A Survey of Deep Learning and Traditional Approaches for EEG Signal Processing and Classification

Memoona Iftikhar
Department of computer and software Engineering
National University of science and Technology
Islamabad, Pakistan
memoonaiftikhar760@yahoo.com

Shoab Ahmad Khan[2]
Department of computer and software Engineering
National University of science and Technology
Islamabad, Pakistan
shoabak@ceme.nust.edu.pk

Ali Hassan[3]
Department of computer and software Engineering
National University of science and Technology
Islamabad, Pakistan
alihassan@ce.ceme.edu.pk

Abstract—**Processing of Electroencephalography (EEG) signals help to detect the specific patterns in a person's ongoing brain activity and then translate these patterns into control command. Brain Computer Interfacing (BCI) is a communication system which uses these commands to interact with outside world. BCI's have extensive applications such as to classify motor imagery tasks, and to detect abnormalities and loss of functions of brain. Selection and optimization of the efficient processing technique to extract discriminative features from the EEG signals is a key to develop high performance and robust system. The main purpose of the feature selection and reduction process is threefold: by extracting relevant features computational complexity can be reduced, it improve performance by avoiding over fitting which may be caused due to irrelevant features, selected features will take less time to process. Some novel methods such as time-domain analysis, power spectral estimation, and wavelet transform have been widely used for the evaluation of the selected subset of features. In this paper first, we survey comparative evaluation of various recent techniques for feature selection and classification applied for BCI at each stage. Second, the use of deep learning techniques for brain signal classification is explored.**

Keywords: Brain Computer Interface (BCI), Electroencephalogram (EEG), Signal Classification, Convolutional Neural Network (CNN).

I. INTRODUCTION

Brain Computer interface (BCI) is a method of communication based on the neuronal activity generated by the brain. This paper summarizes various techniques applied for BCI at pre-processing, feature extraction and classification stage. BCI systems typically consist of six steps: EEG measurement, pre-processing of signal, feature extraction and reduction, classification, interpretation of command and feedback as shown in Fig. 1. The electrical signals of brain are controlled by human intentions and to record these signals effectively is a crucial step. These raw EEG signals are recorded through various types of electrodes which can be invasive or non-invasive and are converted to digital signals for further processing. Once the EEG is recorded accurately the next step is to put the signals for further operations desired to make the signals artefact free and improving signals to noise ratio and boosting the relevant information which is embedded in the signal, and then most discriminative features are extracted from the signal. Once the relevant features are extracted, the next step is to assign particular class to these features to which they belong and passing the command to external devices outputting the intention of the subject [1].

The useful information of brain signals lies in five major brain waves. It is possible to distinguish between these functional states of brain by using EEG signals which separates these brain waves in different frequency bands. The frequency band delta lies in (0–4 Hz) which arise during deep sleep state, Theta band plays role in deepest state of mediation which lies in (3.5–7.5 Hz), alpha band related to the case of dreaming and relaxation having frequency band of (7.5–13 Hz), in case of high attention beta band (13–26 Hz) is dominant, and finally while making decision gamma band (26–70 Hz) is highly dominant. When dealing with mental illnesses states, unexpected disturbances of the brain waves occur leading to the need of considerable signal processing burdens for diagnosis of abnormal states [5].

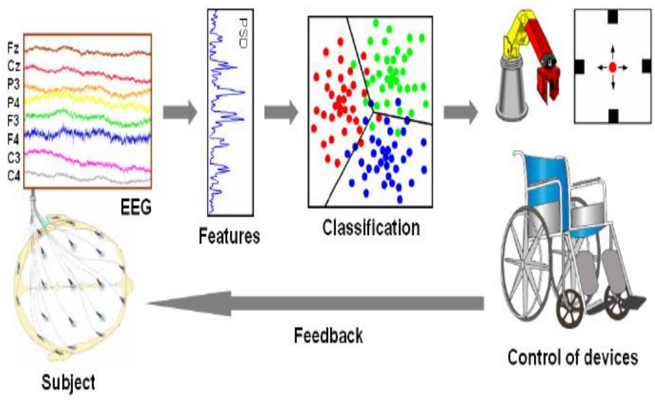

Fig. 1. Fig. 1. General Architecture of BCI [4]

II. PRE-PROCESSING OF EEG SIGNAL

When EEG signals are collected through procurement apparatus then there are some artefacts and noises such as heart beats, eye blinks and other disturbed factors are commonly present in them. To remove such interrupts pre-processing should be done by using following well known filtering techniques.

A. CSP

To operate motor imagery based BCI Common spatial pattern technique is used. EEG signals pattern detection is done through spatial filters due to this variance of each task is either increased or decreased with respect to other task. But, multiple electrodes must be present to utilize this method in an efficient way and similarly by varying the position of electrode results of classification can be enhanced [4].

B. PCA

Principal Component Analysis (PCA) is used for analyzing and dimension reduction of data without losing information. It uses mathematical procedures that utilize orthogonal transformation. These transformations change a set of correlated vectors into a set of linearly uncorrelated vectors called principal components [4].

C. CAR

Common Average Referencing is a spatial filtering technique that removes the noise from position of interest. This noise can be considered as common activity of EEG signal [5]. The artefacts present in the EEG signal causes low signal to noise ratio which can be improved by referencing approaches. Mean of all electrodes will be eliminated in this approach to produce a noise free signal. The classification done by car has good accuracy result in all referencing methods. In common referencing methods problem can be caused in the calculation of the average of electrode due to "finite sample density" and inadequate head coverage of EEG [4].

D. SL

Surface Laplacian of the skull is the estimation of the present density ingoing or departing the scalp via skull. This approach is based on the actual spatial which is used in BCI studies. SL technique can be used for the identification of sources and for the improvement of signal to noise ration of signal [6]. EEG signal can be observed with high spatial resolution using this technique. Electrode cap generates some artefacts at the uncovered points which can be tackled by this method.SL is robust against problems and can solve the electrode referencing problems. The limitation of the method is that when the spline interpolation is carried out it is excessively sensitive to the choice of spline parameters [4].

E. Adaptive Filtering

Adaptive filtering is a filtering technique that models the relationship between two signals iteratively. Adaptive filtering consists of four parts: the input signal diluted with noise, the relationship between input and output signal, the parameter that varies with respect to the input and output of filters. A new adaptive filtering design which outperform for sensorimotor rhythms (SMR) based BCI has been introduced in [3] which are known as adaptive Laplacian (ALAP) filter. An ALAP filter works as a Gaussian filter which construct a smooth spatial gradient of channel weights. The optimal kernel radius and regularization parameters of spatial filter are also obtained in this technique. This optimization is based on gradient descent method which minimizes the leave-one-out cross-validation error. The benefit of adaptive filters is that it works well for the signals with overlapping spectra and is able to modify signal features according to signal being analysed [4].

F. ICA

Noise from the EEG signals is split into independent components using ICA i.e. Independent Component Analysis on the basis of data characteristic without depending on the reference channel. Multi-channel EEG data decomposed into two components which are spatial-fixed and other is temporal independent using ICA algorithm in [4]. This algorithm is computationally efficient and has capability for the particular class of signals to remove the artefacts and has ability of source extraction. ICA has better performance than PCA when the data we want to decompose is larger in size [4].

G. SVD

SVD is the more generalized form of the eigenvalue decomposition (EVD). SVD is the most stable process to divide the system into linearly independent set of modules which have their specific energy influence. SVD has countless benefits, it is capable to change the image in base of two unique subspaces which are noise subspace and data subspace that are used for the filtering of noise and it has maximum energy packing advantage too [7].

H. Digital Filters

In artefact handling, specifically for the removal of muscle artefacts digital filters like band-pass, low pass and high pass which are frequency selective are used. This has the limitation that this technique needs EEG signal and artefact that inhabit distinctive frequency bands which do not exist in reality [8]. Band pass and notch filters are the commonly used digital filters used in EEG signal [5]. Noise produced in electrical line which is caused by the grounding of EEG can be easily removed by the notch filter [9].
In Table I comparison of the pre-processing approaches are shown.

III. Feature Extraction Method

Using different techniques like Independent Component Analysis (ICA), Fast Fourier Transform (FFT), Wavelet Transformations (WT), Auto Regressive (AR), Principal Component Analysis (PCA) and Wavelet Packet Decomposition (WPD) useful features can be extracted from the pre-processed EEG signal [7].

A. PCA

In the preprocessing method "Principal Component Analysis" (PCA) is used, and it is also used in features extraction stage .Without loss of information data can be analyzed and dimensions of data can be reduced using this technique. The information at different time series multi-channel can be extracted using this technique. Principle components can be produced and signal dimension can be reduced by the artefacts removal.

TABLE 1. LIST OF PRE-PROCESSING APPROACHES

No.	Method	Advantages	Disadvantages
1.	CSP	Motor Imagery data can be handled.	Numerous electrodes required.
2.	PCA	Dimension of feature can be reduced.	Less efficient than ICA.
3.	CAR	Performs well as compared to other reference methods.	Needs adequate analysis.
4.	SL	Vigorous against the artefacts at bare conductor.	Artefacts sensitive. Spline pattern sensitive.
5.	ICA	Computationally proficient. Large data can be handled.	Computationally complex for decomposition.
6.	SVD	Stable and Maximum energy packing.	
7.	Adaptive filtering	Overlapped spectrum can be handled.	Require binary signals (reference signal included).
8.	Digital Filters	Easily Noise removal.	Distinct Frequency band Required for EEG signal and artefact.

PCA feature extraction, has advantage over the other extraction techniques that the performance can be improved [10]. The limitation of the technique is that it assume the data is linear and continuous and it is not feasible for the processing of data with complicated set of features [7].

B. ICA

Independent component analysis is one of the best techniques for pre-processing of EEG signals and for feature extraction. There are different components of ICA which are totally independent of each other and through these components features are extracted .Blind source separation is the best example of ICA.it is developed using ICA and it is also used to identify signals and to separate noise from brain signals. ICA is computationally very efficient and it is also very helpful to calculate high performance for EEG signals. [3].

C. WT

WT is Wavelet Transform and it is a mathematical technique used on a very large scale for separating information from different types of continuous data like images and speech. Non-stationary signals are processed by using Wavelet transform technique because of representation of signals in time-frequency domain [10]. There are various advantages of WT e.g. analysis of signal with discontinuity using variable window size, localization of signal in frequency and time domain, Extraction of energy, cluster or distance but one of the major pitfall is the lack of suitable method to the pervasive noise application [3].

D. WPD

WPD is Wavelet Packet Decomposition (WPD), it is the advance form of WD (Wavelet decomposition).WD is different from WPD in some manners, as WD is used to get approximation and all the information in details at higher stage by breaking down approximation at each separate resolution stage. Moreover, when comparison is being done, it is observed that a precise and efficient frequency resolution is obtained by WPD technique instead of WD [18]. By using WT coefficients mean, WPD is able to get features in both domains i.e. frequency and time. WPD gives better results in extraction method in case of non-stationary signals but it can cause computation time to increase. [3].

E. Parametric Approach

Parametric approaches include auto regressive (AR) method. One of the best usages of AR is to process EEG signals through signal representation of linear combination at previous levels at respective separate channel [11].it is helpful in time domain for feature extraction technique. To increase best frequency resolution and to minimize spectral loss, AR is used with reduced data records time duration. When input parameters are AR type like in non-stationary EEG signals then this approach gives the best results. There are some disadvantages too in this approach one of them is the establishment of AR parameter model characteristic [3].

F. FFT

FFT is Fast Fourier Transform (FFT). FFT applies mathematical technique to EEG data. Power spectral density (PSD) estimation method is used to calculate characteristics of the obtained EEG signals which have to be analyzed for the purpose of EEG sample signals representation. Signal's features are extracted by converting the signal from time domain to frequency domain. Linear random process and stationary signals give better performance by using FFT. However, non-stationary signals are not processed by FFT and it doesn't compute both time and frequency [3].

G. STFT

The STFT uses FFT to analyses signal in small contents, which is achieved by dividing a signal into short consecutive segments and then computing the Fourier coefficient of each segment. But still, it is a compromise between the time resolution and frequency resolution during the procedure. Whether imprecise the time or the frequency representation, is determined by the window size. The Fourier transform can represent the frequency localization, but lost all information in time domain. While, short-time Fourier transform, as an evolved scheme of Fourier transform, has equal-length intervals in the time domain. STFT has its limit to analyze the signals with low frequencies, because of low-frequency signals take a long time interval to get the whole shape of one period, and if the length of window function cannot cover the whole period, it cannot get significance to estimate the wave-shape of whole period [12].

Table II shows different feature extraction approaches and their advantages and disadvantages.

IV. CLASSIFICATION

The extracted features are fed to some classifier so that similar extracted features can be placed in one class.

A. Traditional Approaches

The majority of published techniques currently use some form of classification techniques with a relatively simple model. These techniques require less implementation time than deep learning, and are less prone to overfitting. These traditional approaches may include following classifiers [13].

1) *LDA:* Linear Discriminant Analysis is a classification technique which produces models that have quite good accuracy as compared to the complex models. It has very low mathematical requirements. It creates linear combination of variables that best splits two classes and that linear combination of variables are presented in form of model of the Probability density function [13]. The complex structure of the data may not preserve if the data is non Gaussian distribution. If the discriminatory function is based on variance of the data instead of mean LDA fails it performance [14].

2) SVM: Support vector machine is a linear classifier and one of the exciting algorithms in machine learning and used by most of the BCI. It was first introduced by Vapnik and was determined by principal of statistical learning theory and structural risk minimization [13]. SVM searches a line that creates a clear gap among the dataset and split it into designated category. The line which splits the data is called hyperplane, it capitalizes the distance between the hyper plane and the points that are nearer from each class that are called as support vectors. The motivation for this method is to reduce the complexity of the learned model and improving performance [13]. Due to this, SVM provides good generalisation.

3) *NB*C: The naïve base classifier produces nonlinear decision boundary. It is based on Bayesian theorem through which posterior probabilities of feature vectors are calculated and assigns the features vector to its class with a highest posterior probability. They often outperform well as compared to the discriminative classifiers and reject uncertain samples more accurately. Hidden Markov Model (HMM) is a Bayesian classifier and it is suitable for the classification of time series. These are not as prevalent as Linear Classifiers in the field of BCI applications [22]. They are suitable when the dimensions are high.

4) *NNC:* Nearest neighbor classifiers is a non-parametric method of classification, it stores all patterns first and then search for the most nearest neighbor of the pattern and assign to it a class. k-Nearest Neighbor (k-NN) classifiers is the vibrant of the NNC. If the feature vector is from the training set then it is called as k-Nearest Neighbor classifiers. It assigns the class membership to the feature vector based on k closest patterns in the feature space. It assigns the label of a test sample with the majority label of its k-nearest neighbors from the training set. K-NN is transparent and easy to understand, and its implementation is quite easy as compared to other techniques [15].

5) *ANN:* Artificial Neural Network is a nonlinear classification approach and is composed of large number of connected cells called neurons. Multi-Layer Perceptron Neural Network (MLPNN) is the most frequently used neural network, there are three layers in MLPN viz., input layer, hidden layer and output layer. The input to the first layer is actually denotes the number of selected features and the output layer shows the number of labels or classes. The AAN will be complex if the hidden neurons are complex. The hidden layer also defines the accuracy of classification. If less number of hidden layers is used there might be a chance for classification error. There is no specific criterion for selection of hidden neurons it is selected by hit and trial method [13] [16].

TABLE II. FEATURE EXTRACTION TECHNIQUES

No.	Method	Advantages	Disadvantages
1.	PCA	Performs Dimensionality reduction without any loss of data.	Unable to process complicated set of data
2.	ICA	Performs well for large sized data. Low computations.	Decomposition of data requires high more computations.
3.	WT	Analyzes time and frequency aspect of signal. Particularly Fit for nonstationary signals.	No specific method to remove prevalent noise.
4.	WPD	Performs analysis of nonstationary signals.	Computational time increases
5.	Parametric approaches	Time localization. Frequency localization without any spectra loses.	Not applicable to stationary signals
6.	FFT	Performs well for stationary signals	Doesn't performs for non-stationary signals Unable to localize both time and frequency
7.	STFT	Give both time and frequency localization	Difficult to select appropriate window size

A. Deep Learning Approaches

Deep learning is a subfield of machine learning that has evolved out of the traditional approaches to artificial neural networks. With the advent of backpropagation, neural networks began to see renewed interest and significant theoretical advancement in the form of recurrent neural networks (RNN), Convolutional neural networks (CNN), Long Short-Term Memory networks (LSTM) and Gradient decent (SGD).

1) *CNN:* Convolutional neural networks learn filter banks that are convolved with the original data. The filters can also be represented as a fully connected layer where the weights of the edges are tied together in way that replicates the convolution operation. This weight sharing structure allows for fewer parameters than having each weight be unique, and directly accounts for structure in the data. Each filter creates a new, processed version of the image.

2) *RNN:* Recurrent neural networks contributed a lot to the success of deep learning in various fields. It outperforms for time series classification because of its intra layer connections and inters layer connections. Training an RNN is a difficult task. Backpropagation must be modified to function in RNNs, since there are cycles in the graph. This is frequently handled through a technique known as backpropagation through time, wherein the network is "unrolled" for a discrete number of steps. This process creates a network with only forward connections, allowing backpropagation to work as normal, at the cost of limiting the impact of the recurrent connections.

3) *LSTM:* Long Short-Term Memory networks have proven successful for sequential data classification because they are specifically designed to classify for each time stamp. At each time point data is processed by individual LSTM unit. The processed result is fed to the next layer and as well as well as remains within the same layer for the processing of next time stamp. The information passed through to the next layer is passed through an activation function [17]. However, the recurrent connection within the layer is not subjected to an activation function. One of the vibrant problems in LSTM is the lack of activation function in recurrent layers to avoid the vanishing gradient problem. Due to this gradient problem it gives gradient a constant value of one. Each LSTM unit has a number of gates which control the flow of information. A gate is a combination of a sigmoidal activation unit and pointwise multiplication. The flow of information from one time stamp to another time stamp is controlled by these gates.

4) *SGD*: It is a derivative of traditional gradient descent, differing in that the error function is calculated using only a subsample of the available data. This is both easier to use and more efficient for training datasets that do not fit in memory. Furthermore, adding randomness to the optimization can aid in breaking through plateaus and avoiding local minima. The addition of momentum terms, which biases the gradient in the direction of recently calculated gradients, greatly improved the ability to train deep models by further increasing speed of convergence [18].

Most studies focus on using machine learning to answer a neuroscience question, rather than on improving the machine learning techniques themselves. Generally only a single technique is applied to the classification of a dataset and Individual datasets have highly varying characteristics, and some are simply harder to classify than others, so it is difficult to draw conclusions on the effectiveness classification system as compared to another.

V. Conclusion

A clear understanding of various signal processing techniques is described in this paper. This survey will give the researchers a way to select the appropriate signal processing method according to their need. EEG classification using deep learning approach is still in its primary stages. There is a need to find architecture and training paradigms for deep earning methods in order to improve classification. However, as research continues into the use of deep learning for classification of EEG signals, best practices will become well known.

REFERENCES

[1] N. B. D. M. G. P. T. V. JR Wolpaw, " Brain-computer interfaces for communication and control," in *Clinical Neurophysiol*, New York State, Elsevier, 2002, pp. 113(6), 767–791 .

[2] T. A. P. J. S. Ebersole, Current Practice of Clinical Electroencephalography, Lippincott Williams & Wilkins, , 2003.

[3] P. W. F. F. G. E. L. JOSÉ DEL R. MILLÁN, "Non-Invasive Brain-Machine Interaction," *International Journal of Pattern Recognition and Artificial Intelligence,* vol. 2, no. 05, August 2008.

[4] D. T. V. P. a. D. V. C. P. M. R. Lakshmi, ""Survey on EEG Signal Processing Methods," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 4, pp. 84-91, 2014.

[5] K. R. W. C. C. J. R. R. S. T. C. B. M. A. D. H. R. M. a. M. K. Leonard J. Trejo, "Multimodal Neuroelectric Interface Development," *IEEE Transactions on Neural Systems and Rehabilitation,* vol. 11, no. 2, June 2003 .

[6] A. R. L. A. a. R. F.-R. S. Amiri, "A Review of P300, SSVEP, and Hybrid P300/SSVEP Brain- Computer Interface Systems," 2013.

[7] D. J. McFarland, "The advantages of the surface Laplacian in brain– computer interface research," *International Journal of Psychophysiology,* 2014.

[8] D. J. M. a. J. R. W. J. Lu, "Adaptive Laplacian filtering for sensorimotor rhythm-based brain–computer interfaces," *Journal of neural engineering,* vol. 10, p. 016002, 2013..

[9] N. M. G. ,. 2. e. B. J. Baars, Cognition, Brain, and Consciousness: Introduction to Cognitive Neuroscience, Elsevier, 2010.

[10] M. M.-T. M. H. C. M. H. a. P. R. W. S. Motamedi-Fakhr, ""Signal processing techniques applied to human sleep EEG signals—A review," Biomedical Signal Processing and Control," *Computational Intelligence and Neuroscience,* vol. vol. 1, no. 3, pp. pp. 21-33, , 2014.

[11] F. L. F. a. F. C. M. N. Mammone, "Automatic Artifact Rejection From Multichannel Scalp EEG by Wavelet ICA," *IEEE,* Vols. vol. 12, , pp. 533-542, , 2012..

[12] P. C. a. K.-B. S. X. Yu, "Analysis the effect of PCA for feature reduction in non-stationary EEG based motor imagery of BCI system," *Optik - International Journal for Light and Electron Optics, ,* vol. vol. 125, pp. 1498-1502, 2014.

[13] L. F. Nicolas-Alonso and J. Gomez-Gil, ""Brain computer interfaces, a review,"," *Sensors,* vol. 12, pp. 1211-1279, 2012.

[14] W. D. H. K. M. M. W. a. K. R. V. Lawhern, "Detection and classification of subject-generated artifacts in EEG signals using autoregressive models," *Journal of Neuroscience Methods,* vol. vol. 208, pp. 181-189, 2012.

[15] H. C. &. Y. ZhangShuanghu., "Analysis Between Fourier Transform and Wavelet Transform.," *HeiBei: Journal of Hebei University of Technology.,* 2005.

[16] L. a. D. Senthilmurugan.M, " Classification in EEG-Based Brain Computer Interfaces Using Inverse Model,," *Internatiional Journel of Computer Theory and Engg., ,* vol. Vol. 3, p. 2, April 2011..

[17] S. B. M. C. a. C. J. Alexandre Barachant, "BCI signal classification using a Riemannian-based kernel," *Proc. of the European Symp. on Artificial Neural Networks, Computational Intell. and Machine Learning (ESANN). ,* 2012.

[18] I. M. D. G. &. H. G. Sutskever, " On the importance of initialization and momentum in deep learning," in *Proceedings of the 30 th International Conference on Machine Learning*, Atlanta, Georgia, USA, , 2013.

[19] S. a. S. J. Hochreiter, "Long short-term memory," Fakult•at f•ur Informatik, , Technische Universit•at M•unchen , 1995.

[20] F. Lotte, " Study of Electroencephalographic Signal Processing and Classification Techniques towards the use of Brain-Computer Interfaces in Virtual Reality Applications. Human-Computer Interaction," HAL , INSA de Rennes, 2009.

[21] M. HEKIM, "ANN-based classification of EEG signals using the average power based on rectangle approximation window," Przegląd Elektrotechniczny (Electrical Review), 2012 .

[22] A. K. V. C. K. R. T. a. A. K. P. K. Padhy, " Feature Extraction and Classification of Brain Signal," World Acad. of Sc., Engg. and Tech, 2011.