Tikhonov Kullback Leibler Vuong Logistic Machine Learning Classifier for Early Disease Diagnosis Over Big Data

DR. A. KALIAPPAN, DR. N. K. PRIYADHARSINI, DR. R. KAVITHA
Assistant Professor(s), Dept. of CSE, P. A. College of Engineering and Technology, Coimbatore, India

Abstract

With big data widening in healthcare groups, precise investigation of medical data conveniences early disease detection. However, the analysis accuracy is reduced when the parallel processing of medical data is not performed. Moreover, with curse of dimensionality as several regions discloses distinctive facets and if not properly filtered, relevant information's are also discarded which may reduce the early prediction of disease outbreaks. To address these issues, in this work, a method using machine learning technique called, Polynomial Tikhonov Entropy and Kullback Vuong Logistic Classifier (PTE-KVLC) is presented. First, Inverse Polynomial Map Reduce Pre-processing is applied to the input data that both minimizes the signal to noise ratio and obtains computationally efficient features via parallel processing. This is turn provides a mean for early detection of epileptic seizures. Second, the feature extraction model is based on Entropy Tikhonov Regularization and is applied to the pre-processed features to identify the features pertinent to seizures. These features are then selected and fed into a Kullback–Leibler Vuong and Logistic Regressive Machine Learning Classifier for early epileptic seizure recognition. Experimental results demonstrate that the proposed method significantly classifies the epileptic seizure classes by means of specificity, sensitivity, and accuracy.

Keywords: *Big Data, Map Reduce, Inverse Polynomial, Tikhonov Entropy, Kullback–Leibler, Vuong, Logistic Regressive*

INTRODUCTION

Due to the rise of healthcare expenditures, early disease prevention has never been important as it is today (Nagavei *et al.* 2018). This is particularly due to the increased threats of new disease variants, bio-terrorism as well as recent improved developments in data collection and computing technology. Increased amount of healthcare data increases the demand to develop an efficient, sensitive and cost-effective solution for disease prevention (Quintero-Rincon *et al.* 2018).

Epilepsy has emerged as a severe disease in recent years, characterised by a persistent risk of developing epileptic seizures. A seizure is a transient occurrence of symptoms or signs due to abnormal excessive or synchronous neuronal activity in the brain (Ahmadi *et al.* 2020). A seizure does not necessarily mean that a person has epilepsy, unless the criteria for diagnosis of epilepsy are met. As there are a number of conditions that can be associated with paroxysmal events that can mimic seizures or epilepsy, described in the section 'epilepsy imitators' in Epilepsy Diagnosis, these should be carefully excluded (Namazi *et al.* 2020). If not properly treated in the early stage also results in severe health issues and even some times to mortality.

The traditional seizure detection method mainly focusses on promotion of healthcare benefits. Presently, Deep Neural Network (DNN) with dataset normalization accurately predicted disease with minimum error (Thara *et al.* 2019). In Successive Decomposition Index (SDI), false

https://doi.org/10.37896/pd91.4/91443

positive rate was reduced while detecting epileptic seizure (Raghu *et al.* 2020). Channel-Embedding spectral-temporal Squeeze-and-Excitation Network (CE-stSENet) aided by a maximum mean discrepancy-based information maximizing loss (Li *et al.* 2020). The visual inspection using deep neural network consumes more time and with a lack of parallel processing, running time of overall process is found to be tedious.

A method called, Polynomial Tikhonov Entropy and Kullback–Leibler Vuong Logistic Classifier (PTE-KVLC) is presented to address the above said issues. First, Inverse Polynomial Map Reduce pre-processing is applied to the input data that minimizes the signal to noise ratio and obtains computationally efficient features via parallel processing. This is turn provides a mean for early detection of epileptic seizures. Second, the feature extraction model based on Entropy Tikhonov Regularization is applied to the pre-processed features to identify the features pertinent to seizures. The features are selected and fed into a Kullback–Leibler Vuong Logistic Regressive Classifier for early epileptic seizure recognition.

Our paper proposes the use of a machine learning classified based on logistic regressive model for early epileptic disease diagnosis and evaluates several feature data points to generate the inputs to the logistic regressive model. The best previous results were obtained from the automated seizure detection using deep neural network [18] and successive decomposition index (SDI) [14]. These previous works are our baseline in this paper.

OVERVIEW OF MAP REDUCE BASED PRE-PROCESSING

The MapReduce based preprocessing step for improving the performance of the classifier is presented. Preprocessing and feature extraction from EEG signal have great affect in maximizing prediction time and True Positive Rate (TPR). Preprocessing is performed for removing noise from the signals and to increase the Signal-to-Noise Ratio (SNR) with the help of filtering techniques (Chu *et al.* 2007). When medical data is not processed in parallel, the accuracy of the analysis is reduced. Furthermore, with the curse of dimensionality, several regions reveal distinct features, and if not correctly filtered, essential information is also eliminated, potentially reducing early disease outbreak prediction (Gillick *et al.* 2006). This provides a parallel preprocessing strategy for decreasing the classifier's complexity.

MapReduce is a parallel programming model consists of two functions Mapper and Reducer, runs on all machines in a cluster. The input and output of these functions must be in form of key, value pairs. The Mapper takes the input (k1, v1) pairs from Distributed File System (DFS) and produces a list of intermediate (k2, v2) pairs. An optional Combiner function is applied to reduce communication cost of transferring intermediate outputs of mappers to reducers. Output pairs of mapper are locally sorted and grouped on same key and feed to the combiner to make local sum (Low *et al.* 2014). The intermediate output pairs of combiners are shuffled and exchanged between machines to group all the pairs with the same key to a single reducer. This is the only communication step that occurs and is handled by the MapReduce platform. There is no other way for mappers and reducers to communicate. The Reducer takes (k2, list (v2)) values as input, make sum of the values in list (v2) and produce new pairs (k3, v3). Figure 3.1 illustrates the work flow of MapReduce.

MapReduce is a simplified programming model since all the parallelization, inter machine communication and fault tolerance are handled by run-time system. The Inverse Polynomial Map Reduce pre-processing is applied to the input data that both minimizes the signal to noise ratio and obtains computationally efficient features via parallel processing. This is turn provides a mean for early detection of epileptic seizures.

BASICS OF ENTROPY TIKHONOV REGULARIZATION

Feature extraction is a part of the dimensionality reduction process, in which, an initial set of the raw data is divided and reduced to more manageable groups. The most important characteristic of large data sets is the large number of variables (Nixon & Aguado 2019). These variables require a lot of computing resources to process them. So, Feature extraction helps to get the best feature from those big data sets by selecting and combining variables into features, thus, effectively reducing the amount of data. These features are easy to process, but still able to describe the actual data set with the accuracy and originality. Figure 1 shows the MapReduce model.

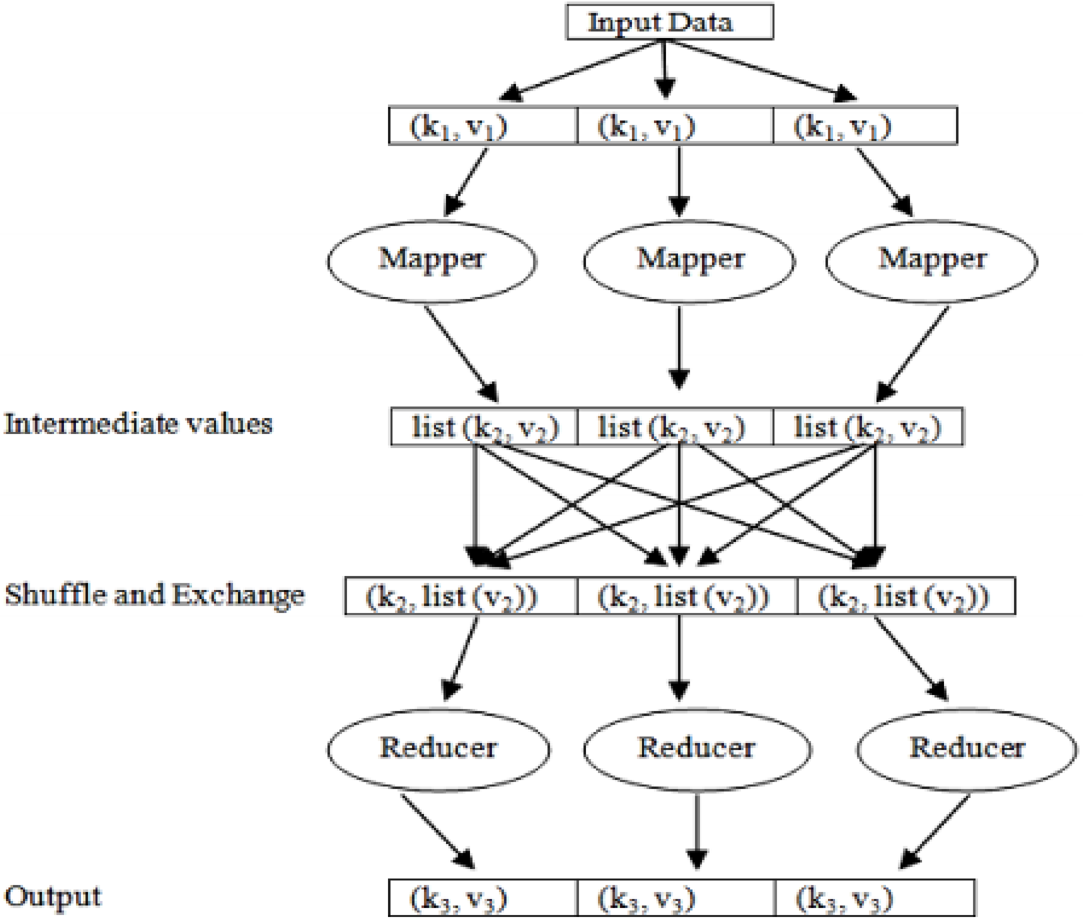

Figure 1 MapReduce Model

Entropy Tikhonov Regularization based feature selection approach has been proposed for increasing the classifier`s performance. Tikhonov based regularization is a well-known technique, where smoothness of the unknown function is searched (Calvetti & Reichel 2003). Similarly, to Tikhonov's regularization, the maximum entropy formalism searches for global regularity and yields the smoothest reconstructions that are consistent with the available data. The maximum entropy principle has been proposed as a general inference on the basis of Shannon's axiomatic characterization of the amount of information. This principle has successfully been applied to a variety of fields.

Recently, the non-extensive entropic form (S_q) has been used as a new regularization operator, using only q = 0.5. The q parameter plays a central role in the Tsallis' thermo statistics, where q = 1 the Boltzmann-Gibbs-Shannon's entropy is recovered. As mentioned, the non-extensive entropy includes as a particular case the extensive entropy (q = 1): where the maximum entropy principle can be used as a regularization method. Another important particular case of the non-extensive entropy occurs when q = 2. In such case, the maximum non-extensive entropy principle to the S_2 regularization operator is equivalent to the standard Tikhonov regularization or zeroth-order Tikhonov regularization.

Two methods were investigated for determining the regularization parameter: The Morozov's discrepancy principle, and the maximum curvature scheme of the curve relating smoothness versus fidelity, inspired in Hansen's geometrical criterion. The regularization techniques are effective methods to deal with the ill-posed problems. In recent years, entropy-based regularization techniques have been proposed one after the other, and this technique have been successfully applied to disease prediction. The Entropy Tikhonov Regularization based feature selection approach is proposed.

PRINCIPLE OF LOGISTIC CLASSIFIER

Detecting the appearance of preictal state predicts the seizure. Therefore, the purpose of this investigation is to detect the appearance of preictal state for epileptic seizures. Machine learning models are used to predict epileptic seizures. These machine learning models include EEG signal acquisition, signal preprocessing, features extraction from the signals, and finally classification between different seizure states. The objective of the prediction model with machine learning was to detect preictal state’s sufficient time before seizure onset starts.

Logistic Regression is used for the classification problems, it is a predictive analysis algorithm, and based on the concept of probability. Logistic regression transforms its output using the logistic sigmoid function to return a probability value. Logistic regression analysis studies the association between a categorical dependent variable and a set of independent explanatory variables (Field 2009). The name logistic regression is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. Logistic regression competes with discriminant analysis as a method for analyzing categorical-response variables. Many statisticians feel that logistic regression is more versatile and better suited for modeling most situations than is discriminant analysis. This is because logistic regression does not assume that the independent variables are normally distributed as the discriminant analysis does. Logistic regression performs a comprehensive residual analysis including diagnostic residual reports and plots. It can perform an independent variable subset selection search, looking for the best regression model with the fewest independent variables. It provides confidence intervals on predicted values and provides Receiver Operating Characteristic (ROC) curve to help determine the best cut-off point for classification. The proposed research work Kullback–Leibler Vuong Logistic Regressive Machine Learning Classifier integrates log likelihood ratio with Kullback–Leibler function, therefore providing accurate results.

POLYNOMIAL TIKHONOV ENTROPY AND KULLBACK-LEIBLER VUONG LOGISTIC CLASSIFIER

Polynomial Tikhonov Entropy and Kullback-Leibler Vuong Logistic Classifier (PTE-KVLC) method for early disease diagnosis is proposed and splitted into three sections. All samples were shuffled into different chunks with each chunk comprising different data points at distinct time periods and sent as input. In addition, the machine learning classifier based on Logistic is constructed to perform early disease diagnosis.

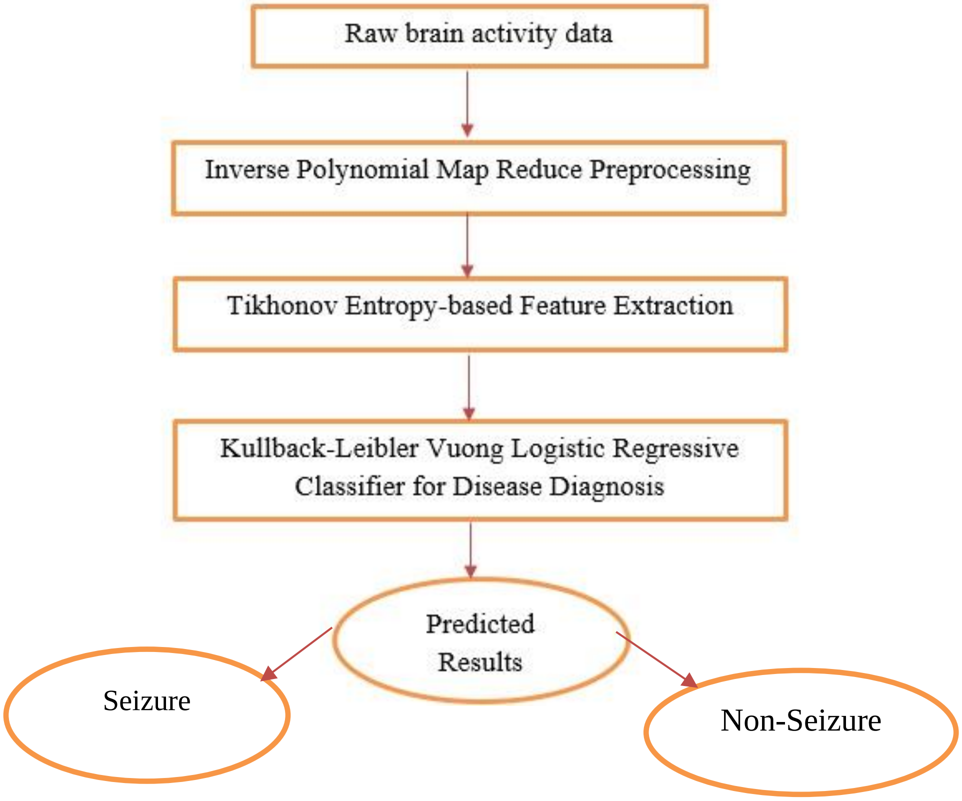
Raw brain activity data
Inverse Polynomial Map Reduce Preprocessing
Tikhonov Entropy-based Feature Extraction
Kullback-Leibler Vuong Logistic Regressive Classifier for Disease Diagnosis
Predicted Results
Seizure
Non-Seizure

Figure 2 Block diagram of Polynomial Tikhonov Entropy and Kullback-Leibler Vuong Logistic Classifier (PTE-KVLC)

The steps involved in the PTE-KVLC method are as follows:

i. An efficient model, called, Inverse Polynomial Map Reduce is proposed to pre-process raw brain activity suitable for logistic classifier.

ii. The robust feature extraction is done by combining permutation-based entropy and Tikhonov regularization function into the pre-processed features to help the classifier perform well, therefore paving way for early disease diagnosis.
iii. Early disease diagnosis is performed via Kullback–Leibler Vuong Logistic Regressive Classifier that integrates log likelihood ratio with Kullback–Leibler function, therefore providing accurate results.

PTE-KVLC method is applied to brain activity data for early disease diagnosis. The proposed PTE-KVLC method comprises of three phases; pre-processing, feature extraction and classification for epileptic disease diagnosis. The working procedure of PTE-KVLC is depicted in Figure 2.

First, Inverse Polynomial Map Reduce pre-processing is applied to the brain activity dataset. Here, pre-processing is applied to both minimize the signal to noise ratio and obtained computationally efficient features. Second, an entropy-based automated features extraction or automated data point extraction using Tikhonov Entropy-based Feature Extraction model is used for extracting features or data points with high optimal class variances. Once the features or data points have been extracted, finally, classification between seizure recognition is performed by applying KVLC.

Dataset Description

The Epileptic Seizure Recognition Data Set is a commonly used dataset featuring epileptic seizure detection. The dataset from reference comprises 5 different folders, each with 100 files. Each file represents a single person with recording of brain activity for 23.6 seconds. The corresponding time-series is sampled into 4097 data points. Each data point is the value of the EEG recording at a different point in time, with a total of 500 individuals with each has 4097 data points for 23.5 seconds. The dataset is divided and shuffled every 4097 data points into 23 chunks, each chunk comprises 178 data points for 1 second, and each data point is the value of the EEG recording at different point in time. So, a total of 23 x 500 = 11500 pieces of information represented in rows. Each information contains 178 data points for 1 second represented in columns. The last column represents the label y {1,2,3,4,5}. The response variable is y in column 179, the Explanatory variables are X1, X2, ….., X178 and y contains the category of the 178-dimensional input vector.

Specifically, y in {1, 2, 3, 4,5}: 5 - eyes open, EEG signal of the brain has been recorded when the patient`s eyes are open, 4 - EEG signal of the brain has been recorded when the patient`s eyes are closed, 3 - Identifies region of the tumor in the brain and recording the EEG activity from the healthy brain area, 2 - EEG from the area where the tumor is located, 1 - Recording of seizure activity. All of the subjects in classes 2, 3, 4, and 5 have never had an epileptic seizure. Only class 1 subjects develop epileptic seizures.

Comparison analysis of four metrics to evaluate the disease diagnosis performance by means of Sensitivity, Specificity, Accuracy and Time complexity is evaluated. An elaborate comparison is made with the proposed Polynomial Tikhonov Entropy and Kullback Vuong Logistic Classifier (PTE-KVLC) against the existing CE-stSENet and DNN.

Inverse Polynomial Map Reduce Pre-processing

During the acquisition of raw brain activity data, significant amount of noise is found to be added that minimizes the signal to noise ratio resulting in poor classification or compromising early detection. Different types of noise, like, power line noise, baseline noise electrical activity or human activities including eye movement and pulse of heart compromises the overall result. Hence, it becomes paramount to remove noise as preprocessing step from epileptic in order to increase Signal to Noise ratio for early disease detection.

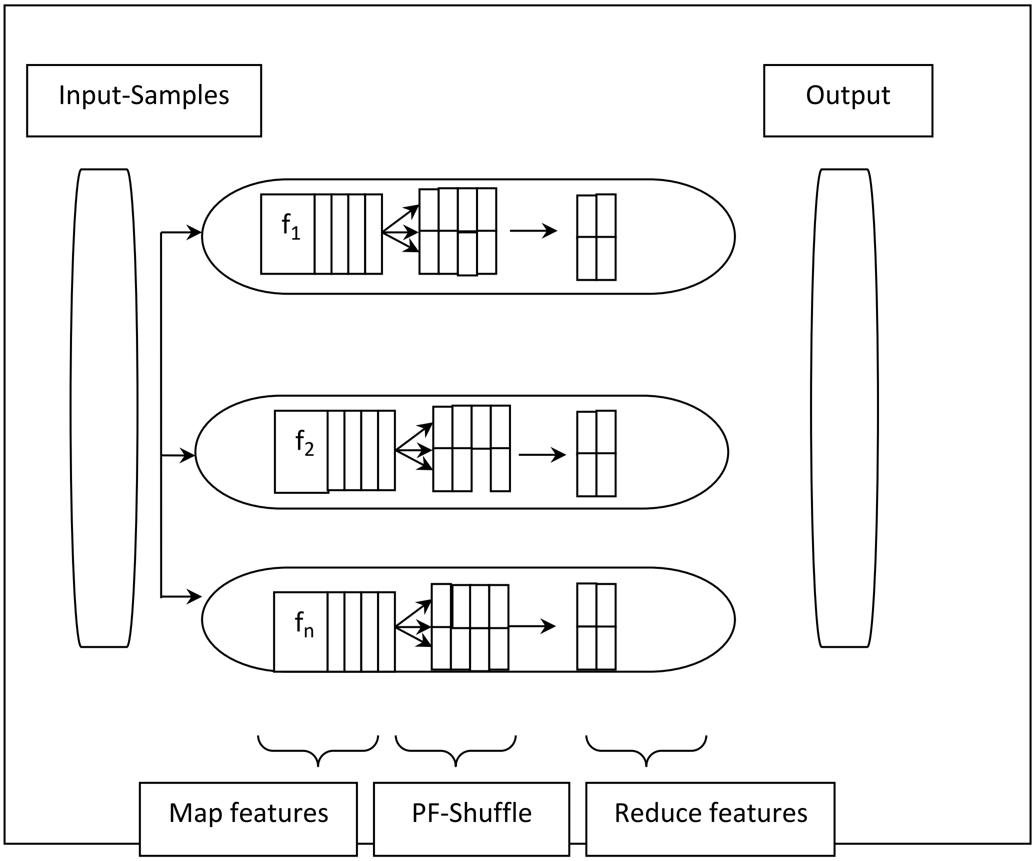

Figure 3 Structure of Inverse Polynomial Map Reduce

Inverse Polynomial Pre-processing model is applied to remove the noise. Figure 3.3. shows the structure of Inverse Polynomial Map Reduce model. As illustrated in the figure, let us consider samples '$S = s_1, s_2, \dots, s_m$' given as input with features i.e., data points represented by '$F = f_1, f_2, \dots, f_n$' in the map phase. Here 'm' represents the number of single subject or persons in the dataset and 'n' represents the number of data points for each single subject or person. Then, the Data Point Matrix (DPM) corresponding to 'm' samples and 'n' features are mathematically formulated as given in the equation (3.1).

https://doi.org/10.37896/pd91.4/91443

$$DPM = \begin{bmatrix} s_1 f_1 & s_1 f_2 & \ldots.. & s_1 f_n \\ s_2 f_1 & s_2 f_2 & \ldots.. & s_2 f_n \\ \ldots.. & \ldots.. & \ldots.. & \ldots.. \\ s_m f_1 & s_m f_2 & \ldots.. & s_m f_n \end{bmatrix} \tag{3.1}$$

DPM a polynomial function is applied due to the reason that the brain activity changes at a fraction of second and monitoring those changes and pre-processing those changes remains the key in denoising the signals. This is performed in the shuffle section and is mathematically formulated as given in the equations (3.2) and (3.3).

$$PF_k(f,t) = f_0 + \sum_{i=1}^{n} f_i t_i \, [DPM\,] \tag{3.2}$$

$$W_j = PF_k(f,t)\sin(\omega t_i + \alpha) \tag{3.3}$$

The polynomial function 'PF_k' is first equated for each features 'f_i' i.e. data points recorded at different time intervals 't_i' is given in the equation (3.2). Then from equation (3.3), the wave for each sample is measured using frequency 'ω' and phase shift 'α' corresponding to 'k' degree polynomials. The pre-processed features or data points recorded in the reduced phase are mathematically formulated using equations (3.4) and (3.5).

$$DPM_i^T g = (DPM_i^T, DPM_i) f * W_j \tag{3.4}$$

$$g = \frac{(DPM_i^T, DPM_i) f}{DPM_i^T} = (DPM_i) f * W_j \tag{3.5}$$

'g' returns the pre-processed features from equations (3.4) and (3.5) i.e., pre-processed data points in the reduced phase and is recorded as output. The pseudo code representation of Inverse Polynomial Map Reduce Pre-processor is given in the Figure 3.4.

```
Input: samples '$S = s_1, s_2, \ldots, s_m$', features '$F = f_1, f_2, \ldots, f_n$'
Output: Computationally efficient pre-processed features '$g$'
Initialize '$m$', '$n$' frequency '$\omega$', phase shift '$\alpha$'
Begin
For each sample '$S$' with features '$F$'
Obtain the data point matrix as given in equation (3.1)
Evaluate polynomial function using equation (3.2)
Measure wave corresponding to '$k$' degree polynomials using equation (3.3)
Evaluate pre-processed features using equation (3.5)
Return (pre-processed features)
End for
End
```

Algorithm of Inverse Polynomial Map Reduce Preprocessing

Inverse Polynomial Map Reduce Pre-processor is used to reduce the signal to noise ratio by means of Inverse Polynomial function. Since, big data is involved for analysis, a parallel processing using MapReduce is used and the Inverse Polynomial function is applied to pre-process computationally efficient features by minimizing the signal to noise ratio. It is achieved by deriving the polynomial function for each sample set and inversing the data point matrix. Feature or data point extraction with optimal class variance i.e., high inter-class variance and low intra-class variance are performed using the resultant pre-processed features.

Tikhonov Entropy-based Feature Extraction

After preprocessing of raw brain activity data, features or data points are extracted for early prediction. Features are said to be extracted in two ways, one is by extracting hand-crafted features and the other is automated feature extraction. In proposed work, automated features extraction or automated data point extraction using Tikhonov Entropy-based Feature extraction model is performed for extracting features or data points with high inter-class variance. Figure 4 shows the flow diagram of Tikhonov Entropy-based Feature extraction model.

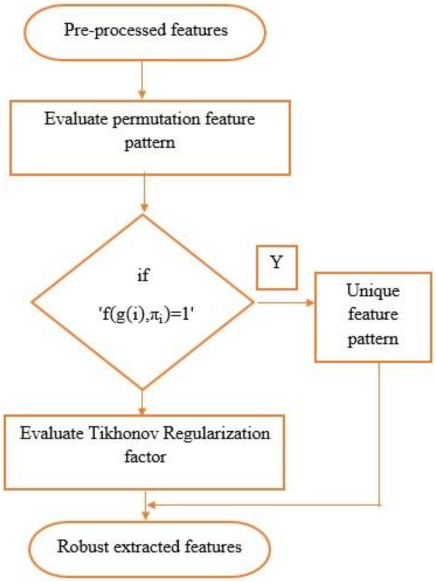
Pre-processed features
Evaluate permutation feature pattern
if
'f(g(i),π_i)=1'
Y
Unique feature pattern
Evaluate Tikhonov Regularization factor
Robust extracted features

Figure 4 Flow diagram of Tikhonov Entropy-based Feature Extraction

As illustrated in the flow diagram, with the pre-processed vector ‘$g(i)$’ and a permutation pattern ‘π_i’ with ‘$i = 1,2, \ldots n!$’, the permutation probability for all ‘$i = 1,2, \ldots n!$’ is defined as the probability that a pre-processed vector has the same feature pattern or data points as the permutation feature pattern. This is mathematically formulated as given in the equation (3.6).

$$Prob\ (\pi_i) = \sum f(g(i), \pi_i) \tag{3.6}$$

where '$f(g(i), \pi_i) = 1$', when '$g(i), \pi_i$' possess the same feature pattern and zero otherwise. In case of zero, Tikhonov regularization is applied with the objective of obtaining only few feature patterns. By applying both permutation-based entropy and regularization, optimal class variance is ensured. Let us consider a pre-processed feature matrix '$(DPM_i)f$' and vector 'g' in the equation (3.7).

$$(DPM_i)f * W_j = g \tag{3.7}$$

The Tikhonov regularization is represented in the equation (3.8).

$$\left[(DPM_i)f * W_j - g\right]_P^2 + [q - q_0]_Q^2 \tag{3.8}$$

where, 'P' refers to the inverse covariance matrix of 'g', 'q_0' is the expected value of 'q' and 'Q' referring to the inverse covariance matrix of 'q'. Then, with the aid of regularization, an optimal solution ReliefF '(RF)' is obtained using the equation (3.9).

$$RF = \left[(DPM_i)_f^T P + Q\right]^{-1} \tag{3.9}$$

where, 'RF' returns the extracted features i.e. extracted data points and is recorded as output. The pseudo code representation of Tikhonov Entropy-based Feature extraction is given in the Figure 3.6.

```
Input: Pre-processed features '$g = g_1, g_2, \dots, g_n$'
Output: Robust feature extraction '$RF = rf_1, rf_2, \dots, rf_n$'
Begin
    For each Pre-processed feature '$g$'
        Evaluate permutation feature pattern using equation (3.6)
        If '$f(g(i), \pi_i) = 1$'
            Similar feature pattern
        End if
        If '$f(g(i), \pi_i) = 0$'
            Evaluate Tikhonov regularization using equation (3.8) and equation (3.9)
        End if
        Return (robust features)
    End for
End
```

Algorithm of Tikhonov Entropy-based Feature Extraction

Kullback–Leibler Vuong Logistic Regressive Classifier for Disease diagnosis

Once the features or data points have been extracted from pre-processed features, the final step is to perform classification between seizure recognition or not. Kullback–Leibler Vuong Logistic Regressive Classifier is applied to the extracted features for early disease diagnosis. Let us consider two predictors 'p_1' and 'p_2' i.e., response variable is 'y' in column 179, '$p_1 = 1$' and '$p_2 = 4$' and one Bernoulli response variable, '$Y = 1$' ($r = Prob(Y = 1)$). Equations (3.10) and (3.11) represents the logit function.

$$l = log_b \left[\frac{r}{1-r}\right] \tag{3.10}$$

$$l = \alpha_0 + \alpha_1 p_1 + \alpha_2 p_2 \tag{3.11}$$

where, 'l' refers to the logit function with 'b' denoting the base of the logarithm and 'α' referring to the robust features 'RF' respectively. The likelihood ratio is evaluated by means of Kullback–Leibler Function (KLF) in the equation (3.12).

$$KLF = \frac{LR_N(\alpha_{p_1}, \alpha_{p_2})}{\sqrt{N}\omega_N} \tag{3.12}$$

The Kullback–Leibler function 'KLF' is evaluated via Likelihood Ratio '(LR)' with respect to two predictors 'α_{p_1}' and 'α_{p_2}' to the sample variance 'ω_N' respectively. Finally, the Log Likelihood Ratio '$(\mathrm{LLR_i})$' for accurate classification forming for early diagnosis is mathematically formulated as given in the equation (3.13).

$$\mathrm{LLR_i} = \log \frac{f_1(y_i | \, RF_i, \alpha_{p_1})}{f_2(y_i | \, RF_i, \alpha_{p_2})} \tag{3.13}$$

Early epileptic disease diagnosis is accomplished using the aforementioned Log Likelihood Ratio and accurate classification. The pseudo code representation of Kullback–Leibler Vuong and Logistic Classifier is illustrated in the Figure 3.7.

```
Input: feature extracted '$RF = rf_1, rf_2, \dots, rf_n$'
Output: Early disease prediction
Initialize '$p_1$' and '$p_2$'
Begin
For each feature extracted '$RF$'
Evaluate logit function using equations (3.10) and (3.11)
Evaluate Kullback–Leibler function using equation (3.12)
Measure log likelihood ratio using equation (3.13)
Return classified results
End for
End
```

Algorithm of Kullback–Leibler Vuong Logistic Classifier

As stated in the Kullback–Leibler Vuong Logistic Classifier algorithm, for each extracted feature, three functions are applied for early epileptic disease diagnosis. First, a logit function that differentiates between the predictors precisely. Second, with the application of Kullback–Leibler function likelihood ratio is evaluated for the overall sample variance. Finally, accurate classification is done via log likelihood ratio, paving way for early disease diagnosis.

RESULTS AND DISCUSSION

In the experiments, the Kullback–Leibler Vuong and Logistic Classifier algorithm is implemented by java program language and Map Reduce parallel programming model via Cloudsim Simulating environment. The version of 1.1.2 is adopted for Hadoop cluster. In the clusters, one node acts as the master and the others act as slaves having the hardware configuration, namely Core2 Duo CPU @ 2.20GHz, 2 CPUs and 8GB of RAM. The nodes are connected by the network with the bandwidth of 100M/s. SUN JAVA JDK1.6.0_24 trained on: https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition.

Performance Analysis of Sensitivity

The first metric to analyze the diagnosis of disease is sensitivity. The sensitivity evaluates the ratio of actual positives that are correctly identified as such e.g., the percentage of diseased samples identified as having the condition.

$$Sensitivity = \frac{TP}{TP+FN} \tag{3.14}$$

From the above equation (3.14), sensitivity '$Sensitivity$', is measured as the ratio of True Positive '(TP)' to the summation of True Positive and False Negative (FN), '$TP + FN$' respectively. Here, True Positive refers to the not diseased samples correctly identified as not diseased and False Negative refers to the not diseased samples incorrectly identified as diseased samples. The results of sensitivity are reported in Table 3.2.

Table 1 Tabulation of Sensitivity

Number of Samples	Sensitivity (%)		
	DNN	CE-stSENet	PTE-KVLC
1100	95.06	96.47	96.87
2200	92.30	94.59	94.91
3300	91.30	94.23	95.20
4400	91.07	94.04	95.67
5500	89.47	92.68	94.23
6600	89.02	93.0	93.89
7700	87.71	91.59	92.31
8800	86.88	90.47	91.60
9900	85.82	89.85	90.48
11000	85.71	89.74	91.03

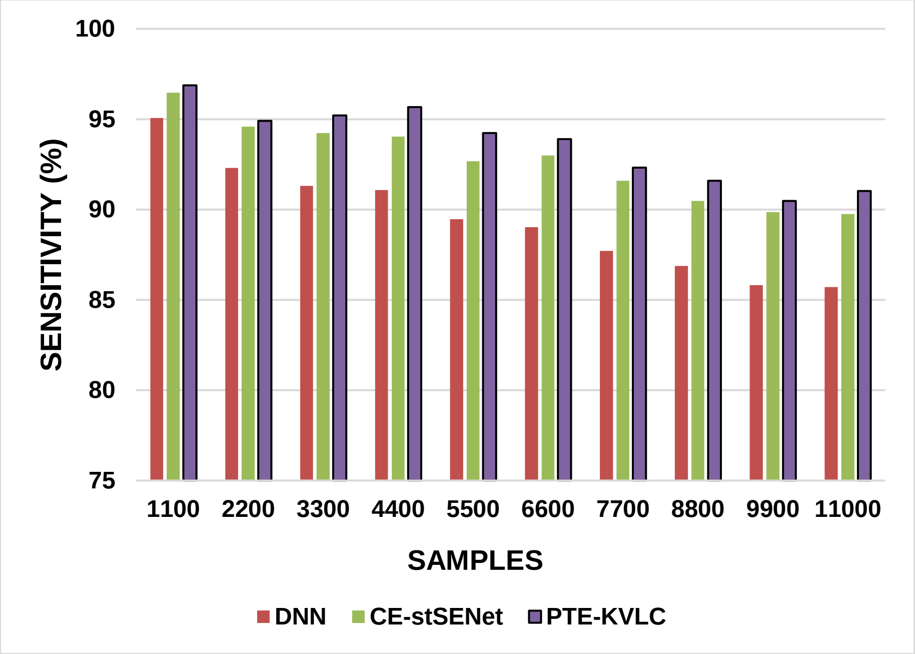

Figure 5 Graphical representation of Sensitivity

Table 1 and Figure 5 depicts the rate of sensitivity with respect to 11000 samples provided as input using Epileptic Seizure Recognition dataset. Here 11000 samples first refer to the five different folders each with 100 files and 4097 data points in 23 chunks, 23 * 500 = 11500 samples. For simulation purpose, 11000 samples have been considered and the sensitivity is measured accordingly. From the Figure 3.8, it is illustrative that comparison made with the proposed method PTE-KVLC shows better results with 96.87% non-diseased samples correctly identified as non-disease out of 100, and 96.47%, 95.06% using CE-stSENet and DNN respectively. From these results improvement is observed using PTE-KVLC than when compared to CE-stSENet and DNN. This is because of the application of Inverse Polynomial Map Reduce Pre-processor algorithm. By applying this algorithm, not only the signal to noise ratio is minimized via Inverse Polynomial function, but also computationally efficient features are obtained via optimal class variance. With these two features, sensitivity is improved using PTE-KVLC by 1.03% compared to CE-stSENet and 4.67% compared to DNN.

Performance Analysis of Specificity

The second metric to diagnose disease is specificity. Specificity measures the ratio of actual negatives that are correctly identified as such e.g., the percentage of diseased samples that are correctly identified as not having the condition.

$$Specificity = \frac{TN}{TN+FP} \tag{3.15}$$

https://doi.org/10.37896/pd91.4/91443

From the above equation (3.15), specificity '$Specificity$' is measured based on the True Negative 'TN' and the summation of True Negative 'TN' and False Positive 'FP' respectively. It is measured in terms of percentage (%). Here, True Negative refers to the normal samples correctly identified as normal and False Positive refers to the normal samples incorrectly identified as diseased samples. The results of specificity are reported in Table 3.7.

Table2 Tabulation of Specificity

Number of Samples	Specificity (%)		
	DNN	CE-stSENet	PTE-KVLC
1100	79.31	80.0	85.39
2200	77.77	83.33	83.97
3300	70.00	71.42	73.06
4400	71.87	75.20	78.57
5500	70.58	71.42	72.13
6600	68.00	68.75	69.18
7700	65.00	65.71	65.86
8800	66.66	68	69.07
9900	68.75	70	70.68
11000	67.64	68.75	69.54

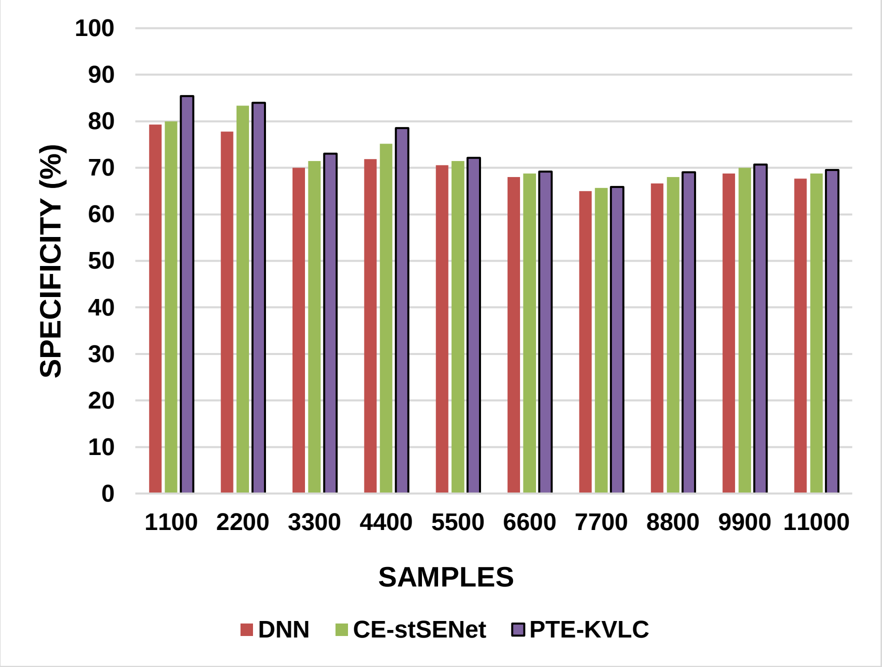

Figure 6 Graphical representation of Specificity

Table 2 and Figure 6 shows the specificity rate with respect to 11000 different samples collected at different time periods, with each file recording brain activity for 23.6 seconds using

Epileptic Seizure Recognition dataset. Figure 3.9 illustrates that the specificity rate is inversely proportionate to the samples considered for simulation. Despite, improvement is found using PTE-KVLC when compared to CE-stSENet and DNN via simulations. With 1100 samples considered for simulation, 880 normal samples were correctly identified as normal using PTE-KVLC out of 900 normal samples and 870, 860 using CE-stSENet and DNN. From the simulation results it is inferred that specificity is better using PTE-KVLC than CE-stSENet and DNN. This is because of the application of Tikhonov Entropy-based Feature extraction algorithm. The robust features are said to be extracted by applying permutation-based entropy to the pre-processed signals i.e. data points and the entropy results regularization of the features or data points is made via Tikhonov function. With these two features the specificity is said to be better using PTE-KVLC method by 2.06% compared to CE-stSENet and 4.52% compared to DNN respectively.

Performance Analysis of Accuracy

Accuracy is mathematically formulated as given in the equation (3.16).

$$Acc = \frac{TN+TP}{TN+TP+FN+FP} \tag{3.16}$$

where, accuracy 'Acc' is measured based on the ratio of the summation of True Negative, True Positive '$TN + TP$' and the summation of True Negative, True Positive, False Negative and False Positive '$TN + TP + FN + FP$' respectively. It is measured in terms of percentage (%).

Table 3 Tabulation for Accuracy

Number of Samples	Accuracy (%)		
	DNN	CE-stSENet	PTE-KVLC
1100	90.90	92.72	95.18
2200	86.36	90.90	93.67
3300	84.84	89.39	94.01
4400	84.09	88.86	93.15
5500	83.63	87.27	92.76
6600	81.06	87.12	90.81
7700	81.81	85.71	88.29
8800	80.68	84.09	85.94
9900	80.30	83.83	85.39
11000	80.90	83.63	83.87

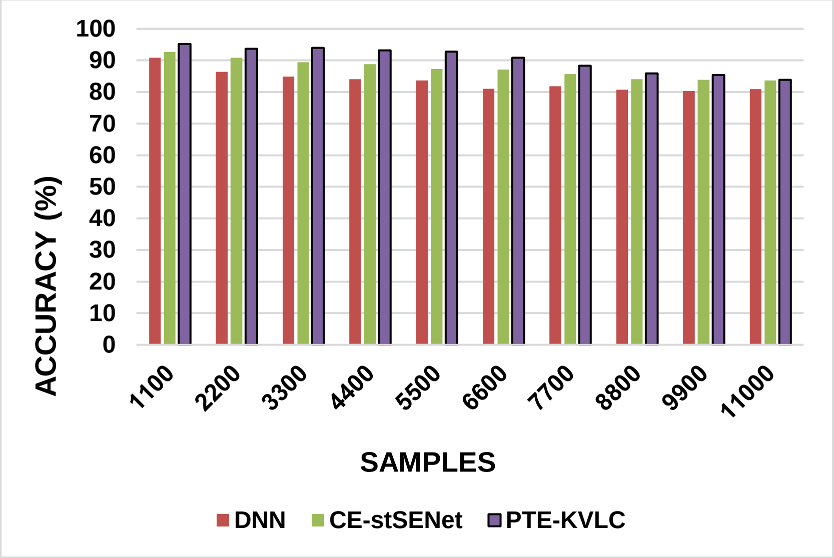

Figure 7 Graphical representation of Accuracy

Table 3 and Figure 7 represent the accuracy. Accuracy is measured based on the True Positive i.e., non-diseased samples correctly identified as non-diseased, True Negative i.e., diseased samples correctly identified as diseased, i.e., False Positive diseased samples incorrectly identified as non-diseased and False Negative i.e., non-diseased samples incorrectly identified as diseased samples. The accuracy results from the graphical representation provide the better results using PTE-KVLC than compared to CE-stSENet and DNN.

The simulations performed for 11000 samples shows better accuracy results using PTE-KVLC with the incorporation of Kullback–Leibler Vuong Logistic Classifier algorithm. By applying the algorithm, accurate and precise classification results are said to be arrived by first applying logit function. With this function, differentiation between the predictors is made in a significant manner. Also, via Kullback–Leibler function with log likelihood ratio, disease diagnosis is made effectively. With this the accuracy rate is improved using PTE-KVLC by 3.39% compared to CE-stSENet and 8.2% compared to DNN respectively.

Comparison of Time Complexity

It is defined as the amount of time taken by the algorithm to identify the disease based on the classification. The overall Time Complexity is measured using the following equation (3.17)

$$TC = n * [time(COS)] \tag{3.17}$$

Where TC denotes a time complexity, n represents the number of samples, $time(COS)$ denotes a time for classifying one sample. The overall time complexity is measured in the unit of milliseconds (ms).

https://doi.org/10.37896/pd91.4/91443

Table 4 Evaluation of Time Complexity

Number of Samples	Time Complexity (ms)		
	DNN	CE-stSENet	PTE-KVLC
1100	45	42	39
2200	55	51	47
3300	69	59	57
4400	75	67	66
5500	80	74	71
6600	89	83	79
7700	95	89	86
8800	101	96	95
9900	110	105	102
11000	120	113	108

Table 4 describes the performance analysis of the time complexity of disease diagnosis versus the number of input samples taken in the counts from 1100 to 11000. While increasing the input samples, the time complexity gets increased for all the classification methods. But comparatively, the proposed PTE-KVLC technique achieves lesser time complexity than the others. Let us consider the 1100 samples, the time taken to classify the input sample is $39ms$ by using the PTE-KVLC technique. The proposed PTE-KVLC technique is compared to the time consumption of the existing results. The average time complexity of the PTE-KVLC technique is considerably minimized by $5ms$ and $12ms$ when compared to CE-stSENet and DNN respectively.

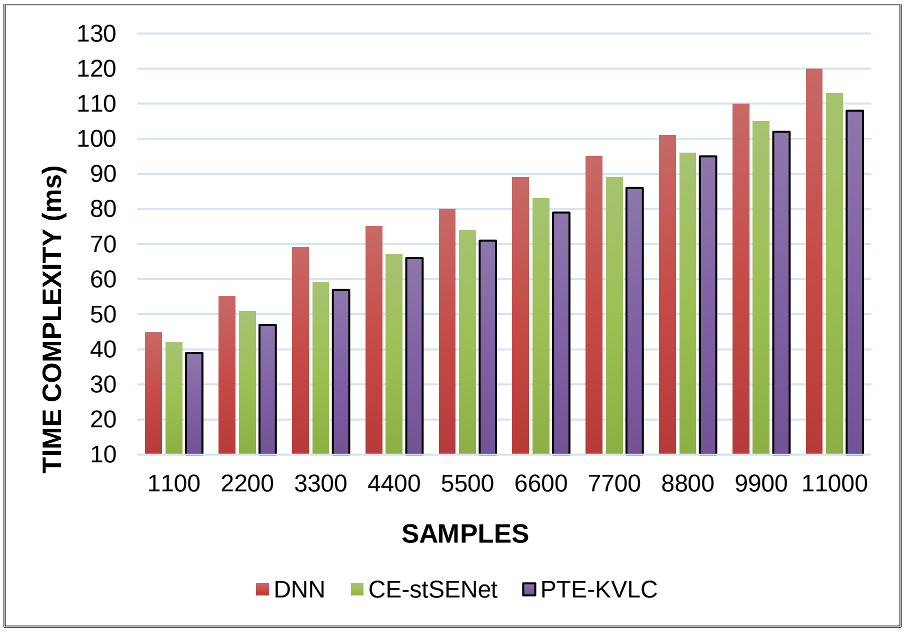

Figure 8 Graphical representation of Time Complexity

Figure 8 exhibits the performance results of the time complexity with respect to a number of samples taken from the Epileptic Seizure Recognition dataset. The graphical result visibly illustrates that the PTE-KVLC technique reduces the prediction time than the existing disease diagnosis methods. Permutation based entropy and Tikhonov regularization function extracts the robust features to identify the Epileptic Seizure Recognition Data Set. Generally, the big dataset consists of a large number of features and it leads to more complexity of disease identification. With the smaller number of selected features, the PTE-KVLC technique consumes minimum time to perform the disease prediction.

CONCLUSION

A Polynomial Tikhonov Entropy and Kullback-Leibler Vuong Logistic Classifier is proposed for epileptic seizure disease diagnosis. First, an Inverse Polynomial Pre-processing model is used to denoise the data points and increase class variance via Map Reduce parallel processing to improve epilepsy disease diagnosis. Second, using a Tikhonov Entropy-based Feature extraction, a significant feature selection model is created. Permutation-based entropy and Tikhonov regularization is utilized to extract robust features with optimal class variance. Besides, Kullback–Leibler Vuong Logistic Regressive classification technique is employed to classify epileptic seizure classes, and the Kullback–Leibler function effectively increases mapping performance when compared to manually predicted features. In the experimental study, the proposed work delivered a comparatively better performance by means of Sensitivity, Specificity, Accuracy and Time Complexity compared to the existing works.

REFERENCES

[1] Aaron N. Richter, Taghi M. Khoshgoftaara, “A review of statistical and machine learning methods for modeling cancer risk using structured clinical data”, Artificial Intelligence In Medicine, Elsevier, 90:1-14, 2018.

[2] Ahmed H. Osman, Ahmad A Alzahrani , “New Approach for Automated Epileptic Disease Diagnosis using an Integrated Self-Organization Map and Radial Basis Function Neural Network Algorithm “, IEEE Access, 7:4741-4747,2018.

[3] Alex Frid, Meirav Shor, Alla Shifrin, David Yarnitsky & Yelena Granovsky, “A Biomarker for Discriminating Between Migraine with and Without Aura: Machine Learning on Functional Connectivity on Resting-State EEGs”, Annals of Biomedical Engineering, Springer, 48:403-412,2020.

[4] Anurag Nishad, Ram Bilas Pachori, “Classification of epileptic electroencephalogram signals using tunable-Q wavelet transform based filter-bank”, Journal of Ambient Intelligence and Humanized Computing, Springer, 2020.

[5] Christopher C. Fesmire, Ross A. Petrella, Callie A. Fogle, David A. Gerber, Lei Xing & Michael B. Sano, “Temperature Dependence of High Frequency Irreversible Electroporation Evaluated in a 3D Tumor Model”, Annals of Biomedical Engineering, springer, 48: 2233–2246, 2020.

[6] Chunxue Wu, Chong Luo, Naixue Xiong, Wei Zhang, Tai-Hoon Kim, “A Greedy Deep Learning Method for Medical Disease Analysis”, IEEE Access, 6:20021-20030, 2018.

[7] Hisham Daoud, Member, IEEE, Magdy Bayoumi , “Efficient Epileptic Seizure Prediction based on Deep Learning”, IEEE Trans. on Biomedical Circuits and Systems, 13(5): 804 – 813, 2019.

[8] Isabell Kiral-Kornek, Subhrajit Roy, Ewan Nurse, Benjamin Mashford, Philippa Karoly, Thomas Carroll, Daniel Payne, Susmita Saha, Steven Baldassano, Terence O'Brien, David Grayden, Mark Cook, Dean Freestone, Stefan Harrer, “Epileptic Seizure Prediction Using Big Data and Deep Learning: Toward a Mobile System”, EBioMedicine, Elsevier, 127:103-111,2018.

[9] Kaliappan. A & Chitra. D,” Kringing Regressive Map reduce Entropy Feature Extraction based Rocchio Adaptive Boost Ensemble Classifier for Early Disease Diagnosis with Big Data”, Journal of Dynamic Systems and Applications, Dynamic Publishers Inc., US, vol.30, no. 6, pp. 964-980.

[10] Kaliappan. A & Chitra. D,” Analysis of Big Data Analytics in Healthcare sector: Applications and Tools”, Journal of Computational and Theoretical Nanoscience, American Scientific Publishers, US, vol.17, no.12, pp. 5605-5612(8).

[11] Kaliappan. A, Dineshkumar. T, Sabarivel. M, "Hierarchical Kriging DNN Model for Epileptic Seizure Detection from EEG Signals”, TNSCST Sponsored 8th International Conference on Latest Trends in Science, Engineering and Technology (ICLTSET'22) organized by Karpagam Institute of Technology, Coimbatore.

[12] Kaliappan. A, “Proficient Valuation and Estimation of Human Performance in Collective Learning Environments using Machine Learning”, South Asian Journal of Engineering and Technology, Vol 8, Supplementary Issue 2, April 2019, ISSN 2320 – 3765.

[13] Kaliappan. A,”Energy Proficient Forecasting of Map Reduce for Big Data Real Time Appliances”, International Journal of Modern Trends in Engineering and Science Vol.4 Issue 03,2017, ISSN 2348-3121.

[14] Kaliappan A,” Improved Bees Hybrid Sensor Coverage Algorithm”, International Journal of Computer Science and Information Technologies, Vol.6 (6),2015, ISSN 0975-9646.

[15] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, Dimitrios I. Fotiadis, “Machine learning applications in cancer prognosis and predict”, Computational and Structural Biotechnology Journal, Elsevier, 13:8-17,2015.

[16] Kuldeep Singh, Jyoteesh Malhotra, “IoT and cloud computing based automatic epileptic seizure detection using HOS features based random forest classification”, Journal of Ambient Intelligence and Humanized Computing, Springer, 2019.

[17] Maisa Daoud, Michael Mayo, “A survey of neural network-based cancer prediction models from microarray Data”, Artificial Intelligence in Medicine, Elsevier, 97:204-214, 2019.

[18] Mohammad Khubeb Siddiqui, Ruben Morales-Menendez, Xiaodi Huang, Nasir Hussain, “A review of epileptic seizure detection using machine learning classifiers”, Brain Informatics, Springer, 2(5):1-7, 2020.

[19] Ozlem Karabiber Cura, Sibel Kocaaslan Atli, Hatice Sabiha Türe and Aydin Akan, “Epileptic seizure classifications using empirical mode decomposition and its derivative”, Bio Medical Engineering, 19:1-22, 2020.

[20] Raghu.S, Natarajan Sriraam, Shyam Vasudeva Rao, Alangar Sathyaranjan Hegde, Pieter L. Kubben, “Automated detection of epileptic seizures using successive decomposition index and support vector machine classifier in long-term EEG”, Neural Computing and Applications, Springer, 32: 8965-8984, 2020.

[21] Rubén San-Segundo, Manuel Gil-Martína, Luis Fernando D'Haro-Enríqueza, José Manuel Pardoa, “Classification of epileptic EEG recordings using signal transforms and convolutional neural networks”, Computers in Biology and Medicine, Elsevier, 109:148-158, 2019.

[22] Shadi Aljawarneh, Aurea Anguera, John William Atwood, Juan A. Lara and David Lizcano4, “Particularities of data mining in medicine: lessons learned from patient medical time series data analysis”, Eur.J. Wireless Communications and Networking, Springer Open, 1-29, 2019.

[23] Sriraam.N, Raghu.S, Kadeeja Tamanna, Leena Narayan, Mehraj Khanum, A. S. Hegde and Anjani Bhushan Kumar, “Automated epileptic seizures detection using multi-features and multilayer perceptron neural network”, Brain Informatics, Springer , 5:1-10, 2018.

[24] Thara D.K., PremaSudha B.G, Fan Xiong, “Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques”, Pattern Recognition Letters, Elsevier, 128: 544-550, 2019.

[25] William C. Stacey, “Seizure Prediction Is Possible–Now Let's Make It Practical”, EBioMedicine, Elsevier, 2018.

[26] Xiaoyan Wei, Lin Zhou, Ziyi Chen, Liangjun Zhang and Yi Zhou, “Automatic seizure detection using threedimensional CNN based on multi-channel EEG”, BMC Medical Informatics and Decision Making, 18:72-127, 2018.

[27] Yang Si, “Machine learning applications for electroencephalograph signals in epilepsy: a quick review”, BMC, Oct 2020.

[28] Yash Paul, “Various epileptic seizure detection techniques using biomedical signals: a review”, Brain Informatics, Springer, 5:1-9, 2018.