



International Journal of Intelligent Information and Database Systems

ISSN online: 1751-5866 - ISSN print: 1751-5858
<https://www.inderscience.com/ijids>

Hybrid jellyfish search sine cosine optimisation-based deep learning for big data classification using MapReduce framework on epileptic seizure data

Jamunadevi Chandrasekar, Arul Ponnusamy

DOI: [10.1504/IJIDS.2025.10068155](https://doi.org/10.1504/IJIDS.2025.10068155)

Article History:

Received:	19 July 2023
Last revised:	15 February 2024
Accepted:	13 March 2024
Published online:	23 December 2024

Hybrid jellyfish search sine cosine optimisation-based deep learning for big data classification using MapReduce framework on epileptic seizure data

Jamunadevi Chandrasekar* and
Arul Ponnusamy

Department of Computer Science,
Government Arts College,
Tiruchirappalli-620 022, India
and

Affiliated to: Bharathidasan University,
Trichy-24, India

Email: jamu83@gmail.com

Email: phdarul2004@yahoo.co.in

*Corresponding author

Abstract: The increase in the amount of big data with the technical advances makes the traditional software tools face difficulties and unable to handle them. In the medical field, big data technologies require new frameworks to leverage them. This paper proposes a novel big data classification using a MapReduce framework on epileptic seizure data is proposed. Here, the big data classification is accomplished in a MapReduce framework, wherein the mapper phase is applied with the data partitioned by deep embedded clustering (DEC). The classification is carried out in the reducer phase, where a deep long short-term memory (DLSTM) trained using the jellyfish search sine cosine (JSCS) algorithm is used for epileptic seizure detection based on the salient features determined from the EEG data. The JSCS-DLSTM is investigated for its efficiency based on accuracy, specificity, and sensitivity and is found to record superior values of 0.915, 0.927, and 0.919, respectively.

Keywords: deep embedded clustering; DEC; jellyfish search sine cosine algorithm; JSCS; deep long short-term memory; big data classification; epileptic seizure detection.

Reference to this paper should be made as follows: Chandrasekar, J. and Ponnusamy, A. (2025) 'Hybrid jellyfish search sine cosine optimisation-based deep learning for big data classification using MapReduce framework on epileptic seizure data', *Int. J. Intelligent Information and Database Systems*, Vol. 17, No. 1, pp.1-31.

Biographical notes: Jamunadevi Chandrasekar graduated in 2005 with a degree of Masters of Science from the Bharathiyar University, Coimbatore. She obtained her MPhil in Computer Science from the Periyar University, Salem in 2008. She has 18 years of teaching experience starting from a Lecturer to an Assistant Professor. At present, she is working as an Assistant Professor (Senior Grade) in the Department of Computer Technology-PG at Kongu Engineering College, Perundurai, Erode, Tamilnadu. She published around 30 research papers in national, international journals and conferences. She is

currently pursuing her PhD part time in PG and Research Department of Computer Science, Government Arts College (Affiliated to Bharathidasan University), Trichy-22, Tamil Nadu, India. Her research interests include bigdata in healthcare domain.

Arul Ponnusamy completed his MSc, MCA, B.Ed., MPhil, PhD degrees in Computer Science. Since then, he has been an Associate Professor in the Department of Computer Science at Government Arts College (Affiliated to Bharathidasan University), Trichy-22, Tamilnadu, where his research interests include network security, big data, machine learning, IoT, data mining and image processing. He has the teaching experience for more than 24 years and most of his work has been published in numerous national, international publications and Scopus journals. He has produced five PhD scholars likely in the field of data security. At present, four scholars are pursuing their PhD under his guidance in the field of data security. He has been acting as a NSS Programme Officer for UNIT I. One of his positive traits is constantly willing to learn.

1 Introduction

Over the last three decades, there has been a tremendous change in the hospital environment due to the transformation of digital media data. Additionally, the images have been digitised and the data, like EEGs, have been continuously monitored physiologically thus making electrocardiogram or EEG strips and imaging films outdated. This transformation known as a ‘paperless hospital’ was not just visible in information storage but also brought in a new age of digitisation, wherein all categories of patient information were stored in a digital format. The advancement in computing superpowers, data storage, and computer systems has made artificial intelligence (AI) and big data investigation conceivable in different scientific domains (Chung et al., 2022). With the increased utilisation of the internet, the amount of data that has been generated has grown rapidly and has ended up in a humungous quantity of data termed big data. Generally, big data is advocated as ‘5V’s’ indicating value-data worth, velocity-speed of data, variety – unstructured, semi-structured, or structured, volume-huge dimensionality of data, veracity – accuracy, uniformity, and pre-eminence of data. Hence, big data is known as data, which are dissimilar and with huge dimensionality that gets proliferated and modernised in a very short duration (Sathyaraj et al., 2020). In the current era, big data has been embraced by several practitioners and researchers in neuroscience. Brain big data holds excessive possibilities in revealing the inherent mechanisms driving brain activities, and maintains the neural recordings of numerous patients under different scenarios in the long term. The current progress in computational intelligence offers an extraordinary opportunity to use brain big data for probing brain dynamics. Organised measuring has been a hot research topic in neuroscience concerning both the functioning and malfunctioning of the brain (Ke et al., 2018).

Epilepsy is a commonly occurring lingering brain disease triggered by the anomalous discharge of neurons and it leads to transient and sudden onset of seizures. According to a report by the World Health Organization Report (2017), over 50 million individuals are affected by epilepsy globally out of which one-fourth are children. People suffering from epilepsy suffer from great pain and it has transformed into a grave community health

problem (Wei et al., 2019). This lethal disease may cause crucial medical indications, such as unconsciousness, strange sensations, muscle cramps, abnormal behaviour, and so on, and it can result in severe injuries, damage to the brain as well as the death of the subjects while at work in hazardous work environments, or road accidents. These uncontrolled seizures, despite occurring at low frequencies, can significantly affect the life quality of those affected (Singh and Malhotra, 2022). The patients normally suffer from high anxiety owing to the volatile characteristics of the seizure and the associated consequences, which degrade the quality of life. These issues have enthused scholars and medical practitioners to create efficient seizure prediction models (D'Alessandro et al., 2003). Timely prediction of seizures with high precision will make it possible to create personalised epilepsy treatments by utilising customised therapies with minimal side effects. The patients can undertake precautionary measures based on the warning regarding imminent seizures thereby avoiding any potential injuries (Hussein et al., 2019). Though seizure prevention is impossible, prediction of the impending seizure can act as a critical tool to improve the condition of epileptic patients. Hence, medical experts and doctors observe the electrical activity happening in the epileptic patient's brain using EEG signals for studying and diagnosing epileptic seizures (Bhattacharya, 2022).

The EEG recordings normally contain an immense quantity of data and so to process the EEG signals, methods that function well with big data are needed for handling the feature extraction and categorisation processes. This trend has been observed in various domains, like medical image and signal processing, and data mining (Emara et al., 2022). Over the last few decades, there has been a focus on research on automated seizure diagnosis, and several techniques, like nonlinear approaches (Zhang et al., 2015), spatio-temporal domain approaches (Faust et al., 2015), and many machine learning (ML) methodologies have been devised and some have produced superior results. Recently, there has been a gradual growth in the usage of ML techniques for automatically detecting seizures in comparison to conventional signal processing approaches (Wei et al., 2019). Deep learning (DL), a class of ML, is an efficient computational tool, which permits automatic learning of features from the data. The DL approaches are typically employed for training specific algorithmic techniques called deep neural networks (DNN) for carrying out particular tasks. The accessibility of big data has paved the utility of DL for a varied array of issues (Kiral-Kornek et al., 2018). The progress in DL and big data analytics has cemented the way for new techniques that are ideal for seizure prediction making it easier to solve it. This further makes it easier to analyse the huge amount of complicated EEG signals with superior efficiency. Moreover, the analysis of high-level features and the detection of hidden patterns in the signal can be performed. These factors favour DL approaches, like convolutional neural networks (CNN) for predicting epileptic seizures. CNNs are broadly utilised in EEG classification, image processing, and computer vision, and can be utilised for observing the spikes in the EEG signals and it is considered a breakthrough in seizure detection (Bhattacharya, 2022).

1.1 Motivation

People with epilepsy often lose consciousness when a seizure occurs, thus causing severe injuries, damage to the brain as well as the death of the subjects while at work in hazardous work environments, or road accidents early prediction can significantly

minimise the potential threats, thereby minimising the risk of casualty and injuries. Though several works have addressed the issue of creating effective approaches to predict seizures from EEG signals, the huge amount of EEG data poses a great challenge.

1.2 Contribution

This research puts forward an efficient optimised DL model for accomplishing big data classification on epileptic seizure data using the MapReduce framework. The input EEG big data is first partitioned using deep embedded clustering (DEC) and is applied to the MapReduce framework that contains the Mapper and Reducer phases. The partitioned data is subjected to multiple mappers in the mapper phase, where it is pre-processed utilising the Kalman filter. Later, several statistical and spectral features are excerpted from the data. Thereafter, the most relevant features are chosen using the Weighted Euclidean Distance (WED) and Bhattacharyya Distance to minimise the feature dimensions. The chosen features are then applied to the reducer phase, where it is aggregated and processed. In the reducer phase, the deep long short-term memory (DLSTM) tuned by the jellyfish search sine cosine (JSCS) approach predicts epileptic seizures.

The main contribution of this paper is given below:

- *Proposed JSCS-DLSTM for epileptic seizure prediction:* the JSCS-DLSTM is employed for the prediction of epileptic seizures from the EEG data. Here, big data classification is executed by employing the DLSTM for predicting epileptic seizures. The parameters of the DLSTM are adjusted by the JSCS algorithmic approach.
- *Proposed JSCS:* it is formed by amalgamating the jellyfish search optimiser (JSO) and the sine cosine algorithm (SCA).

1.3 Structure of the paper

The residual part of this work has the ensuing arrangement: Section 2 illustrates the available epileptic seizure prediction works, the proposed JSCS-DLSTM is expounded in Section 3, Section 4 summaries the experiment results and assessment, and Section 5 provides the conclusion with scope for future work.

2 Literature review

Ghosh et al. (2021) developed a network-on-chip (NoC) based CNN (NoC-CNN) for classifying epileptic seizures in real-time. This method possessed the advantages of scalability, versatility, and reliability, although it failed to improve classification accuracy. Singh and Malhotra (2022) proposed a two-layer long short-term memory (2L-LSTM) network to predict epileptic seizures from EEG signals. This approach had a minimal loss and was highly appropriate for predicting seizures in real-time, although the technique required a long training time. Hussein et al. (2019) created a multi-scale CNN for determining the features automatically from EEG data for epileptic seizure prediction. This technique was successful in determining the highly complex and local features from the intracranial EEG concurrently but the system's performance was degraded by the existence of outliers in the data. Nejedly et al. (2023) used a temporal autoencoder for

clustering and classifying EEG signals for predicting epileptic seizures. This system had a low cost and was able to attain superior outcomes even with limited labelled samples, although the technique suffered from reliability issues as only one expert scientist was used for scoring the pseudo-prospective testing set. Bhattacharya (2022) created a temporal CNN + LSTM model for predicting epileptic seizures automatically. This model was highly accurate and efficient and produced zero false positive rates, although real-time implementation was not carried out. Emara et al. (2022) presented a frequency-based approach using scale-invariant feature transform (SIFT) to forecast epileptic seizures. This scheme had minimal computational complexity but it suffered from high prediction time. Wei et al. (2022) employed a two-dimensional CNN (2DCNN) for epileptic seizure prediction. The 2D CNN was effective in avoiding overfitting and had good stability, but the performance was low when compared to the 1D CNN. Wei et al. (2019) proposed a CNN for detecting epileptic EEG signals considering the task as big data classification. This approach had good generalisability and was successful in handling the data imbalance issue during training but when considering transient seizures, the segmentation accuracy attained was very low.

2.1 Challenges

The succeeding points brief the issues met by the epileptic seizure prediction methods contemplated in the literature.

- The key problem encountered by the NOC-CNN devised in Ghosh et al. (2021) was that it was futile to use more mechanisms and methods in developing the NoC framework to allow the mapping of more intricate models to improve computation efficiency.
- The 2L-LSTM developed in Singh and Malhotra (2022) for predicting seizures in epileptic patients produced a highly accurate prediction. However, the method suffered from marginal variations in evaluation measure values with various EEG segment durations.
- In Nejedly et al. (2023), a temporal autoencoder was proposed, which permitted unbiased, efficient, and fast exploration of data in a large heterogeneous dataset. However, the method failed to generalise the scalp EEG data for inferring the performance as it has a wide range of neurological diseases.
- The 2D CNN was introduced in Wei et al. (2022) for predicting seizures from multiple lead EEG signals, and the method had a minimal training loss. Although the high dimensionality of the 2D CNN complicated the training process, it remained a key issue.
- Epileptic seizures occur very rarely and hence predicting them based on the limited amount of EEG makes the prediction task a complex process. Further, the variation in the EEG signals between different peoples and the state of the brain of the same individual complicates the process.

3 Proposed JSCS-DLSTM for big data classification on epileptic seizure data

Seizures are the most common problems faced by epileptic patients, which limits their ability to control or move any body parts. Seizure prediction can facilitate proper treatment thereby preventing any fatal incidents. Seizures are normally predicted using EEG signals, however, the huge amount of data created while recording the EEG signals causes large processing delays. Under such scenarios, big data classification methods can be effectively used to alleviate the need for high computational requirements. In this research, a novel big data classification scheme is devised for predicting epileptic seizures from the EEG data. At first, big data EEG signal is accumulated from the dataset and subjected to data partitioning using DEC (Xie et al., 2016) to augment performance, decrease contention, and enhance scalability. Later, the partitioned data is fed to the MapReduce framework which consists of a mapper phase, and a reducer phase. In the Mapper phase, each block of data is fed to a mapper for finding the prominent features in the data. Here, the input data is first applied to the pre-processing phase, wherein the Kalman filter (Fox et al., 2003) is used to eliminate the noises in the data. Thereafter, the enhanced data is subjected to the feature extraction phase, where features, like statistical features (Diykh et al., 2016), such as median, mean, skewness, kurtosis, standard deviation (SD), and fast Fourier transform (FFT), activity, complexity (Bablani et al., 2018), and the spectral features, including power spectral density (PSD), spectral centroid, spectral skewness, and pitch chroma, logarithmic band power (LBP) (Hong et al., 2018), and spectral flux are extracted. Afterwards, feature selection is carried out using WED and Bhattacharyya Distance (Cha, 2007) to minimise the feature dimensions. The features selected in the mapper phase are fed to the reducer phase, wherein big data classification on epileptic seizure data is performed using DLSTM (Zhu et al., 2016). A novel JSCS algorithm created by incorporating the JSO (Chou and Molla, 2022) and the SCA (Mirjalili, 2016) is used to train the DLSTM. Figure 1 demonstrates the structural view of the devised JSCS-DLSTM for big data classification using the MapReduce framework on epileptic seizure data.

3.1 Data acquisition

The foremost process in epileptic seizure prediction is to accumulate the big data from the dataset L , and it is illustrated as:

$$L = \{L_1, L_2, \dots, L_i, \dots, L_l\} \quad (1)$$

wherein l indicates the overall data record count in the EEG dataset, among which the i^{th} big data EEG signal L_i is used for classification.

3.2 DEC for big data EEG signal partitioning

Big data is generally characterised by 3Vs, and the data is highly complex, dynamic, and large, which makes it highly challenging to process the big data. Thus effective techniques are required for mining the big data in order to understand the hidden and beneficial patterns in the input data. Clustering is an effective approach for determining the patterns in big data. Here, the input big data is partitioned into several clusters or

subsets by the unsupervised clustering approach named DEC (Xie et al., 2016). This is mainly carried out to address the shortage of feasibility in checking the entire data and for easy management. DEC is mainly employed in data partitioning due to its better performance and scalability. Further, DEC is effective in learning the feature representations and assigning clusters with the utilisation of DNN. Moreover, an auxiliary target distribution attained from the soft cluster allocation is used for iteratively refining the cluster. Here, the data L_i is given as input to the DEC, which utilises nonlinear mapping to transform the data to carry out partitioning. The input data is mapped into a feature space with low dimensionality for optimising the clustering objective and this is learned by the DEC. Here, two phases, namely:

- a initialisation of parameters, where a deep autoencoder is used
- b optimising parameters, where an iterative process is carried out between reducing the Kullback-Leibler (KL) divergence and computation of an auxiliary target distribution.

Assume that there exists a total of a points $\{z_j \in Z\}_{j=1}^a$ that have to undergo partitioning into b clusters, which are depicted by a centroid $v_i, i = 1, \dots, b$. DEC utilises a nonlinear mapping $f_\phi: Z \rightarrow C$ to transform the data latent feature space C , and ϕ signifies the learnable parameters. DNN is used to parameterise the term f_ϕ . The DEC concurrently learns the DNN parameter ϕ and the b cluster sets in the feature space C .

3.2.1 Clustering with KL divergence

An unsupervised technique, which switches among two steps is used for improving the clustering given an initial value of ϕ and $\{v_i\}_{i=1}^b$. At first, a soft assignment is determined among the embedded point and the cluster centroids. Thereafter, auxiliary target distribution is used to learn the current high-confidence allocations for refining the cluster centroids, the deep mapping f_ϕ is also restructured, and the procedure is iterated until the convergence is attained.

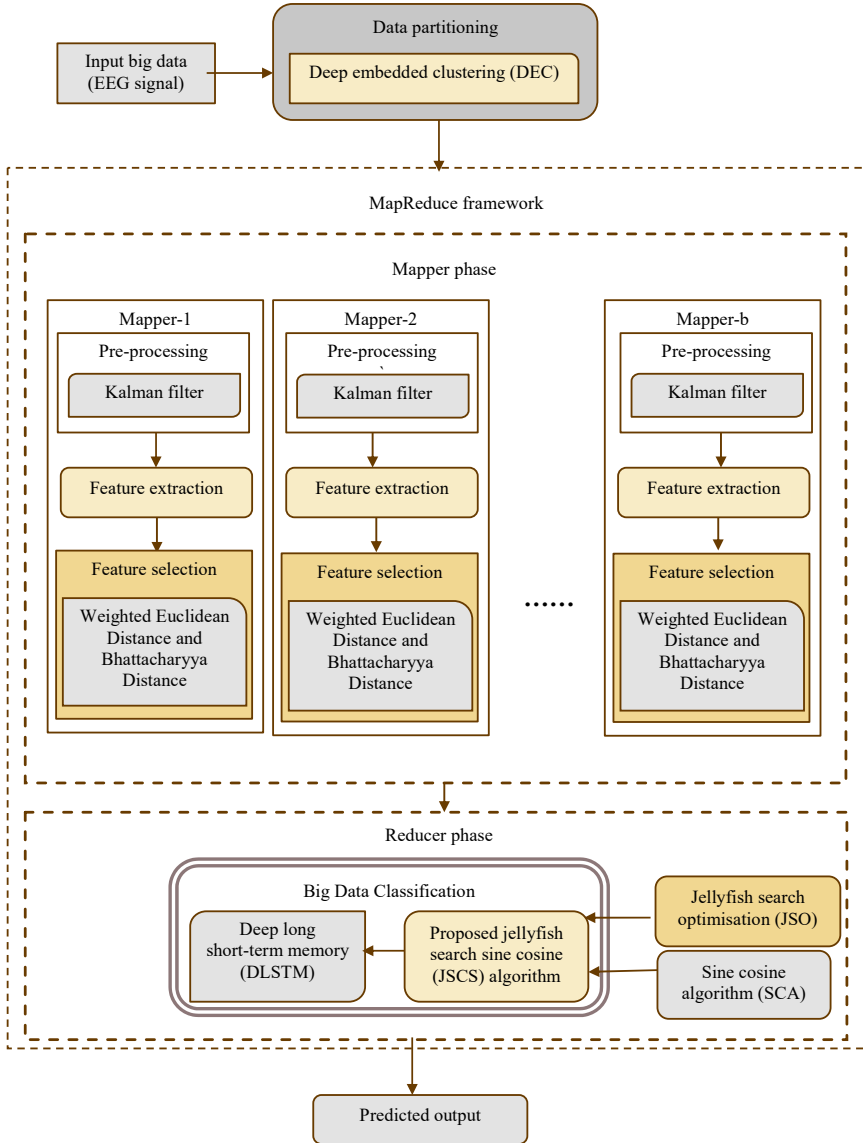
3.2.1.1 Soft assignment

The closeness among the embedded point c_j and the centroid v_i is measured using a kernel namely the student's t-distribution as given below,

$$d_{ji} = \frac{\left(1 + \|c_j - v_i\|^2 / \delta\right)^{-\frac{\delta+1}{2}}}{\sum_{i'} \left(1 + \|c_j - v_{i'}\|^2 / \delta\right)^{-\frac{\delta+1}{2}}} \quad (2)$$

Here, d_{ji} designates soft assignment, which is the likelihood of allocating a sample j to a cluster i , δ signifies the degree of freedom, and $c_i = f_\phi(z_j) \in X$ is associated with $z_j \in Z$ after embedding.

Figure 1 Structural view of the JSCS-DLSTM for big data classification using MapReduce framework on epileptic seizure data (see online version for colours)



3.2.1.2 KL divergence reduction

An auxiliary target distribution is utilised for refining the clusters repeatedly depending on the learned data from the high-confidence allocations. The soft assignments are matched to the target distribution for training the model and the expression given below represents the KL divergence loss.

$$Loss = KL(E\|D) = \sum_j \sum_i e_{ji} \log \frac{e_{ji}}{d_{ji}} \quad (3)$$

Here, E refers to the target distribution, and the term $Loss$ indicates the KL divergence loss among the target distribution e_{ji} and d_{ji} . The term d_j is first raised to the second power and then its normalisation is then performed by frequency per cluster to compute the value e_j as given below,

$$e_{ji} = \frac{d_{ji}^2 / f_i}{\sum_{i'} d_{ji'}^2 / f_{i'}} \quad (4)$$

where the soft clusters are signified by $f_i = \sum_j d_{ji}$. After pre-training the autoencoder, fine-tuning is carried out to minimise the KL convergence loss and thus the data is effectively partitioned by the DEC into various clusters $M_1, M_2, \dots, M_j, \dots, M_b$, and these are applied to the next stage.

3.3 MapReduce framework

The partitioned data $M_j, j = 1, \dots, b$ is fed to the MapReduce framework, which is a model used for processing big data at several nodes in parallel. MapReduce model is mainly used to address the high complexity and quantity of big data. It contains two stages, a mapper and a reducer stage. The mapper phase converts the input data into small chunks by processing it. At the reducer, the data chunks received from the mappers are combined and then processed to obtain the outcome. Due to the huge amount of EEG signals accumulated for epileptic seizure prediction, the MapReduce framework is used in epileptic seizure prediction.

3.3.1 Mapper phase

This is the primary phase in the MapReduce framework, and it comprises multiple mappers. Here the partitioned big data $M_j, j = 1, \dots, b$ is applied to each mapper, which processes the clusters separately in parallel. Every mapper is composed of three modules, such as the pre-processor, feature extractor, and feature selection. The mappers convert the input big data into reduced-size feature vectors, which are fed to the reducer phase. The process occurring at the mapper is detailed below.

3.3.1.1 Pre-processing

At the mapper, the input data M_j is primarily applied to the Kalman filter (Fox et al., 2003) for denoising the signal in the pre-processing step. The Kalman filter is utilised for eradicating the noise in the EEG signal owing to its high computational efficiency and accurate results even with nonlinear inputs and can be utilised even in uncertain situations. It is a type of Bayes filter, where the approximation of beliefs is carried out based on the first and second moments, and it is similar to the unimodal Gaussian filter, and is expressed as:

$$Bel(M_j) \approx P(M_j; \mu_j; \beta_j) = \frac{1}{(2\pi)^{l/2} |\beta_j|^{l/2}} \exp \left[-\frac{1}{2} (M_j - \mu_j)^T \beta_j^{-1} (M_j - \mu_j) \right] \quad (5)$$

where $Bel(M_j)$ indicates the representation of the beliefs, μ_j is the first moment (mean), β_j denotes the second moment (covariance) matrix with dimension $l \times l$, with l denoting the dimension of the state, and $P(M_j; \mu_j; \beta_j)$ designates the probability of M_j given μ_j and β_j . Consider the filtered output thus produced to be depicted as H .

3.3.1.2 Feature extraction

Once the EEG big data is denoised using Kalman filtering, the most discriminatory features from the EEG signals have to be identified to achieve dimensionality reduction and decrease the computational complexity. Here, various features, such as statistical features (Diykh et al., 2016), like median, mean, skewness, kurtosis, SD, and FFT (Murugappan and Murugappan, 2013), activity, complexity (Bablani et al., 2018), and the spectral features, including PSD, spectral centroid, spectral skewness, and pitch chroma, LBP (Hong et al., 2018), and spectral flux are extracted from the filtered EEG data H .

3.3.1.2.1 Statistical features

This includes the simplest features that are used to determine the characteristics of the EEG signal. Here, features, namely mean, kurtosis, median, SD, and skewness are computed from the Kalman filter's output H . Consider the pre-processed EEG signal to be indicated as $H = \{h_1, h_2, \dots, h_R\}$, where R indicates the overall count of data points.

- *Mean*: mean is used to find the average of the samples of the EEG big data, and is given by,

$$F_{mean} = \frac{1}{R} \sum_{l=1}^R h_l \quad (6)$$

where F_{mean} indicates the mean.

- *Median*: it determines the middlemost value in the EEG data and is expressed as

F_{median} , and is determined by finding out the $\left(\frac{R+1}{2}\right)^{th}$ data point.

- *SD*: SD quantifies the measure of fluctuations from its mean value, and is attained by finding the square root of the variance, and is formulated by,

$$F_{SD} = \sqrt{\frac{1}{R} \sum_{l=1}^R (h_l - F_{mean})^2} \quad (7)$$

where F_{SD} denotes the SD feature.

- *Kurtosis*: Kurtosis quantifies the variation of the tail of the EEG signal distribution from the normal distribution, and is expressed as,

$$F_k = \frac{1}{R} \frac{\sum_{l=1}^R (h_l - F_{mean})^4}{F_{SD}^4} \quad (8)$$

where F_k signifies the kurtosis feature.

- *Skewness*: it is used to determine the symmetry of the EEG signal around its mean value and is quantified as,

$$F_{sk} = \frac{1}{R} \frac{\sum_{l=1}^R (h_l - F_{mean})^3}{F_{SD}^3} \quad (9)$$

Here, F_{sk} terms the skewness value.

3.3.1.2.2 FFT

Further, FFT features are also excerpted from the denoised EEG signal, where FFT is the quickest method to compute the discrete Fourier transform (DFT). FFT is useful in extracting features, such as spectral centroid, and spectral entropy, which provide more details about the variations of the EEG signals from four bands, like alpha, gamma, beta, and alpha to gamma. A Butterworth filter of order 4 with distinct cut-off frequencies is employed for determining the four frequency bands, and is given by,

$$V_g = \sum_{i=0}^{A-1} v_i \exp^{-2\pi g \frac{i}{R}} \quad g = 0, 1, 2, \dots, R-1 \quad (10)$$

Here, g indicates the count of FFT points, V_g specifies the FFT coefficient, and the FFT feature thus created is designated as F_{FFT} .

3.3.1.2.3 PSD

PSD (Usman et al., 2019) denotes the power contained in the EEG signal per unit frequency, and is computed by,

$$F_{PSD} = \sum_{l=0}^R \alpha_y(l) e^{-j\omega l} \quad (11)$$

Here, F_{PSD} refers to the PSD feature, α is the interval length, $\alpha_y(l)$ refers to the autocorrelation function, and ω is the frequency.

3.3.1.2.4 Spectral centroid

It (Murugappan and Murugappan, 2013) is one of the effective features in music (sound) identification, and it is attained by utilising the magnitude and frequency of the Fourier transform to evaluate the ‘centre of gravity’. It can be demarcated as the ratio of the mean frequency weighted by amplitude to the aggregate value of the amplitude, and is expressed as:

$$F_{SC} = \frac{\sum_{g=1}^R gI[g]}{\sum_{g=1}^R I[g]} \quad (12)$$

where F_{SC} specifies the spectral centroid, and $I[g]$ signifies the amplitude associated with the FFT spectrum's g^{th} bin.

3.3.1.2.5 Activity

This (Bablani et al., 2018) is computed by finding the squared value of the SD, and is expressed using the following expression,

$$F_{act} = \frac{1}{R} \sum_{l=1}^R (h_l - F_{mean})^2 \quad (13)$$

where F_{act} signifies the activity feature.

3.3.1.2.6 Complexity

This parameter (Bablani et al., 2018) quantifies the proportion of the derivation of the signal mobility to the signal mobility, and is expressed as:

$$F_{com} = \frac{\vartheta(h'_l)}{\vartheta(h_l)} \quad (14)$$

where ϑ indicates mobility and F_{com} denotes complexity.

3.3.1.2.7 Pitch chroma

This feature (Mannepalli et al., 2017) is used for categorising the pitches meaningfully and approximating it to the matching tempered pitch scale. It minimises the noises and transients contained in the signal and is computed by mapping the chroma or semi-tone pitch categories to their respective semi-tone pitches and is expressed as:

$$F_{PC} = \sum_{l': e(l'=p)} J(l'); \quad p \in [0, 12] \quad (15)$$

where $J(l')$ refers to the semi-tone pitch spectrum, and F_{PC} refers to pitch chroma.

3.3.1.2.8 LBP

The LBP (Hong et al., 2018) feature is computed from the logarithms of the power of various EEG data bands, and is computed based on the PSD of all bands and then computing the log of the maximal value, and is given by,

$$F_{LBP} = \log(F_3) \quad (16)$$

where F_{LBP} denotes the LBP feature.

3.3.1.2.9 Spectral skewness

Spectral skewness (Usman et al., 2019) yields the symmetry of the spectrum about its mean and is formulated as:

$$F_{SS} = \frac{\sum_{\omega} ((\omega - F_{SC})/VC)^3 F_{PSD}(\omega)}{\sum_{\omega} F_{PSD}(\omega)} \quad (17)$$

where VC refers to the variational coefficient, and F_{SS} signifies spectral skewness.

3.3.1.2.10 Spectral flux

This feature (Mannepalli et al., 2017) is used to determine the signal's spectral components and it is significant as with time the spectral components vary which degrades the performance of the system. The spectral flux at a time t is determined by computing the squared variation in the signal frames from the normalised magnitudes of successive spectral distributions, and is formulated as:

$$F_{SF} = \sum_{g=1}^{R/2} (|H(g)| - |H_{t-1}(g)|)^2 \quad (18)$$

where R signifies the vector length, and F_{SF} terms the spectral flux.

Later, all the features are combined to form a feature vector F expressed as:

$$F = \left\{ F_{mean}, F_{median}, F_{SD}, F_k, F_{sk}, F_{FFT}, F_{PSD}, F_{SC}, F_{act}, F_{com}, F_{PC}, F_{LBP}, F_{SS}, F_{SF} \right\} \quad (19)$$

where F_{mean} indicates the mean, the median is expressed as F_{median} , F_{SD} denotes the SD feature, F_k signifies the kurtosis feature, F_{sk} terms the skewness value, FFT feature is designated as F_{FFT} , F_{PSD} terms the PSD feature, F_{SC} specifies the spectral centroid, F_{act} signifies the activity feature, F_{com} denotes complexity, F_{PC} refers to pitch chroma, F_{LBP} denotes the LBP feature, F_{SS} signifies spectral skewness, and F_{SF} terms the spectral flux. The vector F is then fed to the feature selection module for choosing the relevant features.

3.3.1.3 Feature selection

Though feature extraction provides a reduced data dimensionality, it may contain some redundant and irrelevant features that degrade the classifier's performance. Feature selection aims at retrieving the most relevant features in the data, thereby reducing computational and storage costs. Further, it also provides better interoperability and readability of the features. Here, feature selection is accomplished with the help of two similarity measures, such as the WED, and Bhattacharya distance. At first, the WED is applied to the determined features F to refine the features and later, the Bhattacharya distance (Cha, 2007) is used to refine the features further thereby attaining a reduced feature set.

- WED

The WED is a similarity measure that computes the distance between two features based on the weighted values of the squared difference among the features, and is expressed as:

$$\chi_1 = \sqrt{\sum_k \tau_k (m_k - n_k)^2} \quad (20)$$

where m_k and n_k indicates the candidate and target features, and τ_k indicates the weight, which is assigned a value of 1 if $m_k = 0$, and $\tau_k = m_k$ otherwise. After determining the WED for all features, the top r features which have the least distance are chosen, thus a reduced dimension feature vector F_{WED} with dimension of $\alpha \times r_2$ is obtained from the input feature vector F with dimension $\alpha \times r_1$, with $r_1 > r_2$.

- Bhattacharya distance:

The features generated F_{WED} are further refined by computing the Bhattacharya distance (Cha, 2007) to compute the similarity between the target and obtained features. The Bhattacharya distance is formulated as:

$$\chi_2 = \left| \log_{10} \left(\frac{\sum_k \sqrt{|m_k - n_k|}}{\sum_k m_k} \right) \right| \quad (21)$$

Once the Bhattacharya distance is measured, the top r_3 features with the least distance value are chosen to obtain a reduced feature set F_{BT} with dimension $\alpha \times r_3$, such that $r_2 > r_3$, and these features are applied to the reducer phase.

3.3.2 Reducer phase

The reducer does the job of combining the feature vector F_{BT} generated from the different mappers in the mapper phase and classifying the data to yield the final output. Here, the data obtained from the various mappers are shuffled, aggregated, and processed to predict epileptic seizures. Consider the aggregated features to be represented as S .

3.3.2.1 Big data classification

The EEG big data is classified for epileptic seizure prediction in the reducer phase using the DLSTM (Zhu et al., 2016) based on the aggregated selected features S obtained from the various mappers. The DLSTM effectively captures the nonlinear aspects of the EEG data. Further, the process of epileptic seizure prediction is augmented by the application of the JSCS algorithm, which is formulated by adjusting the JSO (Chou and Molla, 2022) based on SCA (Mirjalili, 2016).

3.3.2.1.1 Architecture of DLSTM

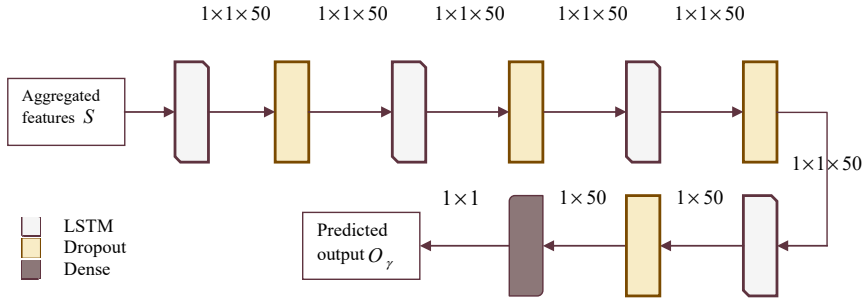
DLSTM (Zhu et al., 2016) comprises a stack of multiple LSTM networks, wherein the output obtained from one network is fed to the next one sequentially. The aggregated selected features S are applied to the DLSTM's input node for predicting epileptic seizures. The LSTM is a category of recurrent neural network (RNN) and is extremely

efficacious in accomplishing sequential learning and determining long-range dependencies. While the selected feature set S is fed to the DLSTM, the output produced at any instance t can be formulated as,

$$k_t = \psi(G_{Sk}S_t + G_{kk}k_{t-1} + p_k) \quad (22)$$

where k_t designates the output response, $\psi(\cdot)$ designates the activation function, G_{Sk} signifies the weight matrix among the input layer and the hidden layer, G_{kk} is the recurrent weight matrix of the hidden layer, and p_k denotes the bias vector. The recurrent weight matrix G_{kk} is employed to explore temporal dependencies.

Figure 2 Structural depiction of the DLSTM model (see online version for colours)



The LSTM encompasses modules, such as the input gate $f(t)$, forget gate q_t , cell w_t , and output y_t . Here, the flow of data inside and out of the cell is controlled by the input and forget gates. The memory cell ensures that the gradient can be propagated back several times without explosion or disappearing, and it have a recurrent edge, which is self-connected and has a unit weight. This guarantees that the problem of vanishing gradient occurring in RNN is completely taken care of. The output gate adjusts the volume of data conveyed to the output response k_t . The computations happening at the different gates are formulated using the following equations. The process occurring at the input gate is expressed as:

$$f_t = \rho(G_{Sf}S_t + G_{kf}k_{t-1} + G_{wf}w_{t-1} + p_f) \quad (23)$$

wherein ρ indicate the sigmoid function expressed by $\sigma(n) = 1/(1 + e^{-n})$, $p_i, i \in \{f, q, w, o\}$ signifies the bias. G_{ab} exemplifies the weight matrix between the units a and b , for example, G_{Sf} designates the weight matrix existing from the input S to the input gate f_t . The response created by the forget gate is articulated as:

$$q_t = \rho(G_{Sq}S_t + G_{kq}k_{t-1} + G_{wq}w_{t-1} + p_q) \quad (24)$$

Furthermore, the response of the memory cell is modelled as:

$$w_t = q_t \Theta w_{t-1} + f_t \Theta \tanh(G_{Sw}S_t + G_{kw}k_{t-1} + p_w) \quad (25)$$

Here, the linear operator is depicted as Θ , and the computation carried out at the output gate is given by,

$$y_t = \rho(G_{Sy}S_t + G_{ky}k_{t-1} + G_{wy}w_{t-1} + p_y) \quad (26)$$

The DLSTM's output response is quantified at the time instance t by,

$$k_t = y_t \odot \tanh(w_t) \quad (27)$$

The above computations are recursively accomplished to produce the predicted epileptic seizure denoted as, and Figure 2 illustrates the structural depiction of the DLSTM.

3.3.2.1.2 JSCS algorithm for training DLSTM

The big data classification for epileptic seizure prediction using the DLSTM can be improved by utilising the proposed JSCS algorithm. The JSCS algorithm is created by adjusting the JSO (Chou and Molla, 2022) and the SCA (Mirjalili, 2016). The JSO is a population-based approach founded on the behavioural aspects of the jellyfish in the ocean. The jellyfish undergoes two kinds of movement, such as pursuing the ocean current, and motion within the swarm in search of food, like larva, phytoplankton, and eggs of fish. The key benefit of using JSO is that it has a better converging rate, thereby determining the ideal solution in a quick span. The SCA (Mirjalili, 2016) is a metaheuristic algorithmic method formulated utilising the sine and cosine functions to determine solutions to various issues. In this approach, the random creation of the solution is accomplished at first and then the solutions are fluctuated to obtain the ideal solution employing a mathematical system based on cosine and sine functions. The SCA effectively determines the best solutions in constrained and unfamiliar search spaces. The inclusion of SCA in JSO can enhance the exploratory ability and convergence rate. The DLSTM can capture long-term dependencies and handle sequential data well. It is good for time series prediction because of its ability to remember previous inputs for a long time. This makes it well-suited for handling sequences with long-term dependencies. Using DLSTM, time series forecasting models can predict future values based on previous, sequential data.

This provides greater accuracy for demand forecasters. The DLSTM could take inputs with different lengths. The JSCS is used to change the weight and bias of the DLSTM and it reduces the error rate and improves the accuracy of the model. The loss function used here is categorical cross-entropy. The process in which the JSCS algorithm is executed is deliberated using the ensuing steps.

Step 1 Initialise the population

The majority of the optimisation approaches initialise the solutions randomly in the search space, which makes the method prone to local optima and they also suffer from poor convergence rate. Thus to handle such situations, the search agents in this technique are initialised using the logistic map, which is a chaotic map based on the expression given below,

$$N_{x+1} = \alpha N_x (1 - N_x), \quad 0 \leq N_x \leq 1 \quad (28)$$

Here, N_x refers to the logistic value of the position of the jellyfish, with location N_{x+1} , and α designates a parameter with the value set at 4.0.

Step 2 Estimate fitness

After the population members are initialised, their fitness is assessed based on the subsequent expression,

$$Fit = \frac{1}{\zeta} \sum_{i=1}^{\zeta} (O_{\gamma}^* - O_{\gamma})^2 \quad (29)$$

Here, the term ζ signifies the count of training samples, O_{γ}^* and O_{γ} represents the projected and obtained output of the DLSTM.

Step 3 Follow ocean current:

Generally, the ocean currents comprise abundant nutrients and this attracts the jellyfish, and they move in the direction of the ocean current. The vectors from all jellyfish, which are currently in the ideal position are averaged to attain the ocean current, and it is expressed as,

$$\overline{tr} = \frac{1}{\kappa} \sum \overline{tr}_x = N^* - v \times rand(0, 1) \times \varepsilon \quad (30)$$

where ε signifies the mean position of jellyfish, the direction of the ocean current is termed as $\overline{tr}$, N^* denotes the jellyfish in the ideal location, and $v > 0$ exemplifies the distribution coefficient linked with the dimension of the ocean current. The location of the jellyfish can be adjusted based on the ocean current as follows:

$$N_x(u+1) = N_x(u) + rand(0, 1) \times \overline{tr} \quad (31)$$

$$N_x(u+1) = N_x(u) + rand(0, 1) \times (N^* - v \times rand(0, 1) \times \varepsilon) \quad (32)$$

Here, u refers to the iteration count, $N_x(u)$ and $N_x(u+1)$ signifies the location of the jellyfish in the u^{th} and $(u+1)^{\text{th}}$ iterations. The exploration of the search space can be boosted by incorporating SCA in JSO. From SCA (Mirjalili, 2016),

$$N_x(u+1) = \begin{cases} N_x(u) + o_1 \sin(o_2) |o_3 Q_x(u) - N_x(u)|, & o_4 < 0.5 \\ N_x(u) + o_1 \cos(o_2) |o_3 Q_x(u) - N_x(u)|, & o_4 \geq 0.5 \end{cases} \quad (33)$$

where o_i , $i = 1, 2, 3, 4$ are main parameters of the SCA, o_1 specifies the movement direction, o_2 depicts the extent of the movement, o_3 provides arbitrary weights for the destination, and o_4 is an arbitrary number in $[0, 1]$, $||$ explicates the absolute value, and $Q_x(u)$ indicates the locale of the destination. Considering $o_4 < 0.5$, and $o_3 Q_x(u) > N_x(u)$ then equation (33) becomes,

$$N_x(u+1) = N_x(u) + o_1 \sin(o_2) (o_3 Q_x(u) - N_x(u)) \quad (34)$$

$$N_x(u+1) = N_x(u) + o_1 \sin(o_2) \times o_3 Q_x(u) - o_1 \sin(o_2) N_x(u) \quad (35)$$

$$N_x(u+1) = N_x(u) (1 - o_1 \sin(o_2)) + o_1 \sin(o_2) \times o_3 Q_x(u) \quad (36)$$

$$N(u) = \frac{N_x(u+1) - o_1 \sin(o_2) \times o_3 Q_x(u)}{(1 - o_1 \sin(o_2))} \quad (37)$$

Applying equation (37) in equation (32),

$$N_x(u+1) = \left(\frac{N_x(u+1) - o_1 \sin(o_2) \times o_3 Q_x(u)}{(1 - o_1 \sin(o_2))} \right) + rand(0, 1) \times (N^* - v \times rand(0, 1) \times \varepsilon) \quad (38)$$

$$N_x(u+1) - \frac{N_x(u+1)}{(1 - o_1 \sin(o_2))} = \left(\frac{-o_1 \sin(o_2) \times o_3 Q_x(u)}{(1 - o_1 \sin(o_2))} \right) + rand(0, 1) \times (N^* - v \times rand(0, 1) \times \varepsilon) \quad (39)$$

$$\frac{(1 - o_1 \sin(o_2) - 1) N_x(u+1)}{(1 - o_1 \sin(o_2))} = \left(\frac{(1 - o_1 \sin(o_2)) [rand(0, 1) \times (N^* - v \times rand(0, 1) \times \varepsilon)] - o_1 \sin(o_2) \times o_3 Q_x(u)}{(1 - o_1 \sin(o_2))} \right) \quad (40)$$

$$N_x(u+1) = \left(\frac{(1 - o_1 \sin(o_2)) [rand(0, 1) \times (N^* - v \times rand(0, 1) \times \varepsilon)] - o_1 \sin(o_2) \times o_3 Q_x(u)}{(-o_1 \sin(o_2))} \right) \quad (41)$$

$$N_x(u+1) = \left(\frac{(o_1 \sin(o_2) - 1) [rand(0, 1) \times (N^* - v \times rand(0, 1) \times \varepsilon)] + o_1 \sin(o_2) \times o_3 Q_x(u)}{o_1 \sin(o_2)} \right) \quad (42)$$

The location of the x^{th} jellyfish while following the ocean current in the iteration $(u + 1)$ is updated utilising the above expression, thereby attaining high exploratory ability.

Step 4 Swarm movement

The jellyfish demonstrates two kinds of movement in the swarm, such as type A, or passive, and type B, or active. In the preliminary phases, a swarm is created and the members exhibit the passive motion. Over time, the active motion is displayed by the members. The jellyfish moves about their position in the type A motion and the position is updated based on the following equation,

$$N_x(u+1) = N_x(u) + \eta \times rand(0, 1) \times (up - low) \quad (43)$$

Here, up and low denotes the higher and lower limits of the search area and $\eta > 0$ depicts the motion linked with the movement length of the jellyfish about its position.

The type B movement is simulated by considering a random jellyfish t instead of the chosen one for identifying the direction of motion. If the food availability at the x^{th} jellyfish's location is less when compared to that at the t^{th} location, then x^{th} jellyfish travels to the t^{th} location or the other way around. This movement is mathematically modelled as:

$$N_x(u+1) = N_x(u) + rand(0, 1) \times \overline{dir} \quad (44)$$

$$\overrightarrow{dir} = \begin{cases} N_t(u) - N_x(u) & \text{if } Fit(N_x) \geq Fit(N_t) \\ N_x(u) - N_t(u) & \text{if } Fit(N_x) < Fit(N_t) \end{cases} \quad (45)$$

The terms $Fit(N_t)$ and $Fit(N_x)$ stipulates the fitness of the t^{th} jellyfish and x^{th} jellyfish.

Step 5 Time control mechanism

The movement of the jellyfish is also influenced by the variation in the wind or temperature and because of this; a new swarm may be created. In the new swarm, both type A and B movements will be displayed by the jellyfish. This can be simulated considering a time control mechanism for movement between the two phases, as well as inside the swarm, and it is formulated as,

$$W(u) = \left| \left(1 - \frac{u}{u_{\max}} \right) * (2 * rand(0, 1) - 1) \right| \quad (46)$$

wherein, $u_{\max}$ terms the maximum iterations and $W(u)$ designates the time control strategy. In addition to this, another term W_0 is taken into account and if $W(u) \geq W_0$ the jellyfish pursue the ocean current, else they move within the swarm. The term $(1 - W(u))$ is used to simulate the movement inside the swarm and if $(1 - W(u)) > rand(0, 1)$ then type B movement is displayed, else type A is showcased.

Step 6 Boundary conditions

Around the globe, oceans can be found everywhere, and as the earth is almost sphere-shaped, the jellyfish moving out of the search space will pass in the opposite bound, and this is depicted as follows:

$$N'_{x,K} = \begin{cases} (N_{x,K} - P_K) + T_K & \text{if } N_{x,K} > P_K \\ (N_{x,K} - T_K) + P_K & \text{if } N_{x,K} < T_K \end{cases} \quad (47)$$

Here, $N_{x,K}$ and $N'_{x,k}$ symbolises the position of the x^{th} jellyfish in the K^{th} dimension before and after applying boundary constraints, P_K and T_K signifies the maximal and minimal bounds in the K^{th} dimensionality of the search space.

Step 7 Evaluate feasibility

The jellyfish's position is adjusted and then, the fitness is measured to find the ideal location using equation (29), wherein the ideal solution corresponds to the low fitness.

Step 8 Termination

The steps deliberated above are executed persistently till u grasps their maximum value $u_{\max}$ and the pseudocode of the JSCS approach is illustrated in Algorithm 1.

The JSCS fashioned by amalgamating the SCA in JSO effectively enhanced the exploratory ability with fast convergence. Thus, the best hyperparameters of the DLSTM

are computed in a very short time, thereby improving the efficacy of epileptic seizure prediction.

Algorithm 1 Pseudocode of JSCS

Begin

Input: Maximum iterations $u_{\max}$, Population size κ

Initiate population with equation (28)

Calculate the objective with equation (29)

Recognise the best jellyfish N^*

For $u = 1: u_{\max}$

For $x = 1: \kappa$

Determine $W(u)$ with equation (46)

If $W(u) \geq W_0$

Update the location using equation (42)

else

If $(1 - W(u)) < rand(0, 1)$

Update the location using equation (43)

else

Update the location using equation (44)

endif

endif

Examine boundary constrictions

Calculate the objective with equation (29)

Change the ideal jellyfish N^* and location of the jellyfish N_x

Endfor

Endfor

Output: Ideal jellyfish

End

4 Results and discussion

The JSCS-DLSTM developed for big data classification for epileptic seizure prediction is analysed for its supremacy in this segment. Further, the experimental setup, metrics, and the algorithm proposed are also examined elaborately.

4.1 Experimental set-up

The big data classification for epileptic seizure prediction using the proposed JSCS-DLSTM is realised by carrying out the implementation of the technique with Python language. Table 1 show the hyperparameters used in the proposed model.

Table 1 Hyperparameters of the methods

<i>Methods/parameters</i>	<i>NoC-CNN</i>	<i>2L-LSTM</i>	<i>Temporal autoencoder</i>	<i>2D CNN</i>	<i>Proposed method</i>
Hidden state neurons	10	10	10	10	10
Epoch	20	20	20	25	30
Batch size	32	64	32	32	64
Learning rate	0.001	0.001	0.001	0.001	0.001

4.2 Dataset description

The EEG big data acquired in this experiment is acquired from the TUEP EEG Resources dataset (https://www.isip.piconepress.com/projects/tuh_eeg/html) and the CHB-MIT Scalp EEG Dataset (2010).

4.2.1 TUH EEG Epilepsy Corpus (TUEP)

The TUEP [TUH EEG Epilepsy Corpus (TUEP), xxxx] dataset is a sub-class of the TUH EEG Corpus (TUEG) dataset, which comprises the EEG recordings taken from the Temple University Hospital (TUH), Philadelphia. This dataset includes more than 1,600 variable-length EEG recordings taken from more than 100 healthy and 100 epileptic individuals of different ages ranging from (17–88) years including both men and women. The TUEP contains more than 400 hours of EEG recordings and is considered to be the most diverse and largest collection of publicly available EEG epileptic signals.

4.2.2 CHB-MIT Scalp EEG dataset

This dataset (CHB-MIT Scalp EEG Dataset, 2010) encompasses EEG signals recorded from paediatric patients with intractable seizures. It comprises scalp EEG recordings for 969 hours with 173 seizures taken from 23 subjects. Different kinds of seizures, such as tonic, atonic, and clonic are included in the database. The paediatric subjects include children aged from 10–22 years including both sexes. The patients were observed for numerous days after cutting back on anti-seizure medication to characterise their seizures and assessing their candidature for surgical intervention.

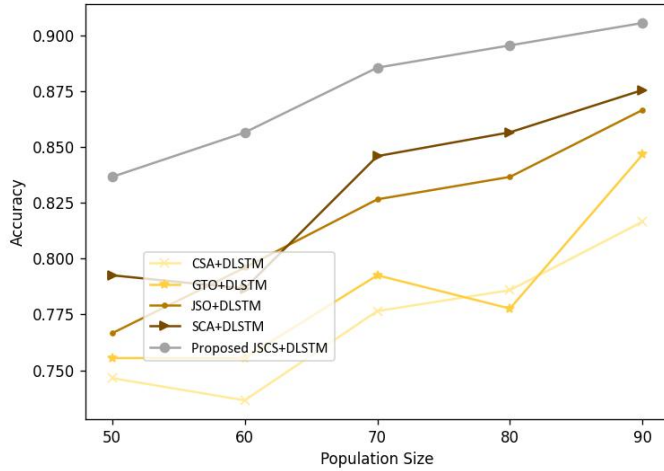
4.3 Algorithmic methods

The efficiency of the JSCS devised in this research is investigated for its supremacy based on three performance metrics concerning available algorithms, such as crow search algorithm (CSA)+DLSTM (Askarzadehv, 2016; Zhu et al., 2016), GorillaTroopsOptimiser (GTO)+DLSTM (Xiao et al., 2022; Zhu et al., 2016), JSO+DLSTM (Chou and Molla, 2022; Zhu et al., 2016), and SCA+DLSTM (Mirjalili, 2016; Zhu et al., 2016).

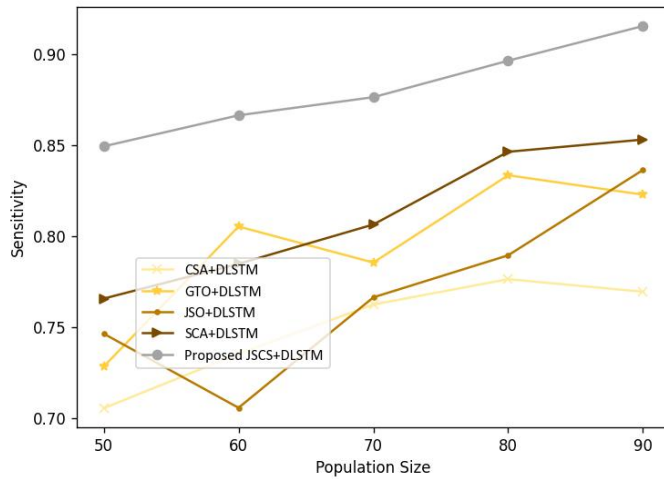
4.4 Algorithmic assessment

The JSCS algorithm is introduced in this research for adjusting the hyperparameters of the DLSTM, and the superiority of this method is investigated based on the data accumulated from the TUEP dataset and CHB-MIT Scalp EEG dataset based on different solution sizes.

Figure 3 Algorithmic assessment of the JSCS-DLSTM, (a) accuracy, (b) sensitivity, (c) specificity with the TUEP dataset (see online version for colours)

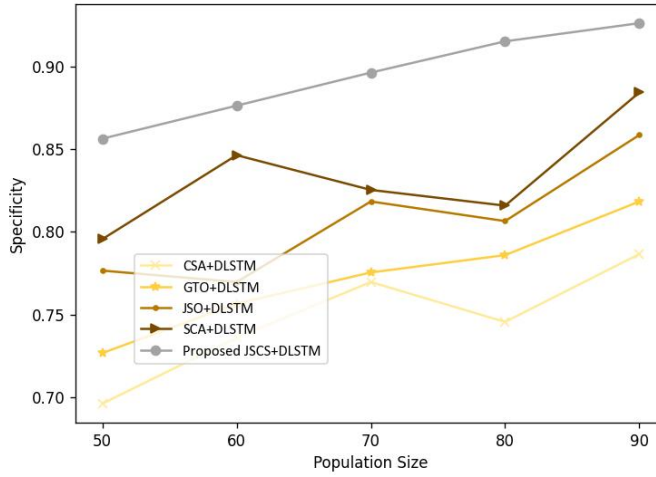


(a)



(b)

Figure 3 Algorithmic assessment of the JSCS-DLSTM, (a) accuracy, (b) sensitivity, (c) specificity with the TUEP dataset (continued) (see online version for colours)



(c)

4.4.1 Using the TUEP dataset

The investigation of the JSCS+DLSTM using the data accumulated from the TUEP dataset is explicated in Figure 3. The examination concerning the accuracy of the JSCS+DLSTM is demonstrated in Figure 3(a). With a solution size of 80, the accuracy figured by the CSA+ DLSTM, GTO+DLSTM, JSO+DLSTM, SCA+DLSTM, and JSCS+DLSTM is 0.786, 0.778, 0.837, 0.857, and 0.895, respectively. Figure 3(b) depicts the examination of the JSCS+DLSTM considering sensitivity. The sensitivity recorded by JSCS+DLSTM is 0.897 with a solution size of 80, while the available algorithmic methods, such as quantified sensitivity values of 0.777 for CSA+ DLSTM, 0.834 for GTO+DLSTM, 0.790 for JSO+DLSTM, and 0.847 for SCA+DLSTM. The JSCS+DLSTM is analysed considering specificity, and this is exhibited in Figure 3(c). The specificity quantified by CSA+ DLSTM is 0.745, GTO+DLSTM is 0.786, JSO+DLSTM is 0.807, SCA+DLSTM is 0.816, and JSCS+DLSTM is 0.916 for solution size of 80.

4.4.2 With CHB-MIT Scalp EEG dataset

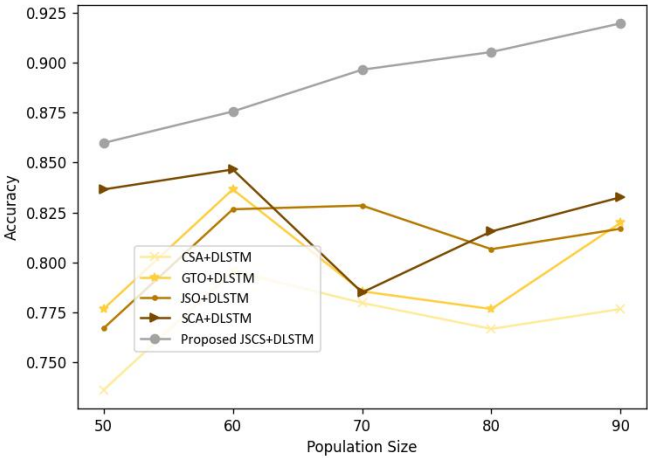
The scrutiny of the JSCS+DLSTM considering the EEG data obtained from the CHB-MIT Scalp EEG dataset is presented. In Figure 4(a), the assessment of the JSCS+DLSTM depending on accuracy is presented. The accuracy computed by JSCS+DLSTM is 0.920, which is higher than the accuracy figured by CSA+ DLSTM at 0.777, GTO+DLSTM at 0.820, JSO+DLSTM at 0.817, and SCA+DLSTM at 0.833 for solution size of 90. Further, the scrutiny of the JSCS+DLSTM founded on sensitivity is illustrated in Figure 4(b). With a solution size of 90, the prevailing approaches, such as CSA+ DLSTM, GTO+DLSTM, JSO+DLSTM, and SCA+DLSTM measured sensitivity of 0.830, 0.805, 0.780, and 0.837, respectively. However, an improved sensitivity of 0.915 is attained by the JSCS+DLSTM. Figure 4(c) exhibits the investigation of the

JSCS+DLSTM concerning specificity. The specificity figured with a solution size of 90 by the CSA+ DLSTM, GTO+DLSTM, JSO+DLSTM, SCA+DLSTM, and JSCS+DLSTM is 0.817, 0.837, 0.830, 0.857, and 0.930, correspondingly.

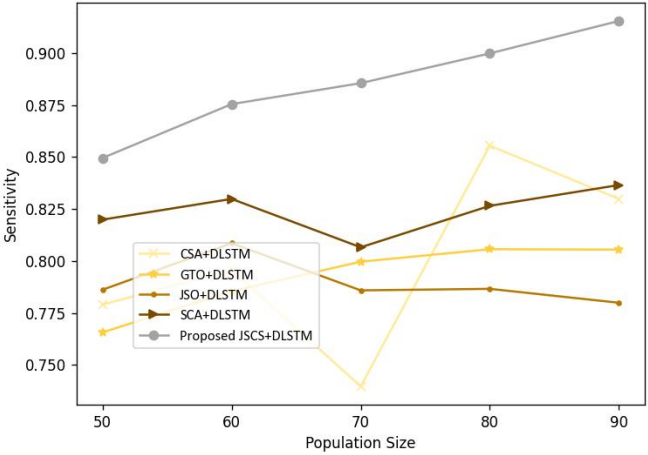
4.5 Comparative techniques

A novel big data classification technique based on epileptic seizure data is proposed in this research using JSCS-DLSTM with MapReduce framework. The efficiency of the JSCS-DLSTM is inspected in comparison to NoC-CNN (Ghosh et al., 2021), 2L-LSTM (Singh and Malhotra, 2022), temporal autoencoder (Nejedly et al., 2023), and 2D CNN (Wei et al., 2022). The JSCS-DLSTM is examined for its competence in big data classification of epileptic seizure data considering the EEG data acquired from the TUEP dataset and CHB-MIT Scalp EEG dataset considering various values of training data.

Figure 4 Algorithmic assessment of the JSCS-DLSTM, (a) accuracy, (b) sensitivity, (c) specificity with CHB-MIT Scalp EEG dataset (see online version for colours)

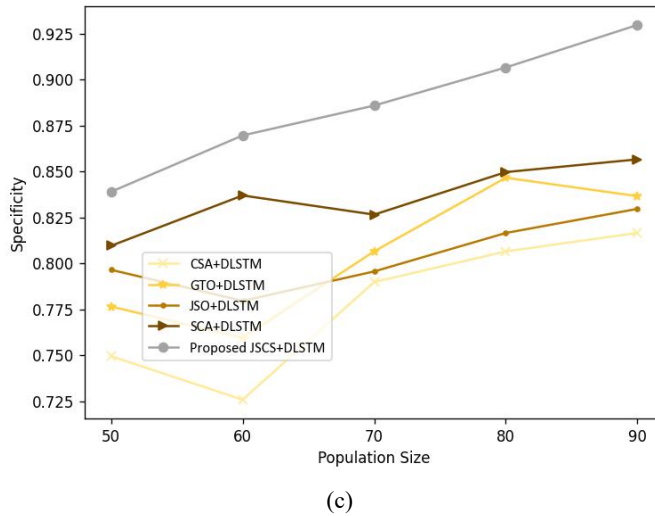


(a)



(b)

Figure 4 Algorithmic assessment of the JSCS-DLSTM, (a) accuracy, (b) sensitivity, (c) specificity with CHB-MIT Scalp EEG dataset (continued) (see online version for colours)



4.5.1 Using the TUEP dataset

This segment explicates the analysis of the JSCS-DLSTM considering the data from the TUEP dataset. Figure 5(a) demonstrates the examination of the JSCS-DLSTM based on accuracy. The JSCS-DLSTM quantified an accuracy value of 0.877, with 70% training data, which is better by 13.35%, 9.23%, 4.59%, and 2.48% than the values attained by NoC-CNN, 2L-LSTM, Temporal autoencoder, and 2D CNN at 0.760, 0.796, 0.836, and 0.855. Further, in Figure 5(b), the investigation based on the sensitivity of the JSCS-DLSTM is explicated. With 70% of training data, the NoC-CNN, 2L-LSTM, temporal autoencoder, 2D CNN, and JSCS-DLSTM measured sensitivity of 0.777, 0.755, 0.835, 0.853, and 0.886, respectively. This reveals that the JSCS-DLSTM has produced a performance enhancement of 12.38%, 14.76%, 5.74%, and 3.80% in comparison to the other methods. The evaluation of the JSCS-DLSTM contingent on specificity is portrayed in Figure 5(c). The specificity figured by the prevailing epileptic seizure detection methods is 0.770 for NoC-CNN, 0.777 for 2L-LSTM, 0.796 for temporal autoencoder, and 0.837 for 2D CNN with 70% training data. However, a higher value of specificity at 0.896 is quantified by the JSCS-DLSTM, which is better by 14.05%, 13.29%, 11.16%, and 6.59%.

4.5.2 With CHB-MIT Scalp EEG dataset

The evaluation of the JSCS-DLSTM based on the information gathered from the CHB-MIT Scalp EEG dataset is presented here considering various evaluation metrics. At first, the examination of the JSCS-DLSTM is executed considering accuracy, which is represented in Figure 6(a). The accuracy figured by the NoC-CNN, 2L-LSTM, temporal autoencoder, 2D CNN, and JSCS-DLSTM are 0.789, 0.794, 0.827, 0.817, and 0.893, respectively, for 80% training data. This shows that the JSCS-DLSTM attained a higher

accuracy by 11.69%, 11.11%, 7.43%, and 8.55% than the other epileptic seizure methods. Figure 6(b) demonstrates the sensitivity-based investigation of the JSCS-DLSTM. The value of sensitivity measured by the existing epileptic seizure detection techniques is 0.787 for NoC-CNN, 0.827 for 2L-LSTM, 0.827 for temporal autoencoder, and 0.806 for 2D CNN. The JSCS-DLSTM is found to attain a higher value of sensitivity of 0.905, which is better by 13.13%, 8.71%, 8.71%, and 11.03%. In Figure 6(c), the evaluation of the JSCS-DLSTM in view of specificity is shown. The JSCS-DLSTM figured a specificity of 0.919 with 80% training data, which is higher than the specificity of NoC-CNN at 0.794, 2L-LSTM at 0.760, temporal autoencoder at 0.814, and 2D CNN at 0.817 by 13.63%, 17.33%, 11.46%, and 11.14%.

Figure 5 Comparative analysis of the JSCS-DLSTM, (a) accuracy, (b) sensitivity, (c) specificity with TUEP dataset (see online version for colours)

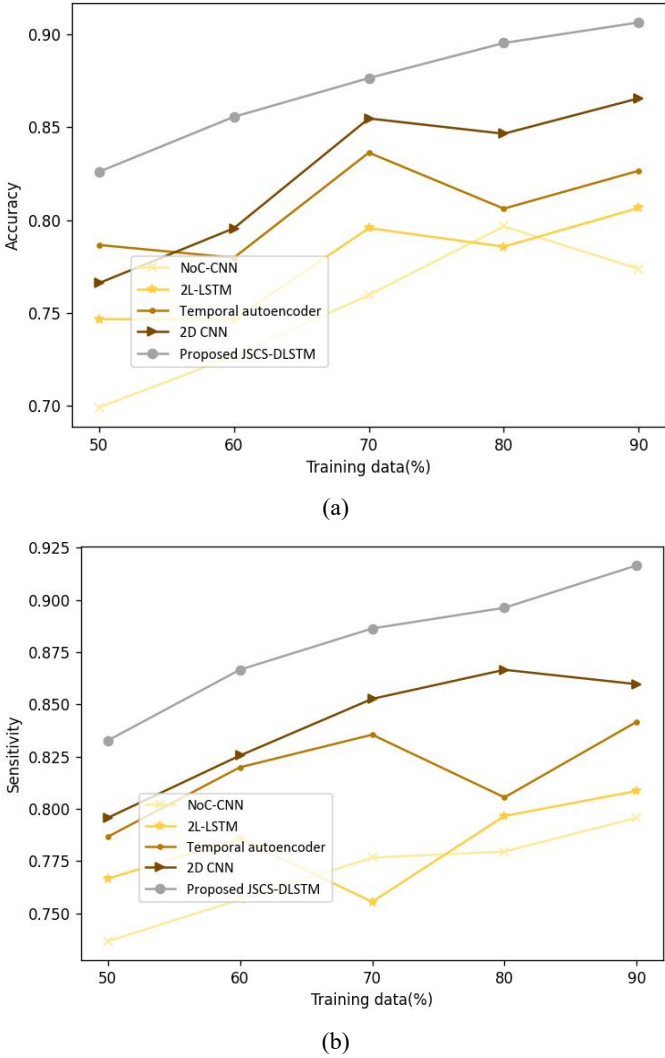
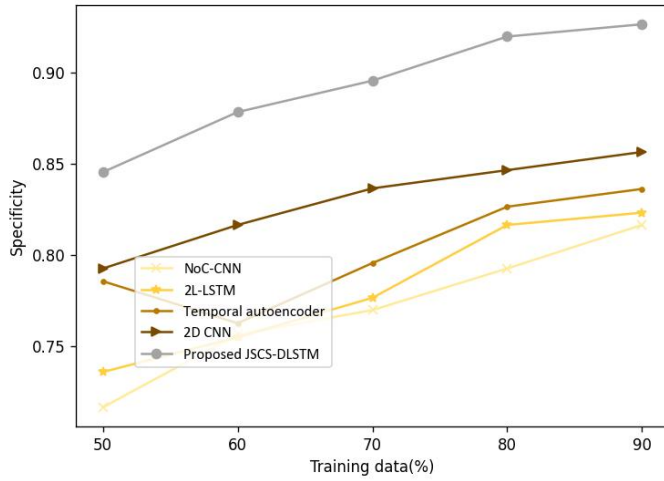


Figure 5 Comparative analysis of the JSCS-DLSTM, (a) accuracy, (b) sensitivity, (c) specificity with TUEP dataset (continued) (see online version for colours)



(c)

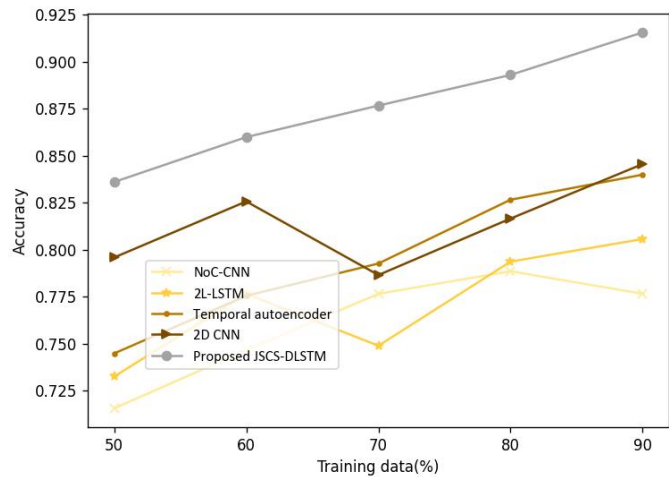
4.6 Comparative discussion

The JSCS-DLSTM developed for big data classification on epileptic seizure data is compared with the other epileptic seizure prediction works to reveal the competence of the proposed method in predicting seizures. The comparison is illustrated in Table 2 with the values of the evaluation measures corresponding to 90% of training data taken from the TUEP and the CHB-MIT Scalp EEG dataset. It can be deduced that the JSCS-DLSTM attained superior values of accuracy at 0.915, sensitivity at 0.919, and specificity at 0.927 when considering the data from the CHB-MIT Scalp EEG dataset. The high value of metrics is accounted for by the usage of DEC for partitioning the data and the JSCS-DLSTM in prediction, as the DLSTM effectively captures the long-term dependencies thereby predicting the epileptic and normal samples correctly. Further, the accuracy attained by the NoC-CNN is 0.777, 2L-LSTM is 0.806, Temporal autoencoder is 0.840, and 2D CNN is 0.845. Likewise, a low value of sensitivity at 0.799, 0.804, 0.837, and 0.827 is measured by NoC-CNN, 2L-LSTM, Temporal autoencoder, and 2D CNN, respectively. The specificity quantified by the prevailing methods is 0.783 for NoC-CNN, 0.807 for 2L-LSTM, 0.826 for temporal autoencoder, and 0.829 for 2D CNN, which reveals the superiority of the JSCS-DLSTM. The proposed model uses the pre-processing technique for improving the quality of the data. This improves the accuracy and reliability of the proposed method. The DEC method improves the performance, reduces storage requirements, and increases the scalability of the model by allowing parallel processing. The ineffective features are removed by the feature selection technique and it reduces the training time. The proposed model is analysed by slicing the dataset into training and testing sets and optimising the parameters to reduce the computation time and enhance the classification accuracy. Feature extraction reduces the amount of data without losing vital information. It increases the efficiency and accuracy of ML models.

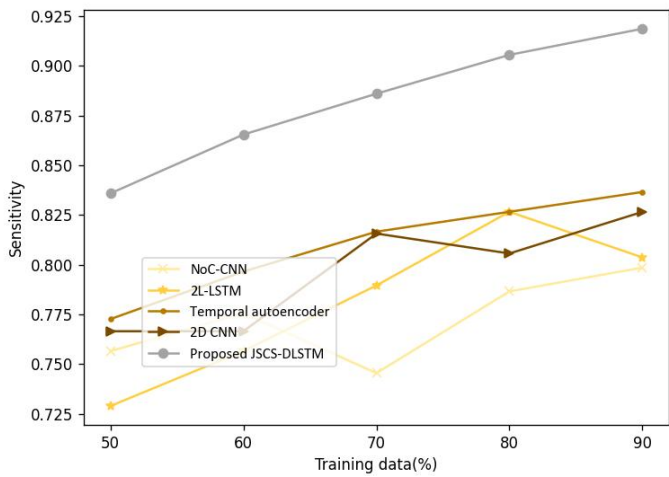
Table 2 Comparative discussion of the JSCS-DLSTM

Dataset	Metrics	NoC-CNN	2L-LSTM	Temporal autoencoder	2D CNN	Proposed JSCS-DLSTM
TUEP	Accuracy	0.774	0.807	0.827	0.866	0.907
	Sensitivity	0.796	0.809	0.842	0.860	0.917
	Specificity	0.817	0.823	0.836	0.857	0.927
CHB-MIT scalp EEG	Accuracy	0.777	0.806	0.840	0.845	0.915
	Sensitivity	0.799	0.804	0.837	0.827	0.919
	Specificity	0.783	0.807	0.826	0.829	0.927

Figure 6 Comparative analysis of the JSCS-DLSTM, (a) accuracy, (b) sensitivity, (c) specificity with CHB-MIT Scalp EEG dataset (see online version for colours)

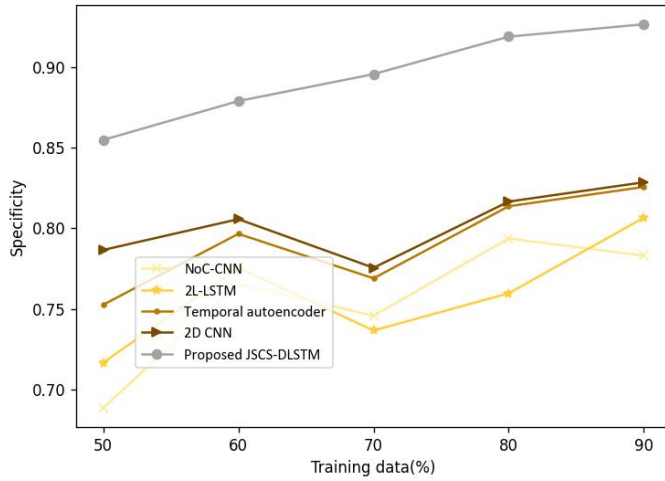


(a)



(b)

Figure 6 Comparative analysis of the JSCS-DLSTM, (a) accuracy, (b) sensitivity, (c) specificity with CHB-MIT Scalp EEG dataset (continued) (see online version for colours)



(c)

5 Conclusions

An efficient big data classification technique for epileptic seizure data is presented in this research, wherein a hybrid optimisation-based DL method, named JSCS-DLSTM is developed. The input big data is partitioned first using DEC and is applied to a MapReduce framework for handling the big data. The MapReduce framework comprises a mapper and reducer phase, wherein the mapper phase, the partitioned data is subjected to each mapper, where it is pre-processed using the Kalman filter, and several statistical and spectral features are determined. Thereafter, WED and Bhattacharya distance is used to select the most relevant features. The selected features are progressed to the reducer phase, where they are aggregated and applied to the DLSTM for epileptic seizure prediction. Here, a novel JSCS algorithm created by combining the JSO and SCA is used to train the DLSTM. Further, the evaluation of the JSCS-DLSTM reveals that superior values of accuracy at 0.915, sensitivity at 0.919, and specificity at 0.927 are attained. In the future, hybrid DL methods with CNNs will be developed to provide higher-order data analytics, thereby improving the detection accuracy.

References

- Askarzadeh, A. (2016) 'A novel metaheuristic method for solving constrained engineering optimization problems: crow search algorithm', *Computers & Structures*, Vol. 169, pp.1–12.
- Bablani, A., Edla, D.R. and Dodia, S. (2018) 'Classification of EEG data using the k-nearest neighbor approach for concealed information test', *Procedia Computer Science*, Vol. 143, pp.242–249.
- Bhattacharya, A. (2022) *Automatic Seizure Prediction using CNN and LSTM*, arXiv preprint arXiv:2211.02679.

- Cha, S.H. (2007) 'Comprehensive survey on distance/similarity measures between probability density functions', *City*, Vol. 1, No. 2, p.1.
- CHB-MIT Scalp EEG Dataset (2010) 9 June [online] <https://physionet.org/content/chbmit/1.0.0/> (accessed 3 March 2023).
- Chou, J.S. and Molla, A. (2022) 'Recent advances in use of bio-inspired jellyfish search algorithm for solving optimization problems', *Scientific Reports*, Vol. 12, No. 1, p.19157.
- Chung, Y.G., Jeon, Y., Yoo, S., Kim, H. and Hwang, H. (2022) 'Big data analysis and artificial intelligence in epilepsy – common data model analysis and machine learning-based seizure detection and forecasting', *Clinical and Experimental Pediatrics*, Vol. 65, No. 6, p.272.
- D'Alessandro, M., Esteller, R., Vachtsevanos, G., Hinson, A., Echauz, J. and Litt, B. (2003) 'Epileptic seizure prediction using hybrid feature selection over multiple intracranial EEG electrode contacts: a report of four patients', *IEEE Transactions on Biomedical Engineering*, Vol. 50, No. 5, pp.603–615.
- Diyykh, M., Li, Y. and Wen, P. (2016) 'EEG sleep stages classification based on time domain features and structural graph similarity', *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 24, No. 11, pp.1159–1168.
- Emara, H.M., Elwekeil, M., Taha, T.E., El-Fishawy, A.S., El-Rabaie, E.S.M., El-Shafai, W., El Banby, G.M., Alotaiby, T., Alshebeili, S.A. and Abd El-Samie, F.E. (2022) 'Efficient frameworks for EEG epileptic seizure detection and prediction', *Annals of Data Science*, Vol. 9, No. 2, pp.393–428.
- Faust, O., Acharya, U.R., Adeli, H. and Adeli, A. (2015) 'Wavelet-based EEG processing for computer-aided seizure detection and epilepsy diagnosis', *Seizure*, Vol. 26, pp.56–64.
- Fox, V., Hightower, J., Liao, L., Schulz, D. and Borriello, G. (2003) 'Bayesian filtering for location estimation', *IEEE Pervasive Computing*, Vol. 2, No. 3, pp.24–33.
- Ghosh, A., Roy, A.P., Patra, R. and Mondal, H.K. (2021) 'Designing efficient NoC-based neural network architectures for identification of epileptic seizure', *SN Computer Science*, Vol. 2, No. 363, pp.1–12.
- Hong, K.S., Khan, M.J. and Hong, M.J. (2018) 'Feature extraction and classification methods for hybrid fNIRS-EEG brain-computer interfaces', *Frontiers in Human Neuroscience*, Vol. 12, p.246.
- Hussein, R., Ahmed, M.O., Ward, R., Wang, Z.J., Kuhlmann, L. and Guo, Y. (2019) *Human Intracranial EEG Quantitative Analysis and Automatic Feature Learning for Epileptic Seizure Prediction*, arXiv preprint arXiv:1904.03603.
- Ke, H., Chen, D., Li, X., Tang, Y., Shah, T. and Ranjan, R. (2018) 'Towards brain big data classification: Epileptic EEG identification with a lightweight VGGNet on global MIC', *IEEE Access*, Vol. 6, pp.14722–14733.
- Kiral-Kornek, I., Roy, S., Nurse, E., Mashford, B., Karoly, P., Carroll, T., Payne, D., Saha, S., Baldassano, S., O'Brien, T. and Grayden, D. (2018) 'Epileptic seizure prediction using big data and deep learning: toward a mobile system', *EBioMedicine*, Vol. 27, No. C, pp.103–111.
- Mannepalli, K., Sastry, P.N. and Suman, M. (2017) 'A novel adaptive fractional deep belief networks for speaker emotion recognition', *Alexandria Engineering Journal*, Vol. 56, No. 4, pp.485–497.
- Mirjalili, S. (2016) 'SCA: a sine cosine algorithm for solving optimization problems', *Knowledge-based Systems*, Vol. 96, pp.120–133.
- Murugappan, M. and Murugappan, S. (2013) 'Human emotion recognition through short time Electroencephalogram (EEG) signals using fast Fourier transform (FFT)', in *Proceedings of 2013 IEEE 9th International Colloquium on Signal Processing and its Applications*, IEEE, March, pp.289–294.
- Nejedly, P., Kremen, V., Lepkova, K., Mivalt, F., Sladky, V., Pridalova, T., Plesinger, F., Jurak, P., Pail, M., Brazdil, M. and Klimes, P. (2023) 'Utilization of temporal autoencoder for semi-supervised intracranial EEG clustering and classification', *Scientific Reports*, Vol. 13, No. 1, p.744.

- Sathyaraj, R., Ramanathan, L., Lavanya, K. and Balasubramanian, V. (2020) ‘Chicken swarm foraging algorithm for big data classification using the deep belief network classifier’, *Data Technologies and Applications*, Vol. 55, No. 3, pp.332–352.
- Singh, K. and Malhotra, J. (2022) ‘Two-layer LSTM network-based prediction of epileptic seizures using EEG spectral features’, *Complex & Intelligent Systems*, Vol. 8, No. 3, pp.2405–2418.
- TUH EEG Epilepsy Corpus (TUEP) [online] https://www.isip.piconepress.com/projects/tuh_eeg/html (accessed March 2023).
- Usman, S.M., Khalid, S., Akhtar, R., Bortolotto, Z., Bashir, Z. and Qiu, H. (2019) ‘Using scalp EEG and intracranial EEG signals for predicting epileptic seizures: review of available methodologies’, *Seizure*, Vol. 71, pp.258–269.
- Wei, X., Zhang, Z. and Zhou, Y. (2022) ‘Epileptic seizure detection from multivariate sequential signals using multidimensional convolution network’, *Neurological Disorders*, Vol. 10, No. 10, pp.1–20.
- Wei, Z., Zou, J., Zhang, J. and Xu, J. (2019) ‘Automatic epileptic EEG detection using convolutional neural network with improvements in time-domain’, *Biomedical Signal Processing and Control*, Vol. 53, No. 3, p.101551.
- Xiao, Y., Sun, X., Guo, Y., Li, S., Zhang, Y. and Wang, Y. (2022) ‘An improved gorilla troops optimizer based on lens opposition-based learning and adaptive β -Hill climbing for global optimization’, *Comput. Model. Eng. Sci.*, Vol. 131, No. 2, pp.815–850.
- Xie, J., Girshick, R. and Farhadi, A. (2016) ‘Unsupervised deep embedding for clustering analysis’, in *Proceedings of International Conference on Machine Learning*, PMLR, June, pp.478–487.
- Zhang, Y., Zhou, W., Yuan, S. and Yuan, Q. (2015) ‘Seizure detection method based on fractal dimension and gradient boosting’, *Epilepsy & Behavior*, Vol. 43, pp.30–38.
- Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L. and Xie, X. (2016) ‘Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks’, in *Proceedings of the AAAI Conference on Artificial Intelligence*, March, Vol. 30, No. 1.