electr. Eng. 40 (2014) 16–28, https://doi.org/10.1016/j.compeleceng.2013.11.024.

[42] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, John Wiley & Sons, 2012.

[43] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (2007) 2507–2517, https://doi.org/10.1093/bioinformatics/btm344.

[44] Q. Gu, Z. Li, J. Han, Generalized Fisher score for feature selection, in: 27th Conf. Uncertain. Artif. Intell., AUAI Press, 2011.

[45] T. Fawcett, An introduction to ROC analysis, Pattern Recogn. Lett. 27 (2006) 861–874, https://doi.org/10.1016/j.patrec.2005.10.010.

[46] T. Hastie, R. Tibshirani, J. Friedman, Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed., Springer, New York, NY, USA, 2009.

[47] J. Neumann, C. Schnörr, G. Steidl, Combined SVM-based feature selection and classification, Mach. Learn. 61 (2005) 129–150, https://doi.org/10.1007/s10994-005-1505-9.

[48] Y. Bae, B.W. Yoo, J.C. Lee, H.C. Kim, Automated network analysis to measure brain effective connectivity estimated from EEG data of patients with alcoholism, Physiol. Meas. 38 (2017) 759–773, https://doi.org/10.1088/1361-6579/aa6b4c.

[49] Y. Kumar, M.L. Dewal, R.S. Anand, Epileptic seizure detection using DWT based fuzzy approximate entropy and support vector machine, Neurocomputing 133 (2014) 271–279, https://doi.org/10.1016/j.neucom.2013.11.009.

[50] L.R. Quitadamo, F. Cavrini, L. Sbernini, F. Riillo, L. Bianchi, S. Seri, G. Saggio, Support vector machines to detect physiological patterns for EEG and EMG-based human-computer interaction: a review, J. Neural. Eng. 14 (2017), 011001, https://doi.org/10.1088/1741-2552/14/1/011001.

[51] Z. Wen, J. Shi, Q. Li, J. Chen, ThunderSVM: a fast SVM library on GPUs and CPUs, J. Mach. Learn. Res. 19 (2018) 1–5.

[52] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms, Chapman and Hall/CRC, 2012.

[53] C. Banerjee, T. Mukherjee, E. Pasiliao, An empirical study on generalizations of the ReLU activation function, in: ACM Southeast Conf., Association for Computing Machinery, Inc, 2019, pp. 164–167, https://doi.org/10.1145/3299815.3314450.

[54] D.P. Kingma, J.L. Ba, Adam: a method for stochastic optimization, in: Int. Conf. Learn. Represent., International Conference on Learning Representations, ICLR, 2015.

[55] D.M.W. Powers, Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, ArXiv Prepr. ArXiv2010.16061 2 (2020) 37–63.

[56] J. Huang, C.X. Ling, Using AUC and accuracy in evaluating learning algorithms, IEEE Trans. Knowl. Data Eng. 17 (2005) 299–310, https://doi.org/10.1109/TKDE.2005.50.

[57] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature. 521 (2015) 436–444.

[58] A. Choromanska, M. Henaff, M. Mathieu, G. Ben Arous, Y. Lecun, The loss surfaces of multilayer networks, Int. Conf. Artif. Intell. Stat. (2015) 192–204.

[59] K. Stavro, J. Pelletier, S. Potvin, H. Louis-H Lafontaine, Widespread and sustained cognitive deficits in alcoholism: a meta-analysis, Addict, Biol. 18 (2013) 203–213, https://doi.org/10.1111/j.1369-1600.2011.00418.x.