IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 08, Issue 6 (June. 2018), ||V (I) || PP 27-40

Gaussian Distributive Stochastic Neighbor Embedding Based Feature Extraction for Medical Data Diagnosis

Nithya C[1], Saravanan V[2]
[1]*PG & Research Department Of M.Sc (IT), Hindusthan College Of Arts And Science, Coimbatore, India.*
[2]*PG & Research Department Of M.Sc (IT), Hindusthan College Of Arts And Science, Coimbatore, India.*
Corresponding Author: Nithya C

Abstract:-Feature extraction is a key process to reduce the dimensionality of medical dataset for efficient disease prediction. The feature extraction technique removes irrelevant features to acquire higher prediction accuracy during disease diagnosis. Few research works are developed to extract the relevant features from dataset using different data mining techniques. But, performance of conventional feature extraction technique was not efficient which reduces the accuracy of disease prediction and also feature extraction time was more. In order to solve these limitations, a Gaussian Distributive Stochastic Neighbor Embedding Feature Extraction (GDSNE-FE) technique is proposed. The GDSNE-FE technique is developed for dimensionality reduction of medical dataset with minimal time and space complexity. The GDSNE-FE technique comprises of two main steps. Initially, GDSNE-FE technique employs Gaussian probability distribution that finds the relationship between features in high dimensional space with help of distance metric. After determining the probability distribution, relevant features have high probability of being picked and irrelevant features have extremely lesser probability of being picked. From that, GDSNE-FE technique extracts only relevant features from medical dataset. After extracting relevant features, GDSNE-FE technique used Jensen–Shannon divergence to evaluate similarity between two feature probability distributions. This helps for GDSNE-FE technique to extracts the more significant features for accurate disease prediction with higher accuracy and minimum time. The GDSNE-FE technique conducts experimental evaluation on factors such as feature extraction accuracy, feature extraction time, true positive rate, false positive rate and space complexity with respect to number of features. The experimental result shows that the GDSNE-FE technique is able to improve the true positive rate and also reduces the time of feature extraction when compared to state-of-the-art-works.

Keywords: -Dimensionality Reduction, Feature Extraction Gaussian Distributive Stochastic Neighbor Embedding, Probability Distribution, High Dimensional Space Jensen–Shannon Divergence

---

Date of Submission: 28-05-2018 Date of acceptance: 10-06-2018

I. INTRODUCTION

In data mining, dimensional reduction reduces the curse of dimensionality and eliminates unrelated attributes in high-dimensional space for efficient disease prediction. The dimensional reduction plays key role in much application as it improves the classification and prediction performance of disease diagnosis. A general problem faced by a data scientists is dealing the very high-dimensional data (i.e. the medical dataset consisting of lots of features) for disease diagnosis. The algorithms designed for disease classification and prediction works well for low-dimensional datasets but does not work well when medical dataset have hundreds of features as the curse of dimensionality. In order to reduce the dimensionality of medical dataset and thereby achieving higher disease diagnosis accuracy, GDSNE-FE technique is developed. The GDSNE-FE reduces the amount of dimensions through extracting the relevant features from medical dataset.

Recently, many research works have been designed for feature extraction from medical dataset. Recursive Feature Elimination was presented in [1] to extract EEG Features for epileptic seizure prediction. However, feature extraction performance was not efficient. An Orthogonal Feature Extraction (OFE) model was presented in [2] to increase the performance of feature extraction for cancer prediction accuracy. But, the amount of time required for feature extraction was not reduced.

An ensemble scheme was introduced in [3] for cancer diagnosis through classification with three stages. But, the prediction accuracy was not improved using ensemble scheme. Survival analysis for high-dimensional medical data was presented in [4] to analyze feature extraction performance. However, the space complexity was remained unaddressed.

Class-wise feature extraction technique was introduced in [5] for choosing features from multimodal data. But, the feature selection was not carried out in efficient manner. A Jointly Sparse Discriminant Analysis

(JSDA) was designed in [6] to find the key features of breast cancer for disease diagnosis. The JSDA takes more computational time for feature extraction.

A novel technique was designed in [7] to extract features for risk prediction of disease. But, this technique does not provide optimal features. A Computer Aided Diagnosis (CAD) system was presented in [8] to extract effective features for diagnosis of Alzheimer's disease. However, true positive rate using CAD system was poor.

A novel Feature Extraction Methods was developed in [9] to diagnosis alzheimer's disease with higher true positive rate. But, computational time taken for feature extraction was higher. A genetic algorithm based feature selection was intended in [10] to perform Coronary Artery Diseases diagnosis. However, feature selection performance was not effectual.

In order to solve the above mentioned issues, Gaussian Distributive Stochastic Neighbor Embedding Feature Extraction (GDSNE-FE) technique is developed. The contributions of GDSNE-FE technique is formulated as follows,

I. To improve the performance of dimensionality reduction of medical data through feature extraction for efficient disease prediction, Gaussian Distributive Stochastic Neighbor Embedding Feature Extraction (GDSNE-FE) technique is designed.
II. To extract only the features that are relevant to disease diagnosis from medical dataset with higher accuracy, Gaussian Probability Distribution applied in proposed GDSNE-FE technique on the contrary to existing scholastic neighbour embedding method.
III. To extract most importance features for efficient disease prediction at an early stage with minimum time, Jensen–Shannon divergence is used in proposed GDSNE-FE technique on the contrary to conventional scholastic neighbour embedding method.

The rest of paper structure is formulated as follows: In Section 2, Gaussian Distributive Stochastic Neighbor Embedding Feature Extraction (GDSNE-FE) technique is explained with assists of architecture diagram. In Section 3, Simulation settings are described and the result discussion is presented in Section 4. Section 5 reviews the related works. Section 6 presents the conclusion of the paper.

II. GAUSSIAN DISTRIBUTIVE STOCHASTIC NEIGHBOR EMBEDDING FEATURE EXTRACTION TECHNIQUE

Feature extraction is a significant process that is employed in machine learning problems. The feature extraction techniques project the data into a low dimensional space to highlight some important features of the scattered data. Selecting significant features for disease prediction is a very challenging process because a lot of features are irrelevant or redundant. The feature extraction technique filters irrelevant features in order to achieve higher prediction accuracy for disease diagnosis. Various models and methods are designed to extract significant features from medical data set. However, the performance of conventional feature extraction technique was not sufficient which impacts the accuracy of disease prediction at an early stage. Besides, the time complexity involved during feature extraction was higher. In order to overcome such limitations, a Gaussian Distributive Stochastic Neighbor Embedding Feature Extraction (GDSNE-FE) technique is introduced.

The objective of GDSNE-FE technique is to achieve dimensionality reduction through extracting significant features from medical dataset for disease diagnosis. The GDSNE-FE technique is a non-linear dimensionality reduction algorithm which mainly well-suited for embedding high-dimensional data into low dimensional data. Therefore, Gaussian Distributive Stochastic Neighbor Embedding is used in proposed GDSNE-FE technique on the contrary to conventional feature extractions techniques.

The GDSNE-FE technique is designed from standard Stochastic Neighbor Embedding with help of Gaussian probability distribution and Jensen–Shannon divergence. On the contrary to existing distributive stochastic neighbor embedding, the proposed GDSNE applies Gaussian probability distribution and Jensen–Shannon divergence which helps to extract more significant features from medical dataset for predicting occurrence of diseases at an early stage. The architecture diagram of Gaussian Distributive Stochastic Neighbor Embedding Feature Extraction (MDSNE-FE) technique is shown in below Figure 1.

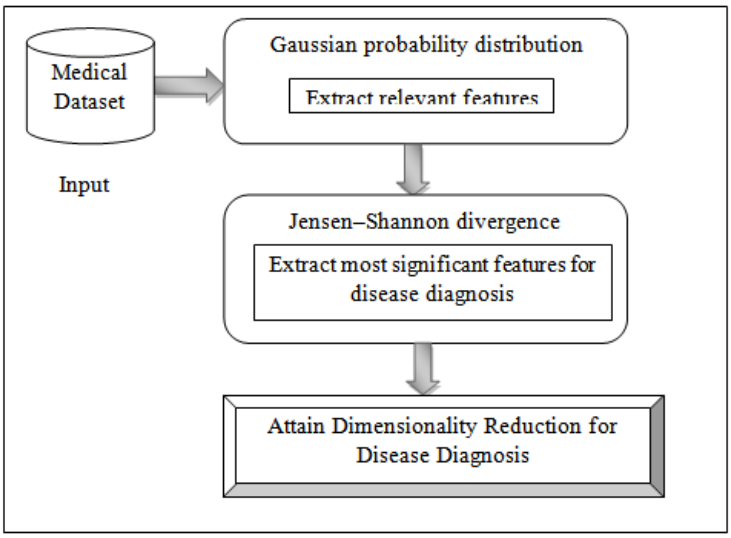

Figure 1 Architecture Diagram of Gaussian Distributive Stochastic Neighbor Embedding Feature Extraction (GDSNE-FE) technique For Disease Diagnosis

Figure 1 demonstrates the process of Gaussian Distributive Stochastic Neighbor Embedding Feature Extraction (GDSNE-FE) technique for efficient disease prediction from medical data. As depicted in figure 1, the GDSNE-FE technique comprises two processes as follows. The GDSNE-FE technique takes medical dataset as input. After taking medical dataset, the GDSNE-FE technique initially applies Gaussian probability distribution that evaluates relationship between features in high dimensional space using distance metrics. With help of Gaussian probability distribution, the GDSNE-FE technique extracts the relevant features from medical dataset in order to perform disease diagnosis. Then, the GDSNE-FE technique used Jensen–Shannon divergence which measures similarity among the relevant features to extract most significant features for predicting disease at an earlier stage with minimum time. Therefore, GDSNE-FE technique improves features extraction accuracy of disease diagnosis with higher true positive rate and minimum space and time. The detailed process of GDSNE-FE technique is explained in below sections.

2.1 Gaussian Probability Distribution Based Relevant Features Extraction

The Gaussian Probability Distribution Based Relevant Features Extraction (GPD-RFE) is designed in GDSNE-FE technique to extracts relevant features from medical dataset. In GDSNE-RFE technique, Gaussian distribution is employed to differentiate random variables whose distributions are unknown. By using this, GPD-RFE measures probability distributions to find relationship between features over a high dimensional space using distance metrics. After computing probability distribution, relevant features have high probability of being picked and irrelevant features have very lesser probability of being picked. Thus, GPD-RFE extracts relevant features from medical dataset. The process involved in Gaussian Probability Distribution for relevant features extraction is shown in below Figure 2.

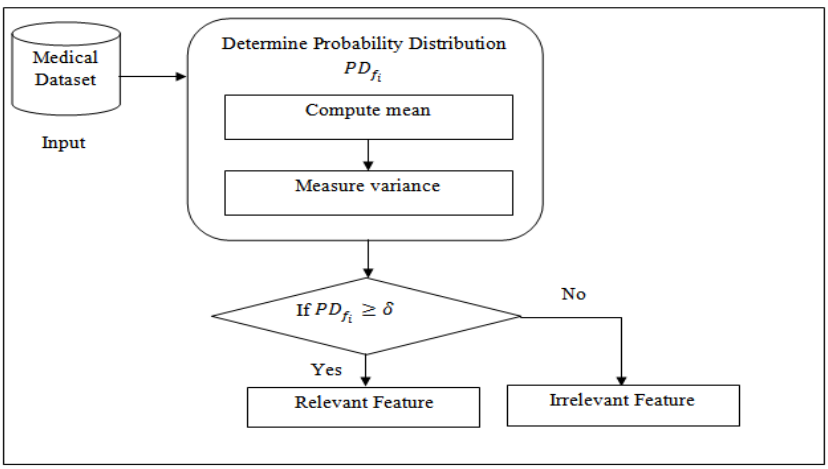

Figure 2 Processes of Gaussian Probability Distribution for Relevant Features Extraction

Figure 2 shows the block diagram of Gaussian Probability Distribution for extracting features relevant to disease diagnosis. As illustrated in figure 1, GPD-RFE measures probability distribution for each feature in dimensional space with help of mean and variance. With help of determined probability distribution, the GPD-RFE take outs only a relevant features from medical dataset for efficient disease prediction.

Let us consider a collection of N features in given medical dataset is represented as $DS = \{f_1, f_2, f_3, \ldots f_N\}$. Here, N denotes the number of features in input dataset. Initially GPD-RFE randomly embeds input medical dataset into a high dimensional space to identify the type of relationship between two quantitative variables. Then GPD-RFE determines probability distribution PD that represents relationship between feature f_i and neighbour feature f_j using distance metric d as similar objects are modeled by nearby points and dissimilar objects are modeled by distant points. From that, probability distribution between features is achieved using below mathematical formula,

$$PD(f_{ij}, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} exp\left(\frac{f_{ij} - \mu}{2\sigma^2}\right) \quad (1)$$

From equation (1), probability distribution is evaluated for each feature f_i and their neighbour feature f_j in high dimensional space. Here, μ denote mean distance where σ refers the standard deviation and σ^2 represent the variance of distance between the features. The mean of distribution is estimated using following mathematical representation,

$$\mu_{f_i} = \frac{1}{N}\sum_{i=1}^{N} f_i \quad (2)$$

From equation (2), mean of distribution is measured for all features in given medical dataset. Followed by, the variance of distribution is obtained as,

$$\sigma_{f_i} = \frac{1}{N}\sum_{i=1}^{N} (f_i - \mu_{f_i}) \quad (3)$$

From equation (3), variance of distribution is determined for every feature in input medical dataset to compute the probability distribution. A feature with n data points has n corresponding distributions that control the density of all features of data points. Specially, the density at the mean point is higher and the variance reflects the variation degree of between features. Thus, density (i.e. length) of feature f_i is measured as the product of each data points which is evaluated using below mathematical expression,

$$PD_{f_i} = \prod_{i=1}^{n} (f_i, \mu_i, \sigma_i) \quad (4)$$

$$PD_{f_i} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i}} exp\left(\frac{f_i - \mu_i}{2\sigma_i^2}\right) \quad (5)$$

From equation (4) and (5), $i = 1 \; to \; n$ represents the data points of features in dataset. By using equation (5), the GPD-RFE measures the probability distribution for each feature. After computing probability distribution, GPD-RFE considers the threshold δ probability distribution value in order to select only relevant features for disease prediction and thereby attaining dimensionality reduction. By defining threshold δ, the feature with a probability distribution higher than a threshold values are selected as relevant features for performing disease prognosis which is expressed as,

$$If \; PD_{f_i} \begin{cases} \geq \delta \rightarrow Relevant \\ < \delta \rightarrow Irrelevant \end{cases} \quad (6)$$

From equation (6), the relevant features in medical data set are selected with minimum time in order to attain the prediction performance for diseases diagnosis. The algorithmic process of Gaussian Probability Distribution for relevant features extraction is shown in below.

// Gaussian Probability Distribution Based Relevant Feature Extraction Algorithm
Input: Medical Dataset, Threshold Probability Distribution Value δ
Output: Relevant features
Step 1: Begin
Step 2: Features in medical dataset are randomly embedded into a high dimensional space
Step 3: **For** each features in high dimensional space
Step 4: Measure mean of distribution using (2)
Step 5: Evaluate variance of distribution using (3)
Step 6: Determine probability distribution PD_{f_i} through considering feature Density using (4)
Step 7: **If** $PD_{f_i} \geq \delta$, **then**
Step 8: The feature is relevant for disease diagnosis
Step 9: **Else**
Step 10: The feature is irrelevant for disease diagnosis
Step 11: **End if**
Step 12: **End For**
Step 13: End

Algorithm 1 Gaussian Probability Distribution Based Relevant Features Extraction

Algorithm 1 shows the step by step process involved in GPD-RFE for extracting relevant features from medical dataset. The GPD-RFE at first randomly embeds medical dataset into a high dimensional space. Then, GPD-RFE estimates the probability distribution for each feature in order to evaluate the relationship among features in high dimensional space with help of distance metric d as similar objects are modeled by nearby points and dissimilar objects are modeled by distant points. If probability distribution of feature is higher than threshold probability distribution value δ, the feature is relevant feature for predicting disease. Otherwise, the feature is irrelevant for performing disease diagnosis. Thus, GDSNE-FE technique significantly reduces the dimensionality of features in order to enhance the performance of disease prediction at an early stage.

After extracting the relevant features, GDSNE-FE technique applied Jensen–Shannon divergence which selects the most important features (i.e. optimal) for effective disease prediction. The elaborate process of Jensen–Shannon divergence is shown in below.

2.2 Jensen–Shannon Divergence Based Optimal Features Extraction

The Jensen–Shannon Divergence Based Optimal Features Extraction (JSD-OFE) is designed in order to extract optimal features from medical dataset for predicting disease at an early stage with minimum time complexity. The Jensen–Shannon divergence estimates the similarity among features probability distributions which helps for proposed GDSNE-FE technique help to determine most significant features which lead to the best prediction performance for disease diagnosis. The block diagram of Jensen–Shannon divergence for optimal feature extraction is shown in below Figure 3.

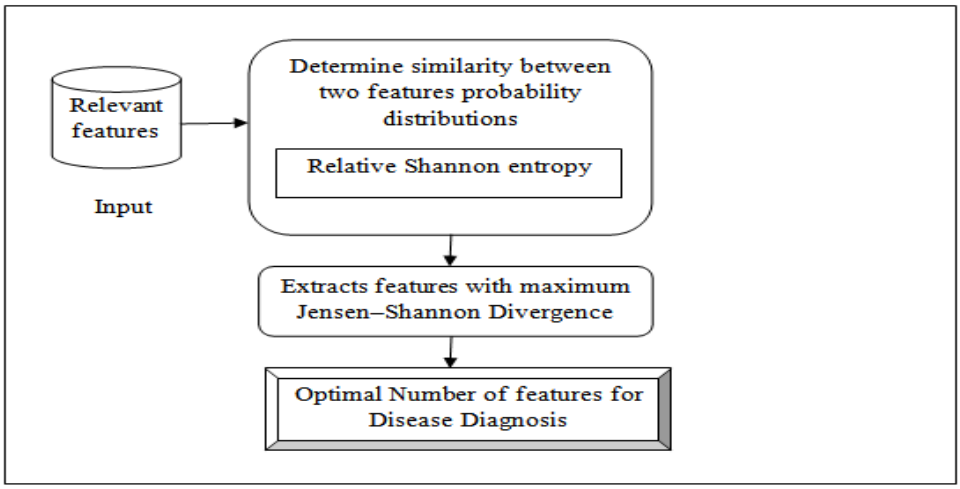

Figure 3 Block Diagram of Jensen–Shannon Divergence Based Optimal Features Extraction for Disease Diagnosis

Figure 3 shows the process involved in Jensen–Shannon Divergence for extracting optimal features from medical dataset to attain higher disease prediction performance. As demonstrated in figure 3, JSD-OFE initially computes similarity among features probability distributions (i.e. Jensen–Shannon Divergence) with help of Relative Shannon entropy. After that, JSD-OFE selects features which have maximum Jensen–Shannon Divergence as optimal for effective disease diagnosis.

Let us consider the features $\{f_1, f_2, f_3, \ldots f_N\}$ with set of N probability distributions $PD_{f_1}, PD_{f_2}, \ldots PD_{f_N}$. The similarity between two features probability distribution (i.e. Jensen–Shannon divergence) is determined using following mathematical formulation,

$$JSD\left[PD_{f_1}, PD_{f_2}\right] = S\left[\frac{PD_{f_1}+PD_{f_2}}{2}\right] - \frac{S[PD_{f_1}]+S[PD_{f_2}]}{2} \qquad (7)$$

From equation (7), $S[PD_{f_1}]$ represents the relative Shannon entropy of feature probability distribution PD_{f_1}. The relative Shannon entropy is measured using below mathematical representation,

$$S\left[PD_{f_1}, PD_{f_2}\right] = \int PD_{f_1} \ln \frac{PD_{f_1}}{PD_{f_2}} df \qquad (8)$$

From equation (8), relative Shannon entropy of each feature probability distribution is measured. By using equation (7) and (8), Jensen–Shannon divergence among the features probability distribution is estimated in order to select the best features for diseases prediction. The output of Jensen–Shannon divergence of feature probability distribution is always ranges between 0 or 1. Here, 1 indicates that the feature is more relevant or significant (i.e. optimal) for performing disease prognosis whereas 0 represents the less significant features. From that, GDSNE-FE technique extracts features which have maximum Jensen–Shannon divergence as

optimal features for efficient disease prediction. The Jensen–Shannon divergence take outs optimal features f^* from medical dataset using below formulations,

$$f^* = arg\ \max JSD[PD_{f_1}, PD_{f_2}, \dots, PD_{f_N}] \quad (9)$$

By using (8), GDSNE-FE technique extracts optimal features for predicting the diseases at an early stage with higher accuracy. The algorithmic process of Jensen–Shannon Divergence for optimal features extraction is shown in below,

// **Jensen–Shannon Divergence based Optimal Features Extraction Algorithm**
Input: Relevant Features with probability distributions $PD_{f_1}, PD_{f_2}, \dots PD_{f_N}$
Output: Optimal Features for Disease Prognosis
Step 1: Begin
Step 2: **For** each feature probability distribution
Step 3: Measure relative Shannon entropy using (8)
Step 4: Determine Jensen–Shannon divergence using (7)
Step 5: Extract optimal features for disease prognosis using (9)
Step 6: **End for**
Step 7:End

Algorithm 2 Jensen–Shannon Divergence based Optimal Feature Extraction

Algorithm 2 depicts the step by step process of Jensen–Shannon Divergence to extract the optimal features for improving diseases diagnosis performance. As demonstrated in algorithm, initially Jensen–Shannon divergence is evaluated for each feature probability distribution with help of relative Shannon entropy. After measuring Jensen–Shannon divergence, JSD-OFE extracts only features with higher Jensen–Shannon divergence as best features for predicting occurrences of disease with higher accuracy and minimum time. As a result, GDSNE-FE technique improves the feature extraction accuracy of disease diagnosis with minimum false positive rate and time complexity.

III. EXPERIMENTAL SETTINGS

In order to analyze the performance, Gaussian Distributive Stochastic Neighbor Embedding Feature Extraction (GDSNE-FE) technique is implemented in Java Language using Epileptic Seizure Recognition Data Set [21]. The Epileptic Seizure Recognition Data Set is obtained from UCI machine learning repository that includes of 5 different folders. The each folder consists of 100 files where each file denotes single patient information. Each file stores brain activity of patients for 23.6 seconds. Each data point is value of EEG recording at a diverse point in time. The corresponding time-series comprises 4097 data points. Therefore, Epileptic Seizure Recognition Data Set contains total 500 patients information with each has 4097 data points for 23.5 seconds.

The GDSNE-FE technique splits every 4097 data points into 23 segments. Each segment includes 178 data points for 1 second where data point denotes the value of EEG recording of patients at a various point in time. Here, each patient's information contains 178 feature data points for 1 second and the last column denotes the label y $\{1,2,3,4,5\}$. In epileptic seizure recognition data set, 178 data points of brain activity are considered as features for brain tumor disease predication. From Epileptic Seizure Recognition Data Set, GDSNE-FE technique extracts most important features (i.e. data points) to improve prediction performance of brain tumor disease at an early stage. The performance of GDSNE-FE technique is measured in terms of feature extraction accuracy, feature extraction time, true positive rate, false positive rate and space complexity. The efficiency of proposed GDSNE-FE technique is compared against with existing Recursive Feature Elimination (RFE) technique [1] and Orthogonal Feature Extraction (OFE) model [2].

IV. RESULTS AND DISCUSSIONS

In this section, the performance of proposed GDSNE-FE technique is discussed. In order to evaluate the efficacy of GDSNE-FE technique, comparisons is made with two existing methods namely Recursive Feature Elimination (RFE) technique [1] and Orthogonal Feature Extraction (OFE) model [2]. The effectiveness of GDSNE-FE technique is evaluated in terms of feature extraction accuracy, false positive rate, true positive rate, and feature extraction time and space complexity. The Experimental results are compared and analyzed with the assist of table and graph.

4.1 Measure of Feature Extraction Accuracy

Feature extraction accuracy is defined as the ratio of number of more relevant features selected as optimal to the total number of features. The formula for feature extraction accuracy is measured as follows,

$$FEA = \frac{Number\ of\ Optimal\ Features\ Extracted}{Total\ Number\ of\ Features} * 100 \quad (10)$$

From equation (10) FEA denotes feature extraction accuracy. When feature selection accuracy is higher, the method is said to more effectual.

Table 1 Tabulation for Feature Extraction Accuracy

Number of Features (N)	Feature Extraction Accuracy (%)		
	RFE technique	OFE model	GDSNE-FE technique
15	66	73	87
30	69	75	88
45	71	76	89
60	73	78	90
75	75	79	91
90	79	83	92
105	80	84	94
120	82	85	95
135	84	87	96
150	85	88	97

The performance result analysis of feature extraction accuracy for brain tumor disease diagnosis with respect to different number of features using three methods is demonstrated in Table 1. The GDSNE-FE technique considers framework with different number of features in the range of 15 to 150 for performing experimental evaluation using Java Language. While considering 75 features from Epileptic Seizure Recognition Data Set for conducting the experimental work, proposed GDSNE-FE technique gets 91 % feature extraction accuracy whereas conventional RFE technique [1] and OFE model [2] obtains 75 % and 79 % respectively. From that, it is clear that the feature extraction accuracy for brain tumor diseases prognosis using proposed GDSNE-FE technique is higher than other existing methods [1], [2].

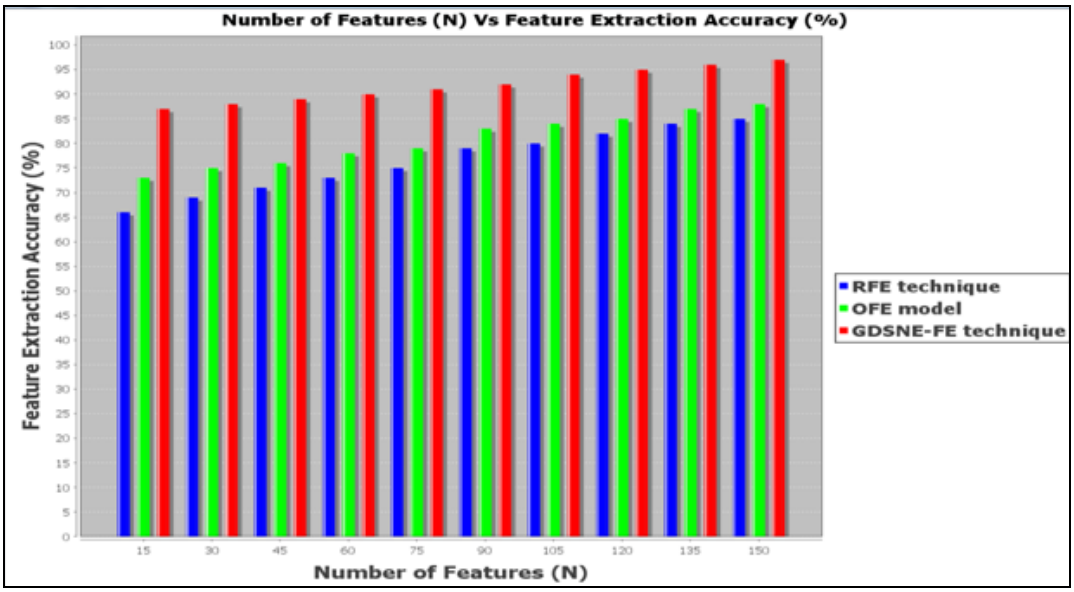

Figure 4 Performance of Feature Extraction Accuracy versus Number of Features

Figure 4 presents the impact of feature extraction accuracy versus number of features using three methods. As demonstrated in figure, the proposed GDSNE-FE technique provides better feature extraction accuracy for predicting brain tumor disease when compared to existing RFE technique [1] and OFE model [2]. Besides while increasing the number of feature for conducting the experimental work, feature extraction accuracy is also gets enhanced using all three methods. But comparatively, feature extraction accuracy using proposed GDSNE-FE technique is higher as compared to other existing methods. This is because of application of GDSNE in proposed GDSNE-FE technique. The GDSNE applies Gaussian probability distribution and Jensen–Shannon Divergence on the contrary to existing Stochastic Neighbor Embedding to attain higher feature extraction performance for brain tumor disease diagnosis. With application of GDSNE, proposed technique extracts more significant features from epileptic seizure recognition dataset for predicting brain tumor disease. This helps for GDSNE-FE technique to improve the feature extraction accuracy. Therefore, proposed GDSNE-FE technique increases the features extraction accuracy to identify brain tumor disease by 21 % and 14 % as compared to existing RFE technique [1] and OFE model [2] respectively.

4.2 Measure of Feature Extraction Time

Feature extraction time is defined as an amount of time required to extract an optimal number of features from the dataset. The formula for feature extraction time is expressed as follows,

$$FET = n * time\ (extracting\ optimal\ features\ ) \quad (11)$$

From equation (11), FET is the feature extraction time and 'n' denotes a number of features. Features extraction time is evaluated in millisecond (ms). When feature extraction time is lower, the method is said to more effective.

Table 2 Tabulation for Feature Extraction Time

Number of Features (N)	Feature Extraction Time (ms)		
	RFE technique	OFE model	GDSNE-FE technique
15	27.8	25.5	18.2
30	32.4	29.7	21.5
45	40.2	32.6	25.6
60	44.9	35.1	28.6
75	48.3	39.5	32.4
90	52.3	42.8	37.1
105	57.1	46.5	41.2
120	60.8	52.6	45.8
135	63.4	56.2	50.7
150	68.7	62.8	54.6

Table 2 depicts comparative result analysis of time needed to mine an optimal number of features from dataset for diagnosing brain tumor disease based on various numbers of features in the range of 15 to 150 using three methods. When taking 105 features from Epileptic Seizure Recognition Data Set for carry outing the experimental process, proposed GDSNE-FE technique obtains 41.2 ms feature extraction time whereas existing RFE technique [1] and OFE model [2] acquires 57.1 ms and 46.5 ms respectively. Thus, it is expressive that the feature extraction time for brain tumor diseases prognosis using proposed GDSNE-FE technique is lower than other existing methods [1], [2].

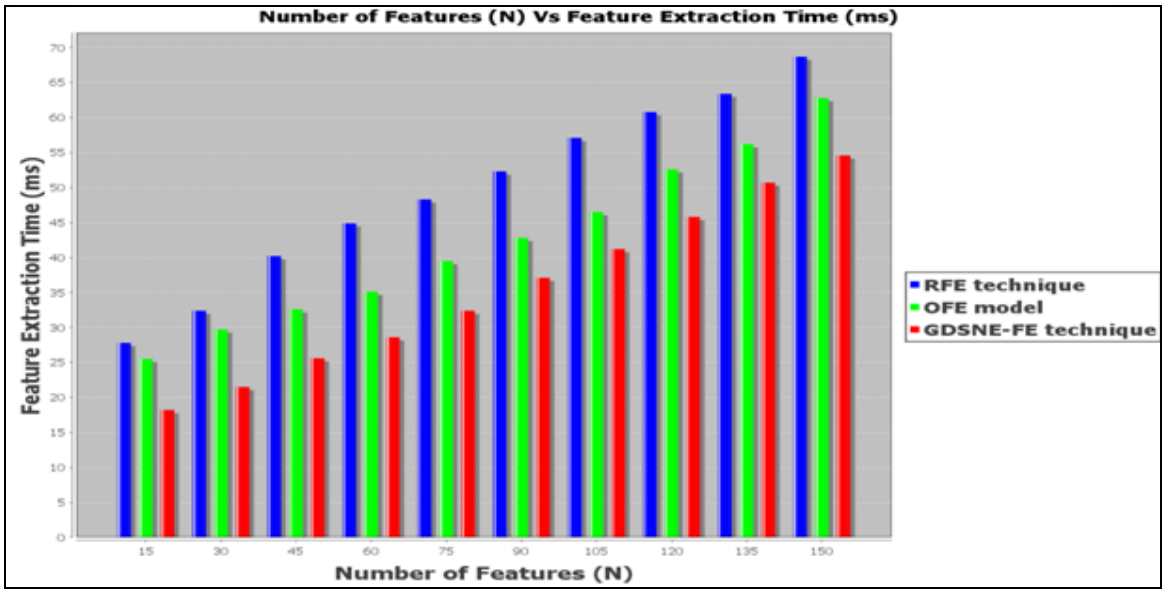

Figure 5 Performance of Feature Extraction Time versus Number of Features

Figure 5 describes the impact of feature extraction time versus number of features using three methods. As exposed in figure, the proposed GDSNE-FE technique provides better feature extraction time for performing brain tumor disease prognosis process when compared to existing RFE technique [1] and OFE model [2]. Further while increasing the number of feature for experimental evaluation, feature extraction time is also gets increased using all three methods. But comparatively, feature extraction time using proposed GDSNE-FE technique is lower as compared to other existing methods. This is owing to application of Gaussian Distributive Stochastic Neighbor Embedding (GDSNE) in proposed GDSNE-FE technique where it employs Gaussian probability distribution and Jensen–Shannon Divergence on the contrary to existing Stochastic Neighbor Embedding to improve feature extraction performance of brain tumor disease diagnosis with minimum time. By using algorithmic process of GDSNE, proposed technique takes outs more important features that presents in

epileptic seizure recognition dataset for brain tumor disease prediction with amount of minimum time utilization. This assists for GDSNE-FE technique to minimize the feature extraction time. As a result, proposed GDSNE-FE technique minimizes the features extraction time of brain tumor disease prognosis by 30 % and 17 % as compared to existing RFE technique [1] and OFE model [2] respectively.

4.3 Measure of True Positive Rate

True positive rate is defined as the ratios of number of features that correctly extracted as optimal to the total number of features in a dataset. The true positive rate is determined as follows,

$$True\ positive\ rate = \frac{Number\ of\ Features\ Correctly\ Extracted\ as\ Optimal}{Total\ Number\ of\ Features} * 100 \quad (12)$$

From equation (12), true positive rate of feature extraction is estimated. It is measured in terms of percentage (%).When true positive rate of feature extraction is higher; the method is said to more effectual.

Table 3 Tabulation for True positive rate

Number of Features (N)	True positive rate (%)		
	RFE technique	OFE model	GDSNE-FE technique
15	55	63	84
30	56	67	85
45	57	68	87
60	60	69	88
75	64	72	89
90	65	73	90
105	67	75	92
120	68	76	93
135	70	78	94
150	72	79	96

Table 3 portrays true positive rate results of optimal features extraction for effective brain tumor disease prognosis with respect to diverse numbers of features in the range of 15 to 150 using three methods. When assuming 120 features from Epileptic Seizure Recognition Data Set for experimental evaluation, proposed GDSNE-FE technique achieves 93 % true positive rate whereas existing RFE technique [1] and OFE model [2] acquires 68 % and 76 % respectively. Hence, it is descriptive that the true positive rate using proposed GDSNE-FE technique is higher than other existing methods [1], [2].

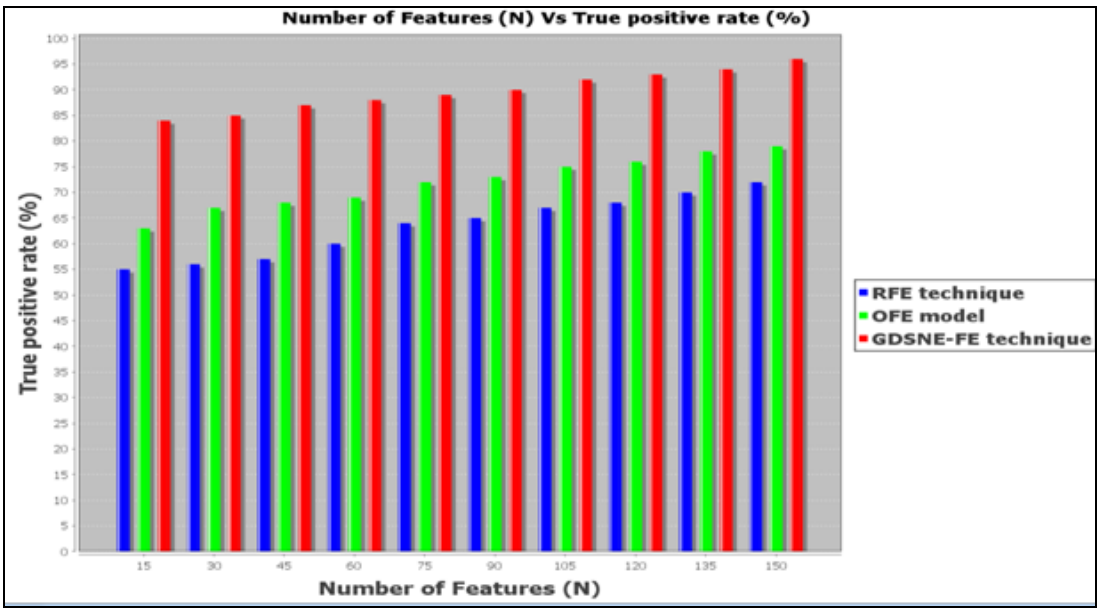

Figure 6 Performance of True positive rate versus Number of Features

Figure 6 demonstrates the impact of true positive rate versus number of features using three methods. As shown in figure, the proposed GDSNE-FE technique provides better true positive rate of feature extraction for brain tumor diseases prediction when compared to existing RFE technique [1] and OFE model [2]. Besides while increasing the number of feature for experimental work, true positive rate is also gets increased using all three methods. But comparatively, true positive rate using proposed GDSNE-FE technique is higher as compared to other existing methods. This is due to application of Gaussian Distributive Stochastic Neighbor

Embedding (GDSNE) in proposed GDSNE-FE technique. The GDSNE used Gaussian probability distribution and Jensen–Shannon Divergence which help to extract more optimal features in epileptic seizure recognition dataset for identifying brain tumor disease at an early stage. This supports for GDSNE-FE technique to attain higher true positive rate. Thus, proposed GDSNE-FE technique enhances the true positive rate of feature extraction to find presence of brain tumor disease by 42 % and 25 % as compared to existing RFE technique [1] and OFE model [2] respectively.

4.4 Measure of False Positive Rate

False positive rate is defined as the ratios of number of features that incorrectly extracted as optimal to the total number of features in a dataset. The false positive rate of feature extraction is measured as,

$$False\ positive\ rate = \frac{Number\ of\ Features\ Incorrectly\ Extracted\ as\ Optimal}{Total\ Number\ of\ Features} * 100 \quad (13)$$

From equation (13), false positive rate of feature extraction is determined and which is measured in terms of percentage (%).When false positive rate of feature extraction is lower, the method is said to more effective.

Table 4 Tabulation for False positive rate

Number of Features (N)	False positive rate (%)		
	RFE technique	OFE model	GDSNE-FE technique
15	45	36	15
30	47	37	18
45	48	39	19
60	51	42	22
75	53	43	24
90	54	45	25
105	57	46	28
120	58	49	30
135	59	50	33
150	62	51	36

Table 4 explains experimental results of false positive rate for feature extraction based on varied numbers of features in the range of 15 to 150 using three methods. When considering 90 features from Epileptic Seizure Recognition Data Set for accomplishing experimental work, proposed GDSNE-FE technique attains 25 % false positive rate whereas existing RFE technique [1] and OFE model [2] gets 54 % and 45 % respectively. From these results, it is expressive that the false positive rate of feature extraction for brain tumor disease prediction using proposed GDSNE-FE technique is lower than other existing methods [1], [2].

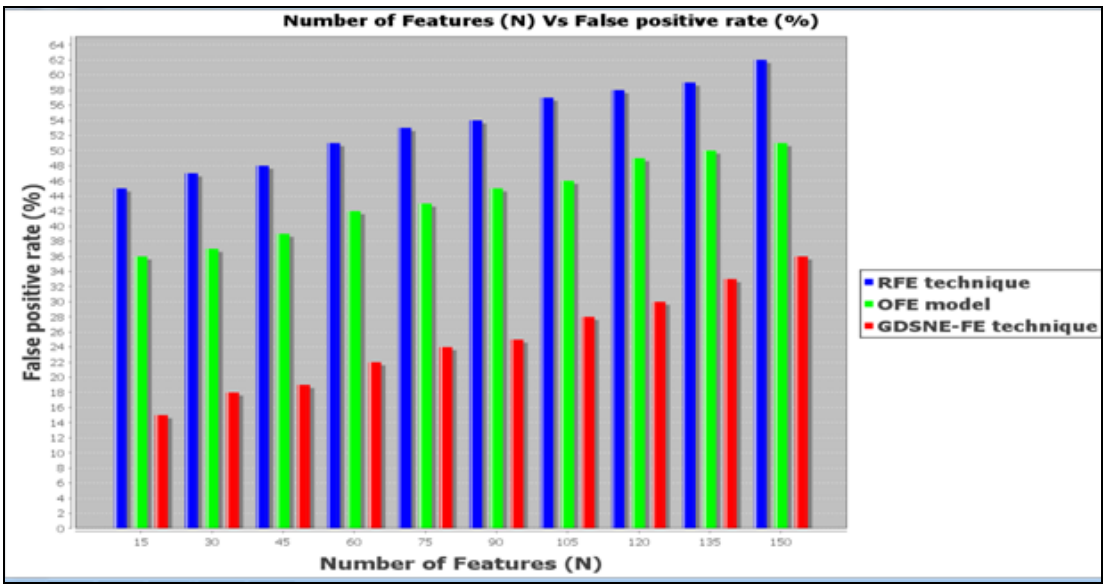

Figure 7 Performance of False positive rate versus Number of Features

Figure 7 reveals the impact of false positive rate versus number of features using three methods. As illustrated in figure, the proposed GDSNE-FE technique provides better false positive rate of feature extraction for brain tumor diseases prognosis when compared to existing RFE technique [1] and OFE model [2]. Also,

while increasing the number of feature for conducting experimental process, false positive rate is also gets increased using all three methods. But comparatively, false positive rate using proposed GDSNE-FE technique is lower as compared to other existing methods. This is due to usage of Gaussian Distributive Stochastic Neighbor Embedding (GDSNE) in proposed GDSNE-FE technique. The GDSNE algorithm calculates the probability distribution to discover relationship among features in high dimensional space with aid of distance metric. This measured probability distribution helps for GDSNE-FE technique to extract relevant feature from epileptic seizure recognition dataset. After extracting the relevant features, GDSNE-FE technique utilized Jensen–Shannon divergence that measures similarity between two features probability distributions to pick optimal features for effective brain tumor disease prediction. This helps for GDSNE-FE technique to get lower false positive rate. As a result, proposed GDSNE-FE technique reduces the false positive rate of feature extraction to discover presence of brain tumor disease by 54 % and 44 % as compared to existing RFE technique [1] and OFE model [2] respectively.

4.5 Measure of Space Complexity

The space complexity measures an amount of memory space required to store extracted optimal features from the medical dataset. The formula for determining space complexity is formulated as below,

$$space\ complexity = n * memory\ (storing\ optimal\ features) \quad (14)$$

From equation (14), 'n' denotes a number of features. The space complexity time is measured in terms of millisecond (ms). When space complexity is lower, the method is said to more efficient.

Table 5 Tabulation for Space Complexity

Number of Features (N)	Space Complexity (KB)		
	RFE technique	OFE model	GDSNE-FE technique
15	20	18	12
30	23	22	15
45	27	23	18
60	29	25	19
75	32	29	21
90	35	32	24
105	36	34	25
120	39	37	27
135	44	41	30
150	47	45	31

Table 5 shows performance results of space complexity for brain tumor disease diagnosis with respect to different numbers of features in the range of 15 to 150 using three methods. When considering 135 features from Epileptic Seizure Recognition Data Set for conducting experimental evaluation, proposed GDSNE-FE technique obtains 30 KB space complexity whereas existing RFE technique [1] and OFE model [2] attains 44 KB and 41 KB respectively. Therefore, the space complexity of brain tumor disease prognosis using proposed GDSNE-FE technique is lower than other existing methods [1], [2].

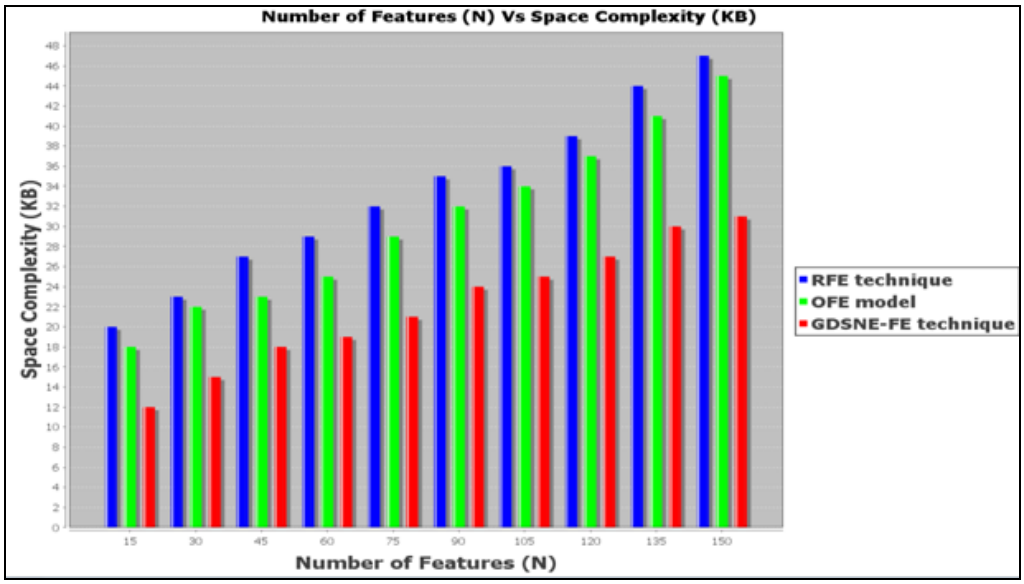

Figure 8 Performance of Space Complexity versus Number of Features

Figure 8 depicts the impact of space complexity for brain tumor disease prediction versus number of features using three methods. As exposed in figure, the proposed GDSNE-FE technique provides better space complexity for identifying brain tumor diseases when compared to existing RFE technique [1] and OFE model [2]. As well, while increasing the number of feature for carry outing experimental process, space complexity is also gets increased using all three methods. But comparatively, space complexity using proposed GDSNE-FE technique is lower as compared to other existing methods. This is because of process of Gaussian Distributive Stochastic Neighbor Embedding (GDSNE) in proposed GDSNE-FE technique where it applied Gaussian probability distribution and Jensen–Shannon Divergence to extract only optimal features in epileptic seizure recognition dataset for predicting brain tumor disease at an early stage. This helps for GDSNE-FE technique to attain minimum space complexity for brain tumor disease diagnosis. Therefore, proposed GDSNE-FE technique lessens the space complexity of brain tumor disease diagnosis by 34 % and 27 % as compared to existing RFE technique [1] and OFE model [2] respectively.

V. RELATED WORKS

A comparative study of feature extraction technique designed for diagnosis of Alzheimer's disease was presented in [11]. Support Vector Machines was employed in [12] for feature selection and predicting diabetes disease. However, more relevant features were not selected. An Artificial Bee Colony (ABC) algorithm was intended in [13] for feature selection of medical data classification. But, feature selection was not effective which lacks prediction accuracy.

Hybrid feature selection (HFS) was presented in [14] in order to perform breast cancer and diabetes disease diagnosis. But, time complexity of diseases forecasting was more. An incremental mechanism was developed in [15] to choose significant feature with aid of rough set model with minimum time. While increasing the size of features samples, incremental mechanism requires more processing time for selecting relevant features from dataset.

Feature extraction method was designed in [16] using mutual information for attaining the dimensionality reduction with minimum computational complexity. False positive rate of feature extraction was higher. The hybridization approach was introduced in [17] with aiming at enhancing diagnostic accuracy based on the extracted significant diabetes features. However, space complexity of disease diagnosis was not considered.

A linguistic hedges neuro-fuzzy classifier was designed in [18] for reducing dimensionality of features. The true positive rate of feature selection was not at required level. A hybrid decision support system was designed in [19] to attain higher classification performance for diabetes disease diagnosis through selecting relevant features. But, feature selection accuracy was poor.

A Random forest classifier was designed in [20] for feature selection and enhances predictive accuracy of breast cancer prognosis. The efficiency of feature selection was not sufficient for effective disease prediction. To solve the above mentioned existing issues, Gaussian Distributive Stochastic Neighbor Embedding Feature Extraction (GDSNE-FE) technique is designed.

VI. CONCLUSION

An effective Gaussian Distributive Stochastic Neighbor Embedding Feature Extraction (GDSNE-FE) technique is developed with objective of reducing dimensionality of medical dataset with minimal time and space complexity for predicting diseases at an early stage. The GDSNE-FE technique includes of two key steps. During first step, GDSNE-FE technique determines the relationship among features in high dimensional space using distance metric with application of Gaussian probability distribution. After evaluating the probability distribution, relevant features have high probability of being chosen whereas irrelevant features have lesser probability of being selected. By using this probability distribution, GDSNE-FE technique removes irrelevant features and extracts relevant features from medical dataset for performing disease diagnosis. Then, MDSNE algorithm utilized Jensen–Shannon divergence that determines similarity between two feature probability distributions during second step. With help of measured Jensen–Shannon divergence, GDSNE-FE technique extracts optimal features in order to precisely predict disease with higher accuracy and minimum time. As a result, GDSNE-FE technique minimizes the dimensionality of medical dataset for disease diagnosis with minimum false positive rate and feature extraction time. The performance of GDSNE-FE technique is estimated in terms of feature selection accuracy, true positive rate, and feature selection time and space complexity using epileptic seizure recognition dataset and compared with state-of-the-art works. With the experimental carried out for GDSNE-FE technique, it is clear that the feature extraction accuracy is higher as compared to state-of-the-art works. The experimental results reveal that GDSNE-FE technique presents better performance with an improvement of true positive rate and minimization of feature extraction time when compared to the state-of-the-art works.

REFERENCES

[1]. Ning Wang, Michael R. Lyu, "Extracting and Selecting Distinctive EEG Features for Efficient Epileptic Seizure Prediction", IEEE Journal of Biomedical and Health Informatics, Volume 19, Issue 5, Pages 1648 – 1659, 2015

[2]. Hao Jiang, Wai-Ki Ching and Wenpin Hou "On Orthogonal Feature Extraction Model with Applications in Medical Prognosis", Applied Mathematical Modelling, Elsevier, Volume 40, Issues 19–20, October 2016, Pages 8766-8776

[3]. Maryam Mollaee and Mohammad Hossein Moattar, "A novel feature extraction approach based on ensemble feature selection and modified discriminant independent component analysis for microarray data classification", Biocybernetics and Biomedical Engineering, Elsevier, Volume 36, Issue 3, 2016, Pages 521-529

[4]. Sebastian Polsterl, Sailesh Conjeti Nassir Navab and Amin Katouzian, "Survival analysis for high-dimensional, heterogeneous medical data: exploring feature extraction as an alternative to feature selection", Artificial Intelligence in Medicine, Elsevier, Volume 72, September 2016, Pages 1-11

[5]. Elias R. SilvaJr, George D. C. Cavalcanti, TsangIng Ren, "Class-wise feature extraction technique for multimodal data", Neurocomputing, Elsevier, Volume 214, November 2016, Pages 1001-1010

[6]. Heng Kong, ZhihuiLai, XuWang, FengLiu, "Breast cancer discriminant feature analysis for diagnosis via jointly sparse learning", Neuro computing, Elsevier, Volume 177, Pages 198–205, 2016

[7]. Truyen Tran, Wei Luo, Dinh Phung, Sunil Gupta, Santu Rana, Richard Lee Kennedy, Ann Larkins, and Svetha Venkatesh, "A framework for feature extraction from hospital medical data with applications in risk prediction", BMC Bioinformatics, Volume 15, Issue 1, Pages 1-9, 2014

[8]. Mohamed M. Dessouky, Mohamed A. Elrashidy, Hatem M. Abdelkader, "Selecting and Extracting Effective Features for Automated Diagnosis of Alzheimer's Disease", International Journal of Computer Applications, Volume 81, Issue 4, Pages 17-28, November 2013

[9]. F. Segovia, J. M. Górriz, J. Ramírez, C. Phillips, "Combining Feature Extraction Methods to Assist the Diagnosis of Alzheimer's Disease", Current Alzheimer Research, Volume 13 , Issue 7 , Pages 831 – 837, 2016

[10]. Sidahmed Mokeddem, Baghdad Atmani and Mostéfa Mokaddem, "Supervised Feature Selection for Diagnosis Of Coronary Artery Disease Based On Genetic Algorithm", Computer Science & Information Technology (CS & IT), Pages 41–51, 2013.

[11]. F. Segovia, J.M.Gorriz, J.Ramirez, D.Salas-Gonzalez,I.A´ lvarez, M.Lo´ pez, R.Chaves, "A comparative study of feature extraction methods for the diagnosis of Alzheimer's disease using the ADNI database", Neuro computing, Elsevier, Volume 75, Pages 64–7, 2012

[12]. Fatma Patlar Akbulut and Aydın Akan, "Support Vector Machines Combined with Feature Selection for Diabetes Diagnosis", Istanbul University-Journal of Electrical & Electronics Engineering (IU-JEEE), Volume 17, Issue 1, Pages 3219-3225, 2017

[13]. Mustafa Serter Uzer, Nihat Yilmaz, and Onur Inan, "Feature Selection Method Based on Artificial Bee Colony Algorithm and Support Vector Machines for Medical Datasets Classification", The Scientific World Journal, Hindawi Publishing Corporation, Volume 2013, Pages 1-10, July 2013

[14]. Divya Tomar and Sonali Agarwal, "Hybrid Feature Selection Based Weighted Least Squares Twin Support Vector Machine Approach for Diagnosing Breast Cancer, Hepatitis, and Diabetes", Advances in Artificial Neural Systems, Hindawi Publishing Corporation, Volume 2015, Pages 1-10, December 2014

[15]. Jiye Liang, Feng Wang, Chuangyin Dang, and Yuhua Qian, "A Group Incremental Approach to Feature Selection Applying Rough Set Technique", IEEE Transactions on Knowledge and Data Engineering ,Volume 26, Issue 2, Pages 294 – 308, 2014

[16]. Liying Fang, Han Zhaoa, Pu Wanga, Mingwei Yud, Jianzhuo Yana, Wenshuai Cheng, Peiyu Chen, "Feature selection method based on mutual information and class separability for dimension reduction in multidimensional time series for clinical data", Biomedical Signal Processing and Control, Volume 21, Pages 82–89, 2015

[17]. Ahmed Hamza Osman and Hani Moetque Aljahdali, "Diabetes Disease Diagnosis Method based on Feature Extraction using K-SVM", International Journal of Advanced Computer Science and Applications, Volume 8, Issue 1, Pages 236-244, 2017

[18]. Ahmad Taher Azar and Aboul Ella Hassanien "Dimensionality reduction of medical big data using neural-fuzzy classifier", soft Computing, Springer, Volume 19, Issue 4, Pages 1115–1127, April 2015

[19]. Ramalingaswamy Cheruku, Damodar Reddy Edla, Venkatanareshbabu Kuppili, Ramesh Dharavath, "RST-BatMiner: A Fuzzy Rule Miner Integrating Rough Set Feature Selection and Bat Optimization for Detection of Diabetes Disease", Applied Soft Computing, Elsevier, Pages 1-50, 2017

[20]. Cuong Nguyen, Yong Wang, Ha Nam Nguyen, "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic", Journal of Biomedical Science and Engineering, Volume 6, Pages 551-560, 2013

[21]. EpilepticSeizure Recognition Data Set:https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition

Nithya C "Gaussian Distributive Stochastic Neighbor Embedding Based Feature Extraction for Medical Data Diagnosis "Optimization of process parameters of Material Removal Rate in Micro hole Machining by Die sinker EDM"IOSR Journal of Engineering (IOSRJEN), vol. 08, no. 6, 2018, pp. 27-40.