Cognitive Neurodynamics
https://doi.org/10.1007/s11571-023-09929-z

RESEARCH ARTICLE

Exploring recurrence quantification analysis and fractal dimension algorithms for diagnosis of encephalopathy

Sreejith Chandrasekharan[1] · Jisu Elsa Jacob[2] · Ajith Cherian[3] · Thomas Iype[4]

Received: 20 October 2022 / Revised: 11 December 2022 / Accepted: 9 January 2023

Abstract

Electroencephalography (EEG) is a crucial non-invasive medical tool for diagnosing neurological disorder called encephalopathy. There is a requirement for powerful signal processing algorithms as EEG patterns in encephalopathies are not specific to a particular etiology. As visual examination and linear methods of EEG analysis are not sufficient to get the subtle information regarding various neuro pathologies, non-linear analysis methods can be employed for exploring the dynamic, complex and chaotic nature of EEG signals. This work aims identifying and differentiating the patterns specific to cerebral dysfunctions associated with Encephalopathy using Recurrence Quantification Analysis and Fractal Dimension algorithms. This study analysed six RQA features, namely, recurrence rate, determinism, laminarity, diagonal length, diagonal entropy and trapping time and comparing them with fractal dimensions, namely, Higuchi's and Katz's fractal dimension. Fractal dimensions were found to be lower for encephalopathy cases showing decreased complexity when compared to that of normal healthy subjects. On the other hand, RQA features were found to be higher for encephalopathy cases indicating higher recurrence and more periodic patterns in EEGs of encephalopathy compared to that of normal healthy controls. The feature reduction was then performed using Principal Component Analysis and fed to three promising classifiers: SVM, Random Forest and Multi-layer Perceptron. The resultant system provides a practically realizable pipeline for the diagnosis of encephalopathy.

Keywords Electroencephalogram (EEG) · Encephalopathy · Recurrence quantification analysis · Higuchi's fractal dimension · Katz fractal dimension · Support vector machine · Random forest · Multilayer perceptron

Introduction

EEG and its chaotic character

The chaotic nature of brain dynamics and of electroencephalogram (EEG) signals is a crucial area of research. EEG signal that records the electrical activity of millions of neurons in the brain (Schomer and Silva 2012), is utilised for the diagnosis of major neurological diseases. These signals clearly portray the dynamics of brain and gives evidence on various neuropathology. It can be used more effectively by employing powerful signal processing techniques, reducing expert interventions through their visual examination alone. Various chaotic and non-linear analysis techniques have been reported for EEG signal assessment and disease diagnosis. This study employs recurrent quantification analysis (RQA) features of EEG signals as well as fractal dimensions and compares their values between the encephalopathy disease group and of normal healthy subjects. RQA has been identified as a novel method for investigating the complex systems not only in biomedical fields, but also in ecology, earth science and finance sector (Marwan 2011). Recurrence plots provide beautiful fancy pictures which give clear information about the level of similarities and hence, complexity of the

✉ Jisu Elsa Jacob
jisuelsa@sctce.ac.in

[1] Thiruvananthapuram, Kerala, India

[2] Department of Electronics and Communication Engineering, Sree Chitra Thirunal College of Engineering, Thiruvananthapuram 695018, Kerala, India

[3] Department of Neurology, SCTIMST, Thiruvananthapuram, Kerala, India

[4] Department of Neurology, Government Medical College, Thiruvananthapuram, Kerala, India

Published online: 03 February 2023

system. Fractal analysis also represents clearly the fractal sets, both of which clearly depicts the complexity of the system analysed.

Encephalopathy and its EEG features

The chaotic and non-linear analysis of EEG signals is performed for identifying the cases of a neuropathological condition called encephalopathy. Normal brain functioning is dependent on the normal neuronal metabolism and is closely dependent on the metabolic balances like levels of glucose, electrolytes, amino acids etc. Thus, when metabolic imbalances occur, diffused brain dysfunction occurs, which is called metabolic encephalopathy. Neuron activities are greatly influenced by metabolic homeostasis and any variation in the metabolic system can lead to brain disorders. The metabolic encephalopathy is the consequence of various systemic disturbances causing diffuse brain dysfunction. Various types of Encephalopathy include hepatic encephalopathy, which can occur due to liver failure (Ferenci et al. 2002; Musgrave and Hilsabeck 2019), hypoglycemic encephalopathy, due to low sugar level in the blood (Blaabjerg and Juhl 2016), hypocalcemia and hypercalcemia, due to variations in calcium level, uremic encephalopathy that occurs due to acute or chronic renal failure and so on. Thus, cerebral activity is affected in metabolic encephalopathy in the absence of gross structural abnormalities of the brain. Unlike epilepsy or Alzheimer's disease, metabolic encephalopathy can be treated as a secondary neurological disorder as brain is affected due to the metabolic imbalances and other organ malfunctioning. Figure 1a and b shows a sample EEG recording of a patient with encephalopathy and of normal healthy subject respectively.

The major features noted in EEG of patients with encephalopathy are generalised slowing of EEG, presence of some periodic patterns such as burst-suppression or background suppression and electrocerebral inactivity, in the presence of triphasic waves. The presence of 'triphasic waves' was reported by Bickford and Butt in EEGs of patients with hepatic encephalopathy (Bickford and Butt 1955; Angel and Young 2011). As the name suggests, the pattern has three phases, mainly a downward deflection both preceded and succeeded by a smaller positive and upward deflection. In metabolic encephalopathy, changes in EEG can very well correlate with the severity of the disease encephalopathy, though EEG lacks specificity in differentiating between various types and states of metabolic encephalopathy (Faigle et al. 2013).

Materials and methods

Data collection

The study was performed on 300 encephalopathy EEGs and 300 EEG epochs of normal healthy individuals. The entire EEG data collection was conducted in EEG lab of Neurology Department, Government Medical College, Thiruvananthapuram, Kerala. All these EEG signals are recorded in identical manner with Nicolet NicVue-v3.0 software using International 10–20 electrode system. The 300 EEG epochs of encephalopathy included the EEG epochs of 30 patients out of which 17 are hepatic and 13 are patients with uremic encephalopathy (Refer Table 1). The type of encephalopathy cases was not fixed prior and was randomly recruited for the study. Table 2 gives the demographic data of the participants of the study.

The steady state paradigm was used for EEG recording. Epochs were collected from common average montages which is illustrated in Fig. 2 which utilises averaged potential of all the electrodes as the reference electrode. EEG recording from fp1 electrode was taken for this study. EEG recording was conducted in resting eyes closed state.

Methods

This study explored all the recurrent quantification analysis (RQA) features and fractal dimensions (FD) of EEG during encephalopathy and studied their variation with that of normal healthy subjects. Thus, we explore the possibility to develop a complete framework for automated diagnosis of encephalopathy based on these features. The flow chart of this study is given in Fig. 3.

First step of the study was to recruit patients with encephalopathy along with age and sex-matched normal healthy individuals for including comparison in the study. The EEG recorded was saved as 12 s epochs in ASCII format. Each epoch of EEG contains 6000 samples as the sampling rate was fixed at 500 Hz. The cut-off frequency for the initial low pass filtering is set to 40 Hz. EEG signals were pre-processed by means of filtering using the technique of low-pass filtering and total variation denoising proposed by Selesnick et al. (2014). Thus, clean EEG data, obtained after LPF-TVD filtering is used to extract the various chaotic and non-linear features.

The RQA features, namely, recurrence rate (RR), determinism (DET), laminarity (LAM), length (L), RQA Entropy (ENT) and Trapping Time (TT) were calculated to get a measure of the complexity of the signal. Fractal dimensions based on Higuchi's (HFD) and Katz's algorithm (KFD) were also calculated to get a measure of the complexity of EEG signals. After calculating these

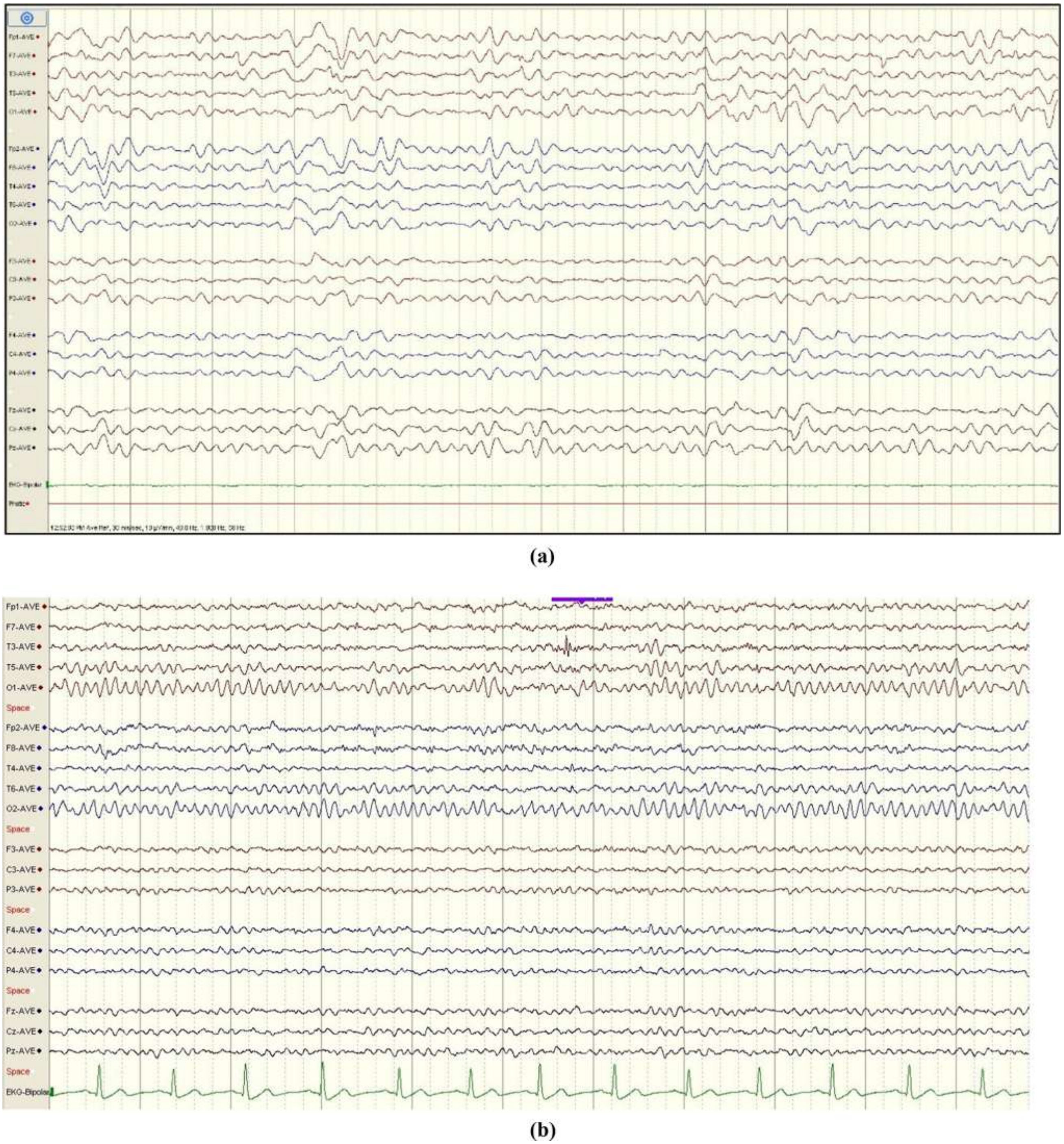

Fig. 1 (**a**) An EEG recording of a patient with encephalopathy. (**b**) An EEG recording of a normal healthy subject

Table 1 Various types of encephalopathy included in the database epochs for representing encephalopathy group

Type of encephalopathy	Number of patients	Number of epochs
Hepatic	17	162
Uremic	13	138

features, Principal Component Analysis (PCA) is applied to improve the discriminative power of the features projecting them to a derived n-dimensional space. These extracted features from PCA stage are fed to the classifiers for creating classifier models. In this study, three classifiers, namely, SVM, Random Forest and Multi-layer Perceptron (MLP) are employed. Performance analysis is performed for all classifiers.

Table 2 Demographic data of disease and normal case participants in the study

Group	No of Participants	No of epochs	Age (Mean; SD)	Gender (M/F)
Normal	30	300	(57.88; 11.2)	17/13
Encephalopathy	30	300	(50.13; 11.3)	16/14

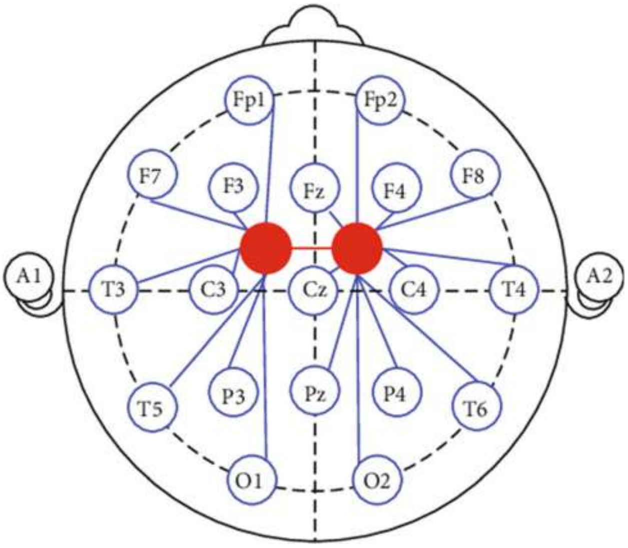

Fig. 2 Common average montage using averaged potential of all the electrodes as the referential electrode (Zhang et al. 2020)

Complexity measures of EEG

Physiological systems that are non-linear and non-stationary in nature requires strong signal processing algorithms. Even though many time-domain, frequency-domain and time–frequency domain analysis and feature extractions have been defined, there is so much focus on the research of chaotic and non-linear analysis in recent years. Chaotic or non-linear signals can be analysed in the best way using non-linear analysis techniques like a measure for the complexity, randomness or chaoticity of the signal. Theiler developed a method to test for the non-linearity nature called surrogate testing where a surrogate time series is developed from the original time series and test statistics are calculated for both series (Theiler et al. 1992). If they are different, data is proven to be non-linear in nature. If the results are same for both original and surrogate time series, it can be proven not to have non-linear property. The chaotic nature of the brain and of EEG signal was proved in many earlier studies (Pritchard and Duke 1995). One of the many complexity analysis methods are used here, namely, Recurrence Quantification Analysis (RQA) and Fractal Dimension (FD).

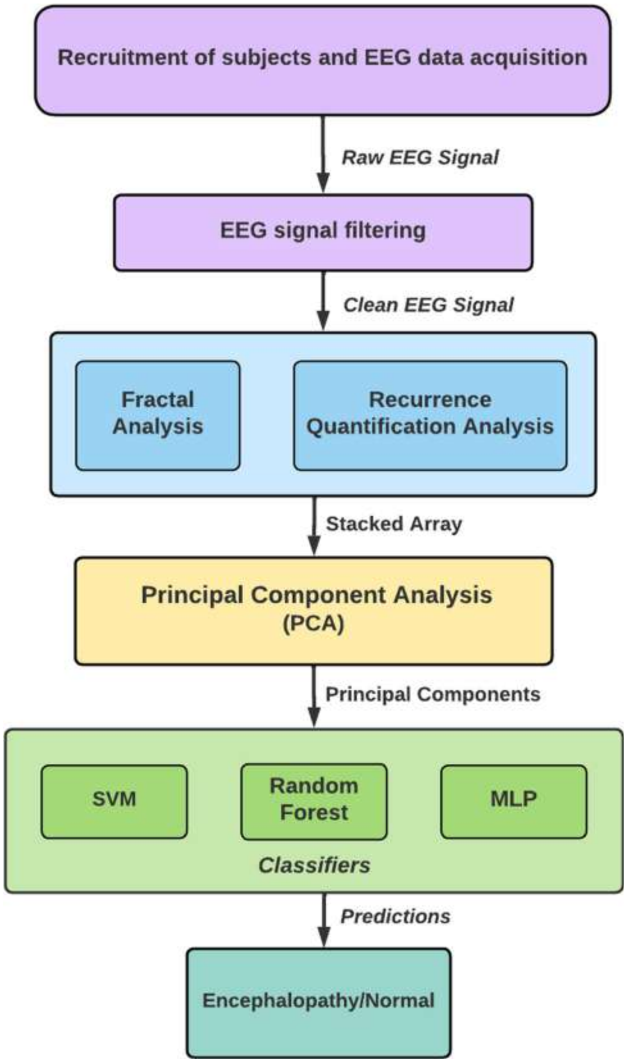

Fig. 3 Block diagram of the proposed system

Recurrence quantification analysis (RQA)

The deterministic dynamics in an EEG time series can be assessed using recurrence quantification analysis. When we consider a deterministic dynamic system, the recurrence of various states can happen more often. These pattern of recurrences of the trajectories in phase space leaves many valuable clues about the dynamic system which generates them. Recurrence states of a system can be visualised in recurrence plot, which was proposed by Eckmann et al. (1987). In 1987, Eckmann et al. introduced the concept of recurrence plots (RP) for visualising the pattern of recurrences in the dynamics of the system that is analysed. From RP, we can assess the time when particular states in the phase space recur where black dots represent recurrence (Ouyang et al. 2008).

From the one-dimensional EEG times series, RQA starts with the reconstruction of phase space (Packard et al. 1980) using delay embedding theorem (Takens 1981). It requires 2 parameters, namely, embedding dimension and time delay. Embedding dimension is chosen using the False Nearest Neighbour (FNN) method and time delay is chosen using mutual information method. Kennel et al. proposed FNN method in which attractor is constructed in m-dimensional phase space and then in m + 1 dimensional phase space (Kennel et al. 1992). True neighbour points are those points which are adjacent in both m and m + 1 phase space. Some points do not follow this condition and become far when dimension is increased from m to m + 1. They are called false neighbours. The number of false neighbours is computed for increasing value of *m*. The optimum value for embedding dimension *m* is the particular value of *m* when the number of false neighbours decrease drastically or become zero. Similarly, time delay is fixed using mutual information method (Fraser and Swinney 1986). Plot of mutual information *s* versus delay λisplotted, where *s* decreases reach a minimum and then again increases. Optimum delay is taken as the time delay when mutual information, *s* reaches its first minimum.

As recurrence plots are not easy for visual interpretation, a better way to analyse them is objective method by defining certain quantitative variables for measuring recurrences and their patterns (Webber and Marwan 2015). RPs consists of small-scale structures including dots which represent chance recurrences and diagonal, vertical and horizontal lines representing deterministic patterns which form the basis for quantitative RQA analysis. The RQA features calculated here are:

(i) *Recurrence Rate (RR)* It gives a measure or rate of recurrences that occur in a dynamic system. It is a measure of density of recurrence points in recurrence plots and it corresponds to the correlation sum. The recurrence rate RR of an RP calculates the probability of occurrence of similar states for a specific value of delay (Webber and Marwan 2015). Higher value of RR implies that the trajectories of the systems travel through same phase space regions, i.e. higher rate of recurrence. It is the probability of system recurrences given by:

$$\mathrm{RR} \; = \frac{1}{N^2} \sum_{\mathrm{i,j=1}}^{\mathrm{N}} \mathrm{R_{ij}}$$

(ii) *Determinism (DET)* It is measured as the rate of recurrences occurring in the dynamic system under analysis. It is calculated as the fraction of recurrence points in the diagonal lines in RP. Longer diagonal lengths in the RP shows a periodic nature of the system, that obey certain rules and hence have higher value of determinism. Shorter diagonals or dots shows lesser recurrences and indicate a stochastic system. Chaotic signals like EEG signals have shorted diagonal lengths. DET measures the predictability of the system, which gives a higher value for periodic behaviours and lower values for chaotic processes. DET is taken as determinism measure which is expected to be higher for disease case as more recurrence is seen in recurrence plot of encephalopathy case.

$$\mathrm{DET} \; = \frac{\sum_{\mathrm{l=}}^{\mathrm{N}} lP(l)}{\sum_{\mathrm{l=1}}^{\mathrm{N}} lP(l)}$$

Here, *l* gives the length of diagonal lines and *P(l)*, the histograms of the lengths.

(iii) *Laminarity (LAM)* It is similar to determinism and the difference is that, in laminarity, the percentage of recurrence points in vertical lines are computed (Marwan et al. 2002). It is given by:

$$\mathrm{LAM} \; = \frac{\sum_{\mathrm{v=v_{min}}}^{\mathrm{N}} vP(v)}{\sum_{\mathrm{v=1}}^{\mathrm{N}} vP(v)}$$

Here, *v* gives the length of vertical lines and *P(v)*, the histograms of the lengths of the vertical lines.

(iv) *Length (LEN)* It gives the average length of the diagonal lines in the RP. Higher the length, higher is the recurrence, indicating more periodic nature of the system. In chaotic systems, length will be lesser indicating lesser chance of recurrences of states in phase space (Webber and Marwan 2015). The length can be calculated as:

$$\mathrm{LEN} \; = \frac{\sum_{\mathrm{l=l_{min}}}^{\mathrm{N}} lP(l)}{\sum_{\mathrm{l=l_{min}}}^{\mathrm{N}} P(l)}$$

(v) *RQA Entropy (ENTR)* It is calculated as Shannon entropy for the probability distribution of diagonal lengths *p(l)*. It reflects the complexity of the system behaviour. It is given by:

$$\mathrm{ENTR} \; = -\sum_{\mathrm{i,j=1}}^{\mathrm{N}} p(l) \ln p(l)$$

(vi) *Trapping Time (TT)* It gives a measure of the amount and the length of the vertical structures in the recurrence plot. It is called trapping time as it gives a measure of how long a state is trapped or the system continues in a state. It is calculated as the average length of the vertical lines given by:

$$\mathrm{TT} = \frac{\sum_{\mathrm{v}=\mathrm{v}_{\min}}^{\mathrm{N}} vP(v)}{\sum_{\mathrm{v}=\mathrm{v}_{\min}}^{\mathrm{N}} P(v)}$$

Another feature defined from recurrence plot is the length of diagonal length which can be considered as a measure of the time for which system evolves similarly. Thus, a large number of diagonal lines in RP depicts a deterministic system with the nature of local predictability. Whereas, most random systems will have mostly single points in their RPs. Trajectories will diverge exponentially in chaotic systems due to which diagonal lengths are very short for chaotic systems.

Fractal analysis

Fractal dimension calculates the complexity of signals by exploiting their stochastic nature. It is a non-integer that is useful in detecting the transients in bio-signals like EEG. FDs were utilised in both ECG and EEG signal analysis to study and identify specific physiological states and various disease conditions (Pradhan and Dutt 1993; Yeragani et al. 1998; Jacob and Gopakumar 2018). Studies have reported fractal dimensions for detecting changes in background EEG activity and for identifying irregular patterns like spikes in brain signals (Pradhan and Dutt 1993; Jacob et al. 2019a).

The two commonly used algorithms for finding FD were defined by Higuchi and Katz. In Higuchi's algorithm (Higuchi 1988), a new time series is redefined from the original time series $x(1), x(2)\ldots x(N)$, as:

$$X_m^k = x(m), x(m+k), x(m+2k), \ldots x\left(m + int\left[\frac{N-m}{k}\right] * k\right)$$

Here, m varies from *1 to k*, where m is the initial time instant and k is the time interval; k varies from *1 to* k_{max}.

Next step is to calculate the length of the curve for each of the k new time series as:

$$L_m(k) = \frac{1}{k}\left[\sum_{i=1}^{int\left(\frac{N-m}{k}\right)} |x(m+ik) - x(m+(i-1)k|\right] \frac{N-1}{k.int\left[\frac{N-m}{k}\right]}$$

Here, N is the total number of samples and $\frac{N-1}{k.int\left[\frac{N-m}{k}\right]}$ is the normalisation factor. The length of the curve *L(k)* is taken as the average value of k values of $L_m(k)$. The average curve length for scale k, *L(k)* is proportional to k^{-D} where D is the fractal dimension. FD can be calculated as the slope of least squares linear best fit of the plot *ln(L(k))* versus *ln (1/k)*. Here, HFD was computed with $\mathrm{k}_{\max} = 6$ and window overlap of 75% as proposed by Accardo et al. (1997).

Katz's algorithm calculates fractal dimension with less computational complexity where the distance between two successive points is calculated for computing FD (Katz 1988).

Katz's FD is computed for the time series *x(1), x(2)… x(n)*, as:

$$Katz's\ FD = \frac{\log L}{\log d}$$

where L is calculated as the sum of distances between all successive points and d gives the maximum distance from the initial point to the farthest point.

$$d = max\left(|x_1 - x_j|\right) where\, j = 2,\ 3\ldots N$$

A normalisation factor a has been introduced for making it independent of particular units of measurement as:

$a = mean\ (distance\ (x_{\ i} — x_{\ i\text{-}1}))$.

i.e. $a = \frac{L}{N-1}$

Katz's Fractal Dimension, $KFD = \frac{\log\frac{\mathrm{L}}{\mathrm{a}}}{\log\frac{\mathrm{d}}{\mathrm{a}}} = \frac{\log(N-1)}{\log(N-1)+\log\frac{d}{L}}$

Principal component analysis

Principal component analysis (PCA) is a commonly used dimensionality reduction technique (Cao et al. 2003; Stork et al. 2001; Subasi and Gursoy 2010). In PCA, a higher dimensional data (of n dimension) is represented in a lower dimension (m dimension, $m < n$) of orthogonal features, thus reducing the complexities in space. Each of the resulting *'m'* orthogonal features are called principal components (PC) and its corresponding eigen value represents the variance. The first PC represents the highest variance, second PC represents the next highest variance and so on. All of them are mutually perpendicular to each other (Acharya et al. 2012). Thus, PCA linearly transforms a high-dimensional input vector into a vector with uncorrelated components of lower dimension.

Steps in PCA are:

(i) First the mean of all data is subtracted from the data to get the data set of zero mean. Then, covariance matrix is calculated as

$$C = \frac{1}{l}\sum_{t=1}^{l} x_t x_t^T$$

(ii) The covariance matrix C was decomposed using Singular Value Decomposition (SVD) to get a matrix of eigen vectors (PCs) in an n-dimension. The corresponding eigen values give the variance in the direction of PCs.

The principal components have the properties that they are uncorrelated and have their variances in increasing order. The principal component corresponding to highest eigen value carries maximum information and that the first few principal vectors convey maximum information and thereby allows reduction in the dimension (Cao et al. 2003).

Classifiers

Three distinct classifier methods are utilized in studying the discrimination power of fractal dimensions and RQA features extracted from EEG signals. Support vector machine (SVM) classifier makes use of a partitioning hyperplane and was crafted by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1963. SVM generates a hyper-plane that divides the data points in two distinct classes—with the use of a training vector consisting of labelled data—for segregating new unknown data. The aim of the SVM classifier is to identify the direction that provides maximization of margins and thus it results in the largest separation of given classes in that direction (Suthaharan 2016). In mathematical terms, it is represented with a cost function as given below:

$$\text{Minimize}, J(\omega,\ \omega_0)\ = ||\,\omega\,||^2$$

Given the constraint, $y_i\ (\omega^T x_i + \omega_0) \geq 1,\ i = 1,2,..N$

The hyper-plane is defined with the help of the direction (indicated by ω) and the position (indicated by ω_0). Minimization of the aforementioned cost function J leads to maximization of the margin. Such a quadratic optimization function is constrained with a linear inequality. However, as the cost function is not dependent on the feature space dimensionality, further efficiency in generalization can be obtained for non-separable classes. With such classes, the cost function requires an update using a component comprising the cost of misclassification on the training data.

On data that is non-linear, a kernel function can import the given input feature vector to a high dimensional feature space, allowing a linear separation of the classes. Out of different kernel functions that are used for SVM, radial basis function (RBF) is used for the classification tasks performed here. RBF kernel is represented by,

$$F(x, x_i) = e^{\frac{-||x-x_i{}^2||}{2\sigma^2}}$$

Here σ is a free parameter that can control the width of the kernel (Chung et al. 2003).

Random forest algorithm is a popular ‘ensemble learning’ classifier method given by Breiman that targets combination of results of a group of unique decision trees (Breiman 1999). In this method, training of each of the trees is performed using bootstrap sampled data taken from the training dataset. The nodes of the trees are built with the best predictors from a random set of predictors at any given node in contrast with the original decision tree algorithm. Breiman proved that such a random selection of predictors provides higher classification accuracy, reduced sensitivity to noise in the data and lesser correlation for the selected features (Fraiwan et al. 2012).

On completion of building of the trees, the out-of-bag data is utilized in performing tests of individual trees and the entire forest as well. The decisions generated by component trees are considered and the class gaining the maximum number of votes in the forest is given as the final decision or output of this classifier. The weights given for votes of each individual tree are adjusted as per misclassification error. The Random Forest classifier used for the experiments here is built with 100 trees. In addition, the depth is set at 10 heuristically, resulting in improved generalization performance.

The multilayer perceptron (MLP) is a popular choice of classification algorithm having a system of interconnected artificial neurons (Lippmann 1987; Madyastha et al. 1994). This neural network makes use of a non-linear mapping with a vector comprising of input data to a vector comprising of output data (Gardner 1998). MLP uses a multilayer feed-forward neural network having at least one hidden layer between input and output layers. Optimization of the weights of this neural network are performed using training algorithms, back propagation algorithm being commonly used. It makes use of the gradient search method in computing the optimized value of weights. The cost function minimization of MLP is nonlinear, leading to a good probability of getting trapped to local minima. Such an approach can cause bad classifications where the resulted local minima is not deep enough.

Results and discussion

This work aims at the RQA and fractal analysis of EEG signals and computed various non-linear features based on the same. RQA analysis was performed with recurrence plots and computing various RQA features based on the plots. Figure 4 shows the RQA plot of an encephalopathy case and Fig. 5 that of a normal healthy subject. The recurrence plots’ patterns are helpful in understanding the time evolution of trajectories of signals in each class—normal and disease. The large-scale patterns in these plots exhibits certain characteristics that are helpful in distinguishing the classes. The RQA depicts variations in processes in the brain into laminar states triggered by key moments of observable events. Figure 3 shows disruptions to the randomness in signal characteristics (Marwan et al. 2007). This is expected as the presence of disease condition

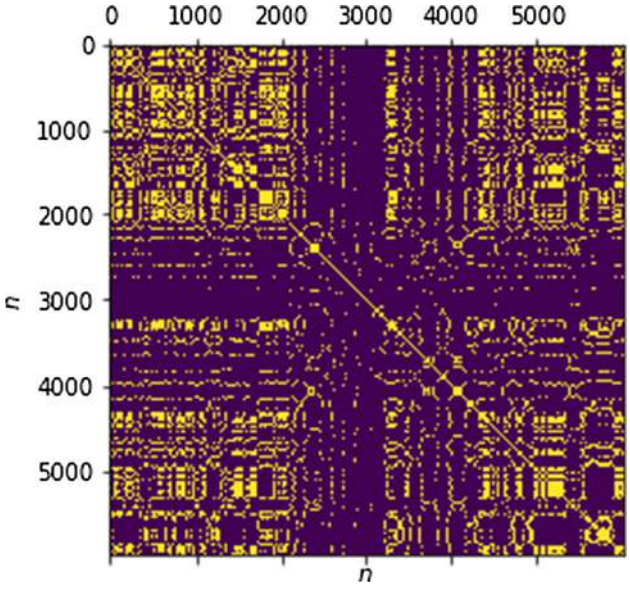

Fig. 4 RQA plot of disease case

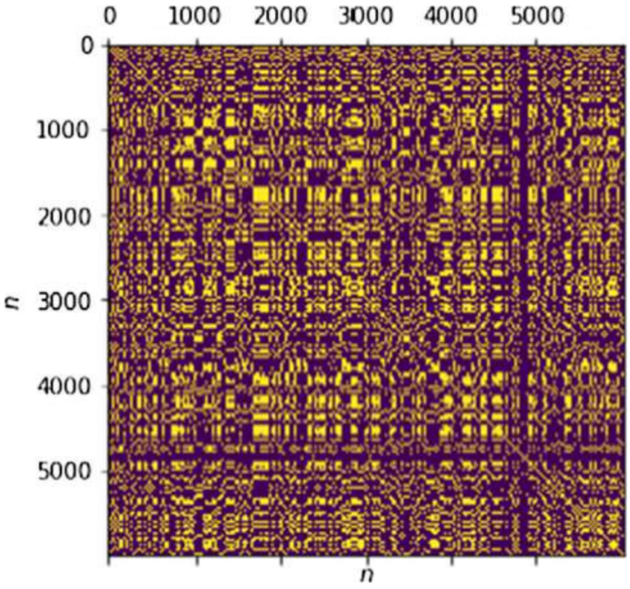

Fig. 5 RQA plot of normal control

would reduce the randomness and thus improve the predictability reducing the chaotic nature of the EEG signals. The continuous dark areas and large bright clusters in the Fig. 4 shows dynamic and unusual extreme events introduced by the disease into the otherwise quasi-stationary EEG signals. On the other hand, the plot for normal subject given in Fig. 5 has a homogenous appearance, where the relaxation times are of less duration in proportion to the time spanned by the whole plot. In Figs. 4 and 5, The value on x and y axes denotes relative time to the forthcoming recurrence points (and not absolute time).

Additionally, the p value analysis helps understand the statistical significance of the concerned features and p value < 0.001 indicates a highly significant feature. Tables 3 & 4 gives the mean values for various RQA features for disease and normal groups. Statistical significance was tested with independent t-test and found that all the computed RQA and FD features were significant with p value < 0.001. The mean values of RQA features (Refer Table 3) clearly show that recurrence is very evident during disease state when compared to that of normal healthy individuals, reporting more predictability and less complexity during encephalopathy. Similarly, Table 4 shows the decreased values of FD for disease compared to that of normal healthy controls in turn proving less complexity of brain dynamics during encephalopathy. This is similar to the trend of other non-linear features like CD, LLE and entropy reported earlier in encephalopathy (Jacob et al. 2018, 2019a). Similar observation was reported in epilepsy and in mild cognitive impairment (MCI) where more periodic dynamics was observed during disease state (Lopes et al. 2021; Timothy et al. 2019). RQA was reported to be effective in epilepsy analysis and reported more recurrences during ictal period compared to normal healthy individuals.

Since there can be non-linear relationships between the features and classes, it is helpful to further assess the features importance using an advanced method—Gini impurity measurement in random forests. The formation of such a tree uses a selected feature at each specific node based on decrease in impurity by such selections. The mean decrease of impurity by the choice of each feature is a measure of importance of the feature. Though the specific values of such scores got no relevance, relative values can be used to understand the features' importance with respect to each other.

Figure 6 shows analysis performed to understand relevance and importance of features. As seen in Fig. 6, all the features seem to be reasonably helpful as none remain unused by the tree and received comparable scores. RQA feature—Recurrence Rate (RR) and the Fractal Dimension (FD) features seem to be playing key role in the classification tasks. Relevance of FDs are proven in the literature in papers from the author and is proven to be superior in terms of class separability (Jacob and Gopakumar 2018; Jacob et al. 2019b). Recurrence Rate measures the relative density of recurring points in this 2-dimensional sparse matrix of recurrence plot. Hence, it is the most distinguishable and visible feature among features generated through RQA methods, making it very useful for distinguishing the classes.

The aforementioned features—6 features from RQA methods and 2 fractal dimensions—are given to Principal Component Analysis for generating derived features.

Table 3 Statistical significance of RQA features in the normal and disease groups

Features	Normal/Disease	No of epochs	Mean	Standard deviation	p value
Recurrence Rate (RR)	Normal	300	0.1656	0.0746	< 0.001
	Encephalopathy	300	0.2426	0.0933	
Determinism (DET)	Normal	300	0.9901	0.0077	< 0.001
	Encephalopathy	300	0.9949	0.0055	
Laminarity (LAM)	Normal	300	0.9934	0.0152	< 0.001
	Encephalopathy	300	0.9982	0.0021	
Length (L)	Normal	300	8.6928	2.072	< 0.001
	Encephalopathy	300	11.1169	2.94	
RQA Entropy	Normal	300	2.9171	0.2861	< 0.001
	Encephalopathy	300	3.1634	0.3103	
Trapping Time (TT)	Normal	300	10.9899	3.6237	< 0.001
	Encephalopathy	300	15.0884	4.7879	

Table 4 Statistical significance of Higuchi's FD and Katz's FD in the normal and disease groups

Features	Normal/Disease	No of epochs	Mean	Standard deviation	p value
Higuchi's Fractal Dimension	Normal	300	1.0895	0.0286	< 0.001
	Encephalopathy	300	1.0566	0.0281	
Katz's Fractal Dimension	Normal	300	2.3104	0.1789	< 0.001
	Encephalopathy	300	1.8344	0.2094	

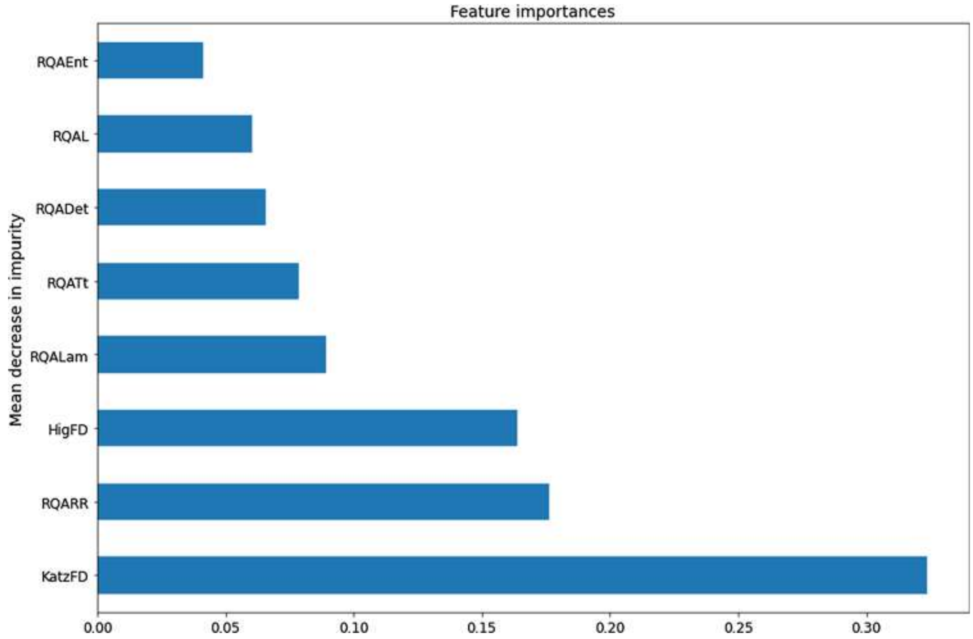

Fig. 6 Feature importance plot

Uncorrelated features generated by PCA is in general helpful in improving the class separability. This is evident from Fig. 7 where top three principal components are used for projecting the data points from two classes—disease (red) and normal (green). As can be seen in Fig. 7, the classes form two clusters overlapping marginally making them separable with efficient use of classifiers.

The choice of principal component is performed heuristically based on the observations from variance plot. The variance plot describes the variance absorbed by each of the principal component as boxes and the line plot shows cumulative sum resulting due to combination of top n number of features, where $n = 2$, 3 etc. Figure 8 makes it evident that an optimal choice of number of principal components can be $n = 3$ as top three components absorb most of the variance and the remaining components' variance—mostly from noises – taken together is negligible. Thus, 3 may be the intrinsic dimension which is

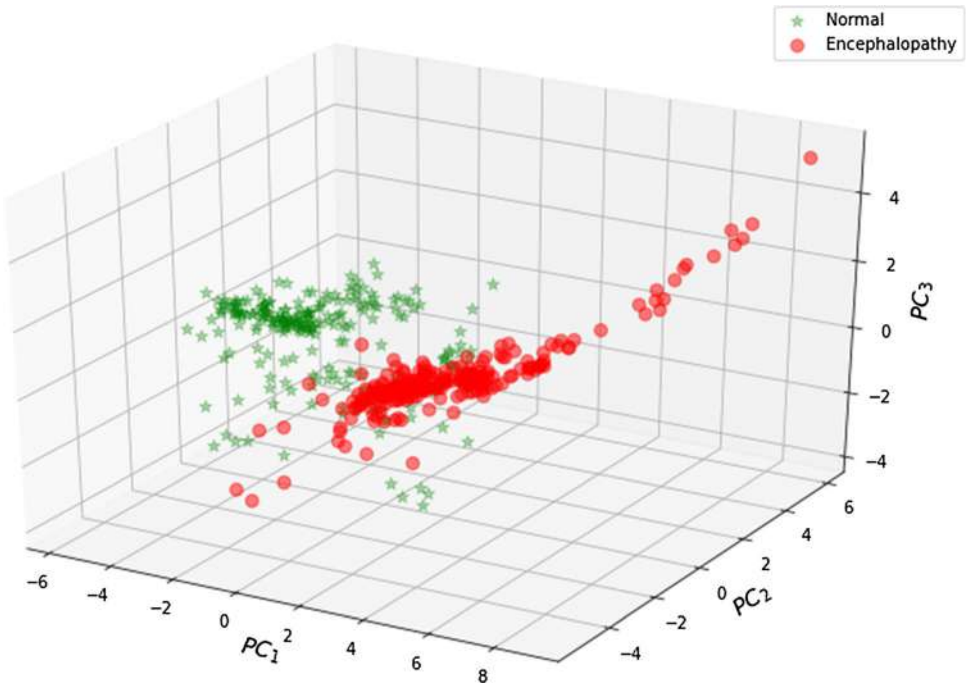

Fig. 7 Data distribution plot (with first three principal components)

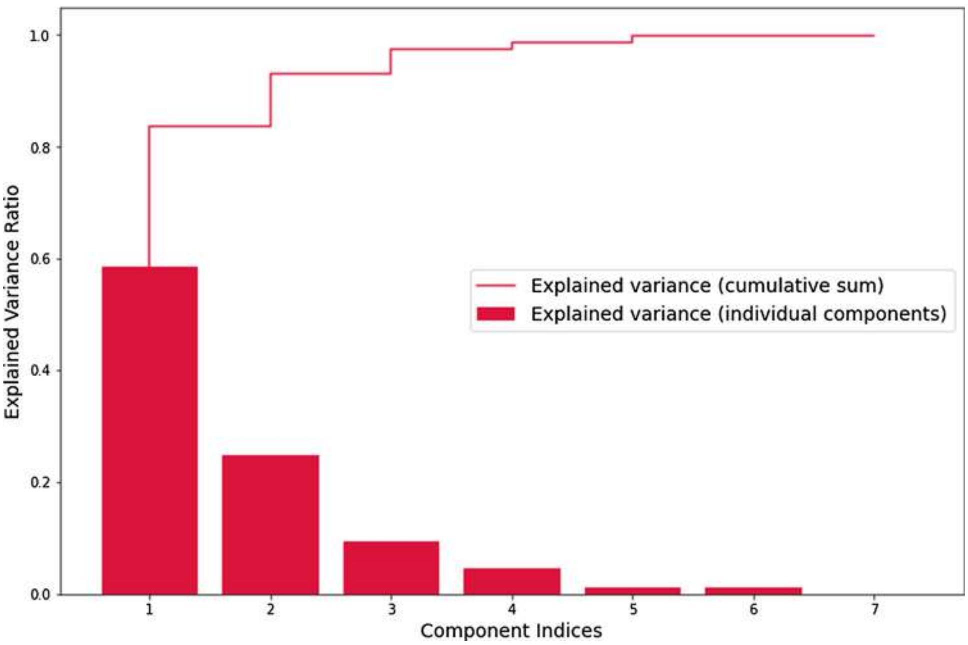

Fig. 8 Variance plot

sufficient enough to explain the data distribution and may make them maximum separable. Anything more than that may not be helpful due to the curse of dimensionality and can increase the complexity of the system unnecessarily, resulting in poor classification performance (Theodoridis and Koutroumbas 2006).

The features generated through PCA is fed to the classifiers completing the pipeline for encephalopathy disease diagnosis. Figure 9 shows variation of classifier performance with respect to increase in number of principal components. All three classifiers—Random Forest, Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP)—all performed really well giving an accuracy of around 95% when peaked.

It can be seen that the classifier accuracy increases with increase in number of principal components until it reaches three and then saturates. Further increase in number of components deteriorates the performance marginally. Since Random Forest classifier internally performs a feature selection which is effective, derived features that removes

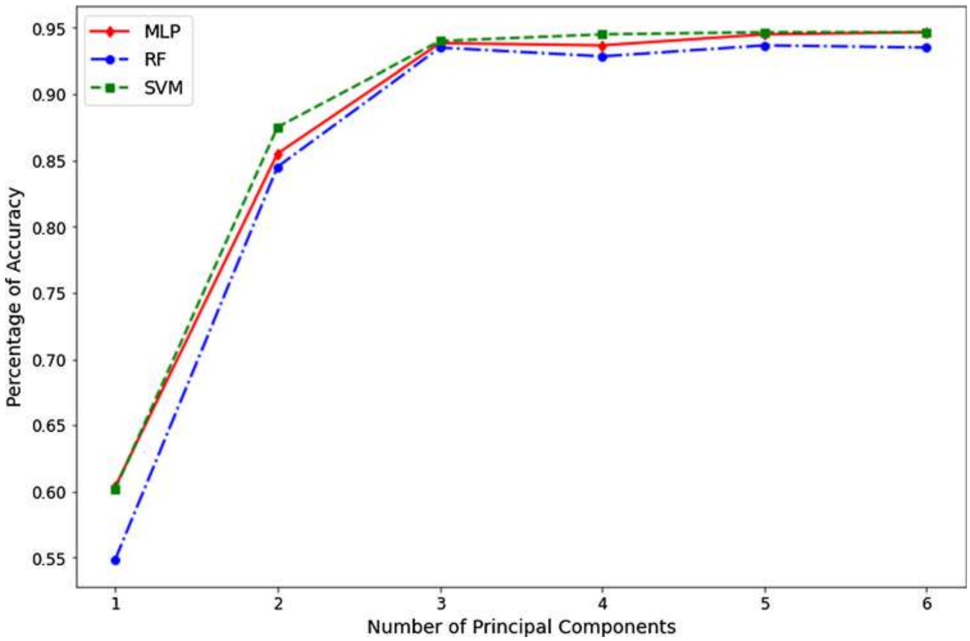

Fig. 9 Performance comparison (Accuracy) of—classifiers RQA and FD features

some information (variance) from the data is not helpful to it. Other two classifiers are getting benefitted by the feature transformation through PCA resulting in improved performance. The ability of all the three classifiers to deal with non-linearly separable classes are reflected in their performances. With further generalizations in terms of the training data representing variabilities, it is possible to use this system for real-life application. Such a realization will be time efficient and can act as a tool ensuring maximum detections an early addressing of the disease.

Area under the curve (AUC) is another measure that is helpful in understanding classifier performance and is robust on skewed distributions as well. Receiver operator characteristics (ROC) plots graphically explains the diagnostic ability of the classifier pipelines at different discrimination thresholds and AUC were calculated for various classifiers. It acts as a mean measure of sensitivity and specificity of the system which has great importance in disease diagnosis. Considering both the scores, SVM classifier seem to be the best performer though by a minimal difference. This should be due to SVM's exceptional capability to deal with high dimensional space when the classes are separable and the hyper-parameters are tuned to optimum values. With AUC curve (Refer Fig. 10) appearing similar to that of accuracy curve, the system seems to be performing equally well on both the classes and is supposed to be reliable for practical applications.

Tables 5 and 6 show performance scores obtained by classifiers with RQA features and combined RQA + FD features respectively. From comparison, it can be seen that the combination of FD and RQA features results in a performance improvement though it is minor. Further, it is also observed that three features (principal components) sufficed to reach maximum performance in case of combination of features whereas five features were required to achieve similar performance in case of RQA features alone. This indicates the availability of more discriminative information to PCA in former case, resulting in considerable separation of distributions of disease and normal patients, in par with the observation from Fig. 6.

The optimal performance is obtained with the use of SVM classifier as seen in Table 6, the performance of which is detailed with the help of a confusion matrix given in Fig. 11. It shows a normalized confusion matrix for the support vector machine giving an accuracy of 94.67% and misclassification rate of 5.33%. Here, the values are normalized over 20 folds—30 data points in each of the 20 test sets—expressed in range [0,1].

Reduction in number of required features can improve the performance in terms of processing power and memory utilization, in addition to the improved classification performance. Hence, the use of combination of features seems justifiable for practical use cases of encephalopathy diagnosis.

Conclusion

Analysis of EEG signals based on Recurrent quantification analysis (RQA) and fractal analysis is found to be promising for the diagnosis of encephalopathy. Both RQA features and FDs were statistically significant for classifying encephalopathy cases from normal healthy controls. The decreased values of RQA features shows more

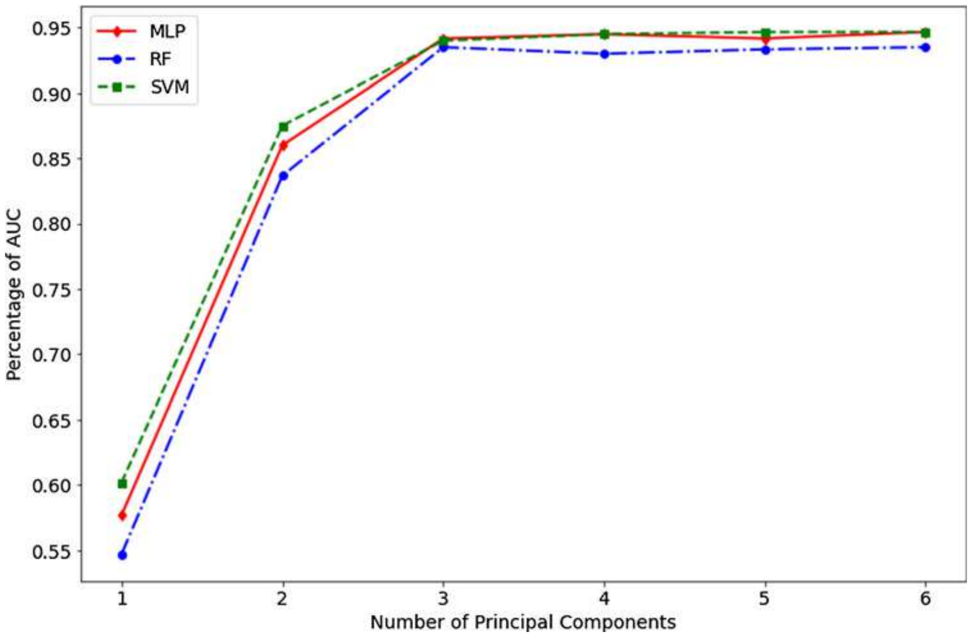

Fig. 10 Performance comparison (AUC) of classifiers—RQA and FD features

Table 5 Performance measurements of classification system using RQA features

Classifier	SVM	Random forest	MLP
Accuracy	94.17	93.17	94.98
AUC	94.17	92.83	94.67

Table 6 Performance measurements of classification system using RQA and FD features

Classifier	SVM	Random forest	MLP
Accuracy	94.67	93.67	94.83
AUC	94.67	93.50	94.67

recurrences and periodic character of brain signals during encephalopathy. Furthermore, lower values of fractal dimensions of EEGs of encephalopathy cases prove the decreased complexity of these signals when compared to that of normal healthy subjects. These results prove the deterministic nature of brain dynamics during encephalopathy. The use of non-linear dimensionality reduction technique—PCA on these chaotic and non-linear features extracted from EEG signals improved discriminative power, resulting in a practically realizable classifier system. The use of aforementioned features with Support Vector Machine resulted in high accuracy of 94.67% for this classification task. With this performance, this study offers a complete framework for the automated diagnosis of encephalopathy based on RQA and FD features of EEG.

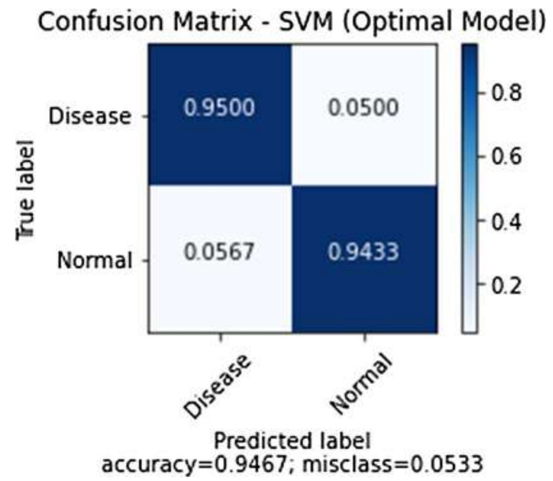

Fig. 11 Normalised confusion matrix for SVM classifier with RQA and FD features which creates the optimal model

Acknowledgements The authors are thankful to the neurologists and EEG technicians for the helpful discussions and for clearing our queries related to this work. This study received the consent from Research Committee and Institutional Ethics Committee prior to the study.

Author contribution JEJ: Conceptualization, Methodology, Validation, Writing–original draft, reviewing and editing. SC: Conceptualization, Methodology, Validation, Writing–original draft, reviewing and editing. AC: Resources, Data Curation. TI: Resources, Data Curation.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability The data are not publicly available as containing information could compromise the privacy of research.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Accardo A et al (1997) Use of the fractal dimension for the analysis of electroencephalographic time series. Biol Cybern 77(5):339–350

Acharya UR et al (2012) Use of principal component analysis for automatic classification of epileptic EEG activities in wavelet framework. Exp Syst Appl 39(10):9072–9078

Angel MJ, Young GB (2011) Metabolic encephalopathies. Neurol Clin 29(4):837–882. https://doi.org/10.1016/j.ncl.2011.08.002

Bickford RG, Butt HR (1955) Hepatic coma: the electroencephalographic pattern. J Clin Invest 34(6):790–799. https://doi.org/10.1172/JCI103134

Blaabjerg L, Juhl CB (2016) Hypoglycemia-induced changes in the electroencephalogram: an overview. J Diabetes Sci Technol 10(6):1259–1267. https://doi.org/10.1177/1932296816659744

Breiman L (2001) Random Forests. Machine Learning 45:5–32. https://doi.org/10.1023/A:1010933404324

Cao L et al (2003) A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. Neurocomputing 55(1–2):321–336

Chung KM et al (2003) Radius margin bounds for support vector machines with the RBF kernel. Neural Comput 15(11):2643–2681. https://doi.org/10.1162/089976603322385108

Faigle R, Sutter R, Kaplan PW (2013) Electroencephalography of encephalopathy in patients with endocrine and metabolic disorders. J Clin Neurophysiol 30(5):505–516. https://doi.org/10.1097/WNP.0b013e3182a73db9

Ferenci P et al (2002) Hepatic encephalopathy—definition, nomenclature, diagnosis, and quantification: final report of the working party at the 11th World Congresses of Gastroenterology, Vienna, 1998. 35(3):716–721

Fraiwan L et al (2012) Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier. Comput Methods Prog Biomed 108(1):10–19. https://doi.org/10.1016/j.cmpb.2011.11.005

Fraser AM, Swinney HL (1986) Independent coordinates for strange attractors from mutual information. Phys Rev A Gen Phys 33(2):1134–1140. https://doi.org/10.1103/physreva.33.1134

Gardner MW, Dorling SJ (1998) Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos Environ 32(14–15):2627–2636 https://doi.org/10.1016/S1352-2310(97)00447-0

Higuchi TJ (1988) Approach to an irregular time series on the basis of the fractal theory. Phys d: Nonlinear Phenom 31(2):277–283. https://doi.org/10.1016/0167-2789(88)90081-4

Jacob JE, Nair GK (2019a) EEG entropies as estimators for the diagnosis of encephalopathy. Anal Integr Circ Signal Process 101(3):463–474

Jacob JE et al (2019b) Application of fractal dimension for EEG based diagnosis of encephalopathy. Analog Integr Circ Signal Process 100(2):429–436. https://doi.org/10.1007/s10470-019-01388-z

Jacob JE and Gopakumar K (2018) Automated diagnosis of encephalopathy using fractal dimensions of EEG sub-bands. In: 2018 IEEE recent advances in intelligent computational systems (RAICS). IEEE. https://doi.org/10.1109/RAICS.2018.8635062

Jacob JE et al (2018) Can chaotic analysis of electroencephalogram aid the diagnosis of encephalopathy?. Neurology Research International. https://doi.org/10.1155/2018/8192820

Jp EJ (1987) Recurrence plots of dynamical systems. Europhys Lett 5:973–977

Katz MJ (1988) Fractals and the analysis of waveforms. Comput Biol Med 18(3):145–156. https://doi.org/10.1016/0010-4825(88)90041-8

Kennel MB, Brown R, Abarbanel HD (1992) Determining embedding dimension for phase-space reconstruction using a geometrical construction. Phys Rev 45(6):3403. https://doi.org/10.1103/PhysRevA.45.3403

Lippmann RJ (1987) An introduction to computing with neural nets. IEEE Assp Mag 4(2):4–22

Lopes MA et al (2021) Recurrence quantification analysis of dynamic brain networks. Eur J Neurosci 53(4):1040–1059. https://doi.org/10.1111/ejn.14960

Madyastha RK et al (1994) An algorithm for training multilayer perceptrons for data classification and function interpolation. IEEE Trans Circ Syst i: Fundam Theory Appl 41(12):866–875. https://doi.org/10.1109/81.340848

Marwan N (2011) How to avoid potential pitfalls in recurrence plot based data analysis. Int J Bifurc Chaos 21(04):1003–1017. https://doi.org/10.1142/S0218127411029008

Marwan N et al (2002) Recurrence-plot-based measures of complexity and their application to heart-rate-variability Data. Phys Rev 66(2):026702. https://doi.org/10.1103/PhysRevE.66.026702

Marwan N et al (2007) Recurrence plots for the analysis of complex systems. Phys Rep 438(5–6):237–329. https://doi.org/10.1016/j.physrep.2006.11.001

Musgrave H, Hilsabeck RC (2019) Hepatic encephalopathy. Handbook on the neuropsychology of aging and dementia. Springer, pp 689–710

Ouyang G et al (2008) Using recurrence plot for determinism analysis of EEG recordings in genetic absence epilepsy rats. Clin Neurophysiol 119(8):1747–1755

Packard NH et al (1980) Geometry from a time series. Phys Rev Lett 45(9):712

Pradhan N, Dutt DN (1993) Use of running fractal dimension for the analysis of changing patterns in electroencephalograms. Comput Biol Med 23(5):381–388

Pritchard WS, Duke DW (1995) Measuring chaos in the brain: a tutorial review of EEG dimension estimation. Brain Cogn 27(3):353–397. https://doi.org/10.1006/brcg.1995.1027

Schomer DL, Da Silva FL (2012) Niedermeyer's electroencephalography: basic principles, clinical applications, and related fields. Lippincott Williams & Wilkins

Selesnick IW et al (2014) Simultaneous low-pass filtering and total variation denoising. IEEE Trans Signal Process 62(5):1109–1124. https://doi.org/10.1109/TSP.2014.2298836

Stork DG, Duda RO, Hart PE et al (2001) Pattern classification. John wiely and sons

Subasi A, Gursoy MI (2010) EEG signal classification using PCA, ICA, LDA and support vector machines. Exp Syst Appl 37(12):8659–8666

Suthaharan S (2016) Support vector machine. Machine learning models and algorithms for big data classification. Springer, pp 207–235

Takens F (1981) Detecting strange attractors in turbulence. Dynamical systems and turbulence, Warwick 1980. Springer, pp 366–381

Theiler J et al (1992) Testing for nonlinearity in time series: the method of surrogate data. Phys d: Nonlinear Phenom 58(1–4):77–94. https://doi.org/10.1016/0167-2789(92)90102-S

Theodoridis S, Koutroumbas K (2006) Pattern recognition. Elsevier

Timothy LT et al (2019) Recurrence quantification analysis of MCI EEG under resting and visual memory task conditions. Biomed Eng: Appl Basis Commun 31(04):1950025. https://doi.org/10.4015/S101623721950025X

Webber C, Marwan N (2015) Recurrence quantification analysis. Understand Compl Syst 10:978–983. https://doi.org/10.1007/978-3-319-07155-8

Yeragani VK et al (1998) Fractal dimension and approximate entropy of heart period and heart rate: awake versus sleep differences and methodological issues. Clin Sci (lond) 95(3):295–301

Zhang Z et al (2020) DWT-Net: seizure detection system with structured EEG montage and multiple feature extractor in convolution neural network. Journal of Sensors 1–13. https://doi.org/10.1155/2020/3083910

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