Kernel Perceptron Feature Selection Based on Sparse Bayesian Probabilistic Relevance Vector Machine Classification for Disease Diagnosis with Healthcare Data

Mr.G.Arun[1], Dr.C N Marimuthu[2]

Assistant Professor, Department of Information technology, MPNMJ Engineering College. Erode [1]
Professor, Department of Electronics and Communication Engineering, Nandha Engineering College [2]

Abstract

Disease diagnosis with big healthcare data is a significant problem to be resolved for finding the presence of disease at an early stage. The conventional classification techniques designed for disease prediction does not provide higher diagnosis rate. Besides, the feature selection accuracy of the existing algorithm is also lower. In order to solve this limitation, a Kernel Perceptron Feature Selection based Sparse Bayesian Probabilistic Relevance Vector Machine (KPFS-SBPRVM) Technique is proposed. The KPFS-SBPRVM Technique is designed for disease diagnosis with higher accuracy and lesser time. The KPFS-SBPRVM Technique comprises two steps, namely feature selection and classification for finding the existence of the disease in big healthcare data. Initially, Kernel Perceptron Feature Selection (KPFS) is performed which is a variant of perceptron learning algorithm with kernel function to extract the significant medical features from input big Healthcare dataset. With the relevant features, then Probabilistic Relevance Vector Machine Classification (SBPRVMC) step is carried out in KPFS-SBPRVM technique to classify the big healthcare data as normal data or abnormal data. SBPRVMC is a machine learning technique which uses Bayesian inference for probabilistic classification. In KPFS-SBPRVM technique, SBPRVMC constructs the hyperplane among the healthcare data to classify as normal data or abnormal data. By this way, the disease gets diagnosed at an early stage with higher accuracy and minimal time consumption. Experimental evaluation of KPFS-SBPRVM technique is carried out on factors such as feature selection rate, disease diagnosis rate, disease diagnosis time, and false-positive rate with respect to a number of patient's medical data.

Keywords: *Big healthcare data, Disease diagnosis, Hyper-parameter Vector, Kernel Perceptron Feature Selection, Patient Medical Data, Similarity, Sparse Bayesian Probabilistic Relevance Vector Machine Classification*

I. Introduction

Big data comprises of a huge amount of structured, semi-structured and unstructured data that are extracted for information employed in machine learning projects [24]. Big data in healthcare represents the abundant healthcare data from different sources such as electronic health records, medical imaging, pharmaceutical and medical devices for decision-making. Medical diagnosis is the process of finding the disease using their symptoms and signs. Feature selection is vital processing in different fields such as data mining, pattern recognition, and machine learning. Feature selection is the method of selecting relevant features that predicts the outputs. Classification is the process of classifying the patient medical data to find the occurrence of disease. However, the existing classification techniques failed to improve the disease diagnosis rate with minimal time. In order to address these problems, KPFS-SBPRVM technique is introduced in this research work by combining the Kernel Perceptron Feature Selection (KPFS) and Sparse Bayesian Probabilistic Relevance Vector Machine classification (SBPRVMC) algorithm.

A Globally optimized Artificial Neural Network Input Gain Measurement Approximation (GANNIGMA-ensemble) technique was designed in [1] using imbalanced healthcare data for diagnosis of brain tumor. However, feature selection accuracy was not sufficient. The Support Vector Machine (SVM) was employed to increase the classification accuracy of disease prediction [2]. But, disease diagnosis time using SVM was not reduced.

A big data analytics-enabled transformation model was introduced [3]. But, the feature selection process was not carried out in an effective manner. A 5G-Smart Diabetes system was introduced for analysis of patients suffering from diabetes [4]. However, the disease diagnosing rate was not improved by using 5G-Smart Diabetes system.

The medical TV controller reader was employed in [5] with the combination of digital device [25] and healthcare products for family health management. However, relevant features are not selected to achieve higher diagnosing rate. A statistical assessment model of the healthcare information system was introduced for Diabetes Analysis with big data [6]. Though accuracy and F-measure were improved, the time consumption was not reduced using statistical assessment model.

Deep neural architectures were designed to get higher classification accuracy results for prediction in healthcare [7]. However, a number of patient's medical data that are wrongly classified was more. The boosted neural network ensemble classification was performed in order to increase the performance of lung cancer diagnosis accuracy for big data [8]. But, disease diagnosis time was very higher.

A multi-modality-based decision framework (BHARAT) was employed for the classification of early Alzheimer's disease with minimal time [9]. However, false-positive rate using BHARAT was lower. A review of different techniques designed to analyze and to perform disease prediction using big healthcare information was presented [10].

In order to addresses the above said existing issues, KPFS-SBPRVM technique is proposed. The main contributions of KPFS-SBPRVM technique are described in below,

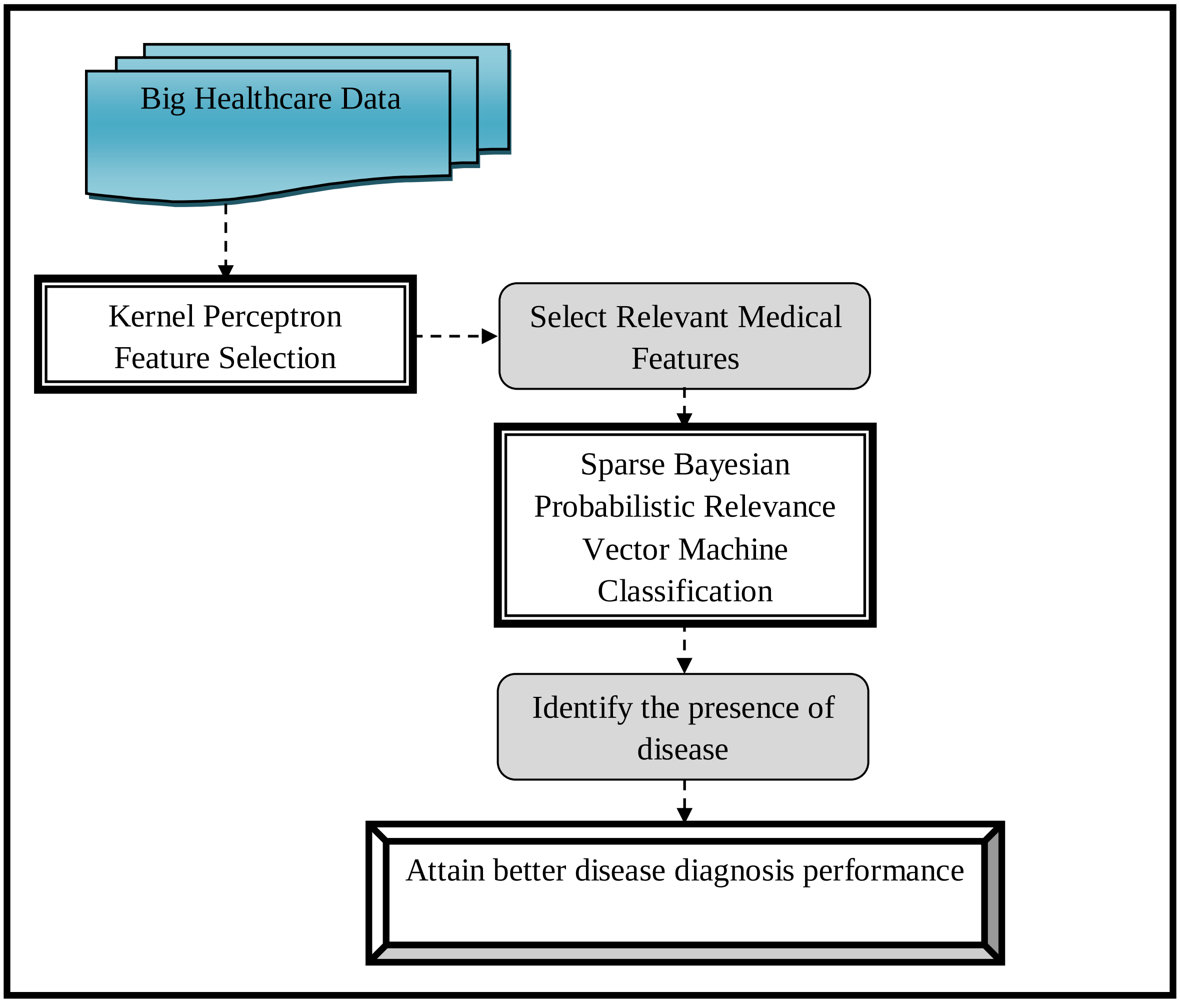
- ✓ To achieve improved disease diagnosis performance with a lower time complexity when compared to state-of-the-art works, KPFS-SBPRVM technique is introduced. On the contrary to conventional works, KPFS-SBPRVM technique is designed by integrating Kernel Perceptron Feature Selection (KPFS) and Sparse Bayesian Probabilistic Relevance Vector Machine classification (SBPRVMC) algorithm.
- ✓ To enhance the feature selection accuracy when compared to conventional works, Kernel Perceptron Feature Selection (KPFS) is performed in KPFS-SBPRVM technique on the contrary to existing works. In KPFS-SBPRVM technique, the KPFS iteratively improves a feature selection performance by running it on training samples, then updating the model whenever it finds it has made an incorrect classification. This helps for KPFS to significantly accomplish optimal feature selection as compared to conventional works.
- ✓ To minimize the false positive rate of disease diagnosis with a minimal amount of time as compared to existing works, Sparse Bayesian Probabilistic Relevance Vector Machine classification (SBPRVMC) algorithm is proposed in KPFS-SBPRVM technique. SBPRVMC attains much interest in the research community because they provide a number of advantages. The SBPRVMC model depends on a Bayesian formulation of a linear model with an appropriate prior which results in a sparse representation. As a result, SBPRVMC model provides disease diagnosis rate with a lower computational cost.

The rest of the paper is prepared as follows. Section 2 describes the detailed processes of KPFS-SBPRVM technique with the assist of architecture diagram. Section 3 and Section 4 shows the simulation settings and details performance analysis with the aid of parameters. Section 5 presents related works. Section 6 concludes the paper.

II. Kernel Perceptron Feature Selection Based Sparse Bayesian Probabilistic Relevance Vector Machine Classification Technique

The Kernel Perceptron Feature Selection based Sparse Bayesian Probabilistic Relevance Vector Machine (KPFS-SBPRVM) Technique is introduced in order to increase the disease diagnosis of big healthcare data with a lower time. The KPFS-SBPRVM Technique is developed with the application of Kernel Perceptron Feature Selection (KPFS) and Sparse Bayesian Probabilistic Relevance Vector Machine Classification (SBPRVMC) on the contrary to conventional algorithms. Therefore, KPFS- BPRVM Technique accurately performs disease diagnosis process at an early stage. The architecture diagram of KPFS-SBPRVM Technique is depicted in Figure 1.

Figure 1 shows the overall processes of KPFS-SBPRVM Technique. As demonstrated in the above figure, the KPFS-SBPRVM Technique initially acquires big healthcare dataset (i.e. Diabetes 130-hospitals for years 1999-2008 Data Set [21] and Epileptic Seizure Recognition Data Set) as input. Next, the KPFS-SBPRVM Technique carried outs Kernel Perceptron Feature Selection process where it discovers and extracts the medical features that are related to diagnosis of disease at an early stage.


Big Healthcare Data
Kernel Perceptron Feature Selection
Select Relevant Medical Features
Sparse Bayesian Probabilistic Relevance Vector Machine Classification
Identify the presence of disease
Attain better disease diagnosis performance

After selecting the medical features, KPFS-SBPRVM Technique performs Sparse Bayesian Probabilistic Relevance Vector Machine Classification. During this process, KPFS-SBPRVM Technique identifies the presence of disease through classification with minimal time. Thus, KPFS-SBPRVM Technique obtains better diagnosis performance for both diabetic and brain tumor disease at an early stage. The detailed process of KPFS-SBPRVM Technique is explained in below subsections.

A. Kernel Perceptron Feature Selection

In KPFS-SBPRVM technique, Kernel Perceptron Feature Selection is proposed in order to select the medical features that are more relevant to diagnose disease at an early stage with higher accuracy. The KPFS is a popular perceptron learning algorithm that employs a kernel function to compute the similarity between the medical features in input big healthcare dataset. In KPFS, kernel function finds relations between the input medical features. The medical features in raw representation have to be explicitly transformed into feature vector representations via a user-specified feature map. On the contrary, a kernel method considers only similarity function over pairs of data features. Prediction for unlabeled inputs, i.e., those not in the training set is treated by the application of a similarity function 'φ', i.e. kernel between the unlabeled input 'f'' and each of the training inputs 'f_i'.

Let us assume a big healthcare dataset comprises of 'n' number of medical features represented as '$f_i = \{f_1, f_2, f_3, .. f_n\}$'. Here, '$n$' denotes the total number of medical features in the input dataset. The KPFS is algorithm operates by a principle called 'error-driven learning'. In proposed technique, the KPFS iteratively improves a feature selection performance by running it on training samples, then updating the model whenever it finds it has made an incorrect classification. At first, KPFS initializes weight '$w_i = 0$' and error threshold '$ER = \alpha$'. Then, KPFS determines a weighted sum of similarities between input medical feature ‘f_i’ and disease symptoms ‘f'’ is mathematically using below,

$$x(t) = sgn \sum_{i=1}^{n} \varepsilon_i \varphi(f_i, f') \qquad (1)$$

From the above mathematical equations (1), ‘x’ represents the predicted result whereas ‘φ’ denotes the kernel function that determines the similarity between input medical feature ‘f_i’ and disease symptoms ‘f'’. Here, ‘ε_i’ indicates the weights for the training input medical feature which is computed by learning algorithm and ‘sgn’ is a sign function which evaluates whether the predicted output is positives or negative. To enhance the feature selection performance through classification as compared to existing works, error ‘E’ is measured in KPFS for each obtained prediction result using below,

$$E = (\overline{x(t)} - x(t)) \qquad (2)$$

From the above mathematical expression (2), misclassification error of feature selection is determined to attain higher accuracy. Followed by, KPFS algorithm updates the weight ‘ε_i’ for each predicted result of medical features according to error using below mathematical expression,

$$\varepsilon_i(t+1) = \varepsilon_i(t) + l.(\overline{x(t)} - x(t)) \qquad (3)$$

From the above formula (3), ‘$\varepsilon_i(t+1)$’ represents an updated weight whereas ‘$\overline{x(t)}$’ denotes the actual output and ‘$x(t)$’ predicted output. Here, ‘l’ indicates the learning rate. The above processes of KPFS is repeated until the iteration error is lower than a user-specified error threshold ‘ER’. As a result, KPFS accurately classify the each input medical feature as relevant or irrelevant with a minimal amount of time consumption.

The algorithmic process of Kernel Perceptron Feature Selection is depicted below,

```
// Kernel Perceptron Feature Selection Algorithm
Input: Number Of Medical Features ‘$f_i = f_1, f_2, .. f_n$’; Disease Symptoms ‘$f'$’; user-specified error threshold ‘$ER$’;Weight ‘$\varepsilon_i$’
Output: Attain higher accuracy for feature selection
Step 1:Begin
Step 2:   Initialize the weight ‘$\varepsilon_i = 0$’ and the threshold ‘$ER = \alpha$’’
Step 3:   For each medical features ‘$f_i$’
Step 4:      While ($ER == \alpha$’’), do
Step 5:         Calculate similarities between features and disease symptoms ‘$x(t)$’ using (1)
Step 6:         If ($x(t) == 1$), then
Step 7:            Medical feature ‘$f_i$’ is classified as relevant
Step 8:         Else if ($x(t) == -1$), then
Step 9:            Medical feature ‘$f_i$’ is classified as irrelevant
Step 10:        End If
Step 11:        Measure misclassification error ‘$E$’ using (2)
Step 12:        Update weight ‘$\varepsilon_i$’ according to ‘$E$’ using (3)
Step 13:     End While
Step 14:  End For
Step 15:End
```

Algorithm 1. Kernel Perceptron Feature Selection

Algorithm 1 shows the step by step process of KPFS. As demonstrated in the above algorithmic steps, KPFS initially defines the weight and error threshold value. After that, KPFS computes similarities between features and disease symptoms. If the similarity value is '+1' then KPFS categorizes input medical features as relevant to effectively find out the disease occurrence. If the similarity value is '-1' then KPFS categorizes input medical features as irrelevant. Subsequently, KPFS measures misclassification error for each result and consequently weight is updated with respect to error. The above process of KPFS is frequent until the iteration error is minimal. By using the above algorithmic process of KPFS, KPFS-SBPRVM technique provides better feature selection performance to choose the medical features that are significant to diagnosis disease at an early stage with a lower time.

B. Sparse Bayesian Probabilistic Relevance Vector Machine

In KPFS-SBPRVM technique, Sparse Bayesian Probabilistic Relevance Vector Machine Classification (SBPRVMC) model is designed to improve disease diagnosis performance by identifying the occurrence of diseases at an early stage with minimal time complexity. The SBPRVMC Model employs a sparse Bayesian modeling approach which is utilized to obtain parsimonious solutions for disease classification at an early stage. The SBPRVMC Model is proposed in KPFS-SBPRVM Technique is a machine learning technique that performs sparse classification via linear weights of fixed small size functions from a large number of potential candidates. The SBPRVMC Model contains a similar functional form to that of support vector machine (SVM). On the contrary to SVM, SBPRVMC Model includes probabilistic classification property. Compare to SVM, the SBPRVMC Model avoids a set of free parameters that necessitate cross-validation post-optimization. Therefore, SBPRVMC Model gives better classification accuracy for disease diagnosis. The process involved in SBPRVMC Model is depicted in below Figure 2.

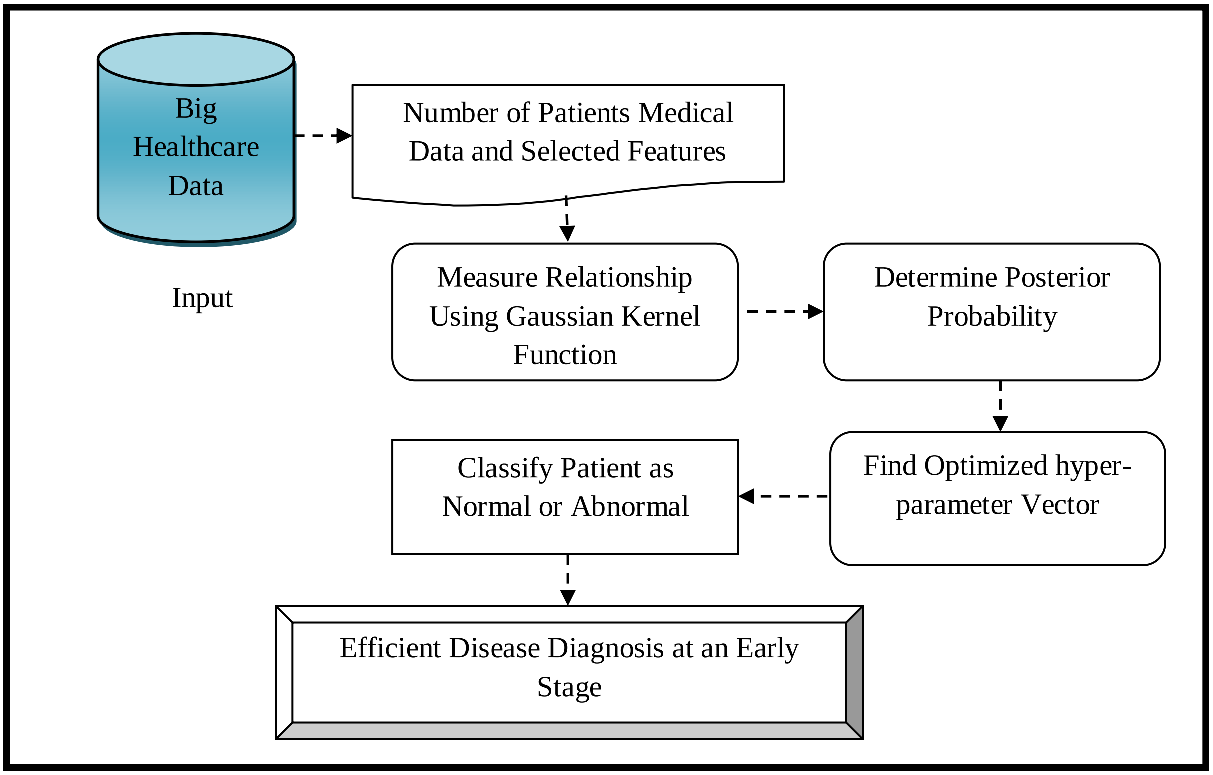

Big Healthcare Data
Input
Number of Patients Medical Data and Selected Features
Measure Relationship Using Gaussian Kernel Function
Determine Posterior Probability
Find Optimized hyper-parameter Vector
Classify Patient as Normal or Abnormal
Efficient Disease Diagnosis at an Early Stage

Figure 2. Flow Process of Sparse Bayesian Probabilistic Relevance Vector Machine Classification Model

Figure 2 presents the block diagram of SBPRVMC Model to achieve better disease diagnosis performance at an early stage. As depicted in the above figure, let us consider a big healthcare dataset includes of many patient medical data with 'n' features denoted as '$DS = \{\beta_1, \beta_2, \beta_3, ..\beta_n\}$'. Here, '$n$' point outs the number of patient medical data in the input big healthcare dataset. In the proposed technique, each medical data of a patient is trained using SBPRVM Model in order to classify the patient as normal or abnormal. The SBPRVMC Model represented as '$\{(\beta_1,\gamma_1),(\beta_2,\gamma_2),\dots(\beta_n,\gamma_n)\}$' where '$\beta_i$ indicates a set of training samples (i.e. input patient medical data) and 'γ_i' refers to the output (diagnosed result). For each training input patient medical data '$\{\beta_i\}_{1=1}^{n}$', '$\{\gamma_i\}_{1=1}^{n}$' gives the corresponding target output value '$\gamma_i \in \{0,1\}$'. From that, SBPRVM classification model is mathematically represented as follows,

$$z(\beta_i,\mu) = \sum_{i=1}^{n} \mu_i\, K(\beta,\beta_i) + \mu_0 \qquad (4)$$

From the above mathematical equations (4), '$K(\beta,\beta_i)$' denotes the kernel function and 'μ_i'indicates the weight of the 'i^{th}' kernel function '$\mu = [\mu_0,\mu_1,\dots\mu_n]^T$' whereas '$\mu_0$' refers to the bias.

In SBPRVMC model, Gaussian kernel function is employed to map '$z(\beta_i,\mu)$' to $(0,1)$ for two class classification i.e. normal class or abnormal class because, the SBPRVMC model returns output as '0' or '1'. Each prediction is independent, the patient medical samples are considered to be independent and identically distributed. To lessen the over-fitting owing to excessive support vectors utilized, the weight vector 'μ' is constrained with the precondition, i.e. all weight vectors satisfy a zero-mean Gaussian prior distribution using below,

$$P(\mu,\vartheta) = \prod_{i=0}^{n} n(\mu_i|0,\vartheta_i^{-1}) = \prod_{i=0}^{n} \frac{\vartheta_i}{\sqrt{2\pi}} exp\left(-\frac{\vartheta_i\mu_i^2}{2}\right)$$

(5)

From the above mathematical representation (5), '$\vartheta = [\vartheta_0,\vartheta_1,\vartheta_2,.\vartheta_n]^T$' represents the hyper-parameter vector which finds the prior distribution of weight vector 'μ' and controls the degree to which the weight deviates from its zero-mean.

Given the prior probability distribution and the likelihood distribution, the Bayes' rule is employed in SBPRVMC Model to estimate the posterior probability of models 'μ' and 'ϑ' using below,

$$P(\mu,\vartheta|t) = P(\mu|t,\vartheta)P(\vartheta|t) \qquad (6)$$

To solve the above mathematical formulation (6), the posterior probability is calculated to ensure the generalization ability. Then, approximation procedure is employed in SBPRVMC Model using Laplace's method which finds maximum hyper-parameter vector 'μ' using below,

$$\mu_{MP}^{new} = \mu_i + \Delta\mu \qquad (7)$$

$$\vartheta_{MP}^{new} = \frac{1-\vartheta_i\Sigma_{ij}}{\mu_{MP}^2} \qquad (8)$$

From the above mathematical expressions (7) and (8), optimized hyper-parameter vector 'μ' is determined in SBPRVMC Model to efficiently classify the each input patient medical data as normal or abnormal. The output of SBPRVMC model is shown in below.

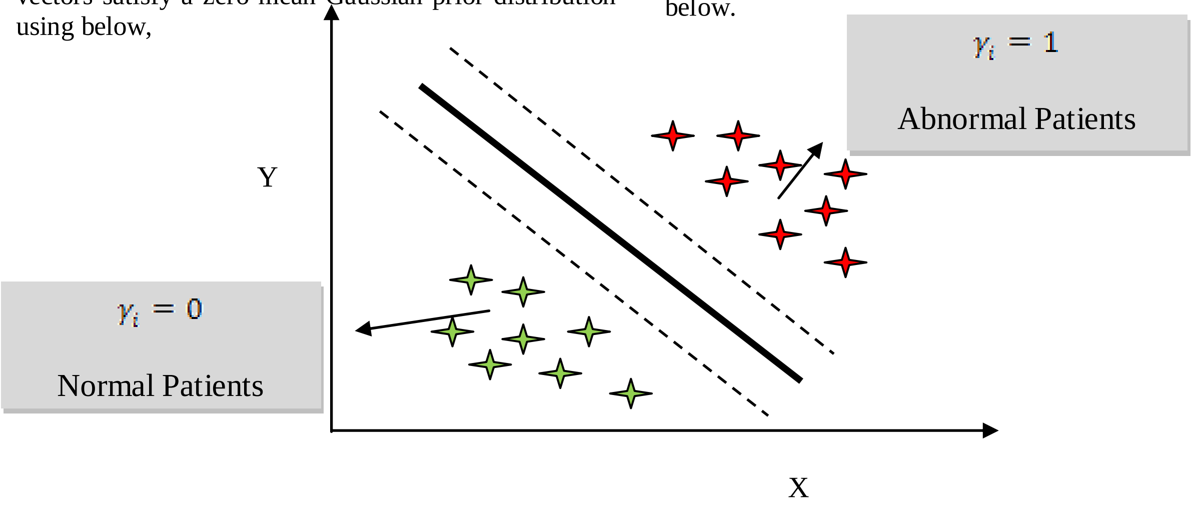


Figure 3. Output of SBPRVMC Model

Figure 3 presents the classification result of SBPRVMC model. The output of SBPRVMC Model is '$\gamma_i \in \{0,1\}$'. In SBPRVMC model, the result '$\gamma_i = 0$' denotes that a patient is normal whereas '$Y_i = 1$' represents that the patient is abnormal. The algorithmic processes of SBPRVMC model is explained below,

```
// Sparse Bayesian Probabilistic Relevance Vector Machine Classification Algorithm
Input: Number Of Patient Medical Data '$DS = \beta_1, \beta_2, ..\beta_n$'.
Output: Obtain higher disease diagnosis rate
Step 1:Begin
Step 2:    For each patient medical data '$\beta_i$' with selected relevant medical features
Step 3:        Apply SBPRVM classification model using (4)
Step 4:        Measure the relationship between patient data and disease features using (5)
Step 5:        Calculate Posterior Probability to ensure the generalization ability using (6)
Step 6:        Discover optimized hyper-parameter vector using (7)
Step 7:        Classify patient medical data '$\beta_i$' as normal or abnormal
Step 8:        Get disease diagnosed result at an early stage
Step 9:    End For
Step 10:End
```

Algorithm 2. Sparse Bayesian Probabilistic Relevance Vector Machine Classification

Algorithm 2 portrays the step by step processes of SBPRVMC model. As shown in the above algorithmic process, SBPRVMC model initially evaluates the relationship between patient data and disease features. After that, SBPRVMC model measure posterior probability to make sure the generalization ability. Followed by, SBPRVMC model determines optimized hyper-parameter vector. By using the identified optimized hyper-parameter vector, finally SBPRVMC model efficiently classifies all the patient medical data with a minimal amount of time consumption. From that, KPFS-SBPRVM technique obtains better disease diagnosed performance to find out the presence of disease at an early stage as compared to conventional works.

III. Experimental Settings

To validate the proposed performance, KPFS-SBPRVM technique is implemented in Java Language with help of two big healthcare dataset namely Diabetes 130-US hospitals for years 1999-2008 Data Set [22] and Epileptic Seizure Recognition Data Set [23]. The diabetic dataset contains 50 features and 100000 instances. This diabetic dataset includes features as patient number, race, gender, age, admission type, and time in hospital, medical specialty of admitting physician, and number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, and number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc. An Epileptic Seizure Recognition Data Set includes of 11500 with 179 features.

In order to accomplish the experimental evaluation, KPFS-SBPRVM technique obtains 50-500 patients medical data from both datasets. The performance of KPFS-SBPRVM is evaluated in terms of feature selection accuracy, disease diagnosis rate, disease diagnosis time and false positive rate with respect to various numbers of features and patient medical data. The experimental result of KPFS-SBPRVM technique is compared against with globally optimized Artificial Neural Network Input Gain Measurement Approximation (GANNIGMA-ensemble) technique [1] and Support Vector Machine (SVM) [2].

IV. Results

In this section, the experimental result of KPFS-SBPRVM technique is discussed. The performance of KPFS-SBPRVM technique is compared against with GANNIGMA-ensemble technique [1] and SVM [2] respectively. The performance of KPFS-SBPRVM technique is determined along with the following metrics with the help of tables and graphs.

A. Feature Selection Accuracy

In KPFS-SBPRVM technique**,** Feature Selection Accuracy 'FSA' computed as ratio of number of medical features precisely selected as relevant or irrelevant to the total number of features. The feature selection accuracy is calculated in terms of percentage (%) and obtained mathematically as follows,

$$FSA = \frac{\tau_{CS}}{n} * 100 \qquad (9)$$

From the above mathematical equation (9), feature selection accuracy for both diabetic and brain tumor disease identification is estimated. Here, 'n' refers to the total number of medical features taken for experimental work whereas 'τ_{CS}' signifies the number of medical features correctly selected.

Sample Calculation for Feature Selection Accuracy of Diabetic Disease:

- **Existing GANNIGMA-ensemble technique:** number of medical features accurately chosen is 2 and the total number of features is 5. Then feature selection accuracy is evaluated as follows,

$$FSA = \frac{2}{5} * 100 = 40\ \%$$

- **Existing SVM:** number of medical features exactly selected is 3 and the total number of features is 5. Then feature selection accuracy is computed as follows,

$$FSA = \frac{3}{5} * 100 = 60\ \%$$

- **Proposed KPFS-SBPRVM technique:** number of medical features correctly selected is 4 and the total number of features is 5. Then feature selection accuracy is obtained as,

$$FSA = \frac{4}{5} * 100 = 80\ \%$$

Table 1. Tabulation for Feature Selection Accuracy of Diabetic Disease

Number of Medical Features (n)	Feature Selection Accuracy (%)		
	GANNIGMA-ensemble technique	SVM	KPFS-SBPRVM technique
5	40	60	80
10	50	70	90
15	60	73	87
20	65	75	85
25	72	80	88
30	77	83	90
35	77	83	89
40	83	88	93
45	82	87	91
50	40	60	80

Table 2. Tabulation for Feature Selection Accuracy of Brain Tumor Disease

Number of Medical Features (n)	Feature Selection Accuracy (%)		
	GANNIGMA-ensemble technique	SVM	KPFS-SBPRVM technique
15	67	73	87
30	70	77	90
45	73	78	91
60	80	83	92
75	84	87	93
90	81	83	90
105	87	89	94
120	89	91	93
135	85	87	90
150	89	90	95

Table 1 and 2 depicts the tabulation result of feature selection accuracy for both diabetic and brain tumor disease with respect to a diverse number of medical features using three methods namely GANNIGMA-ensemble technique [1], SVM [2] and proposed KPFS-SBPRVM technique. When considering the 120 medical features from Epileptic Seizure Recognition Data Set to carry out the experimental process, proposed KPFS-SBPRVM technique gets 93 % feature selection accuracy whereas conventional GANNIGMA-ensemble technique [1], SVM [2] obtains 89 % and 91 % respectively. Hence, feature selection accuracy using proposed KPFS-SBPRVM technique is higher as compared to other works [1] and [2].

This is because of the application of KPFS in KPFS-SBPRVM technique. On the contrary to conventional works, the KPFS iteratively enhances feature selection performance through running it on training samples, subsequently updating the model whenever it discovers it has made an inaccurate classification. This assists for KPFS-SBPRVM technique to improve the ratio of a number of medical features accurately selected when compared to conventional works. Therefore, the proposed KPFS-SBPRVM technique enhances feature selection accuracy of diabetic disease by 34 % and 13 % when compared to the GANNIGMA-ensemble technique [1], SVM [2] respectively. Thus, the proposed KPFS-SBPRVM technique increases feature selection accuracy of brain tumor disease by 15 % and 10 % when compared to the GANNIGMA-ensemble technique [1], SVM [2] respectively.

B. Disease Diagnosis Rate

In KPFS-SBPRVM technique**,** Disease Diagnosis Rate 'DDR' determines ratio of a number of patients medical data exactly classified as normal or abnormal to the total number of patient's medical data. The disease diagnosis rate is evaluated in terms of percentages (%) and calculated mathematically as follows,

$$DDR = \frac{\tau_{AC}}{n} * 100 \quad (10)$$

From the above mathematical formulation (10), the diagnosis rate of both diabetic and brain tumor disease is measured. Here, 'n' point outs to the total number of the patients medical data considered for experimental evaluation whereas 'τ_{AC}' refers to a number of patient accurately classified.

Sample Calculation for Disease Diagnosis Rate of Brain Tumor

- **Existing GANNIGMA-ensemble technique:** number of patients medical data correctly classified is 37 and the total number of patient's medical data is 50. Then the disease diagnosis rate is calculated as follows,

$$DDR = \frac{37}{50} * 100 = 74\ \%$$

- **Existing SVM:** number of patients medical data precisely classified is 40 and the total number of patient's medical data is 50. Then disease diagnosis rate is acquired as follows,

$$DDR = \frac{40}{50} * 100 = 80\ \%$$

- **Proposed KPFS-SBPRVM technique:** number of patients medical data perfectly classified is 44 and the total number of patients medical data is 50. Then disease diagnosis rate is determined as follows,

$$DDR = \frac{44}{50} * 100 = 88\%$$

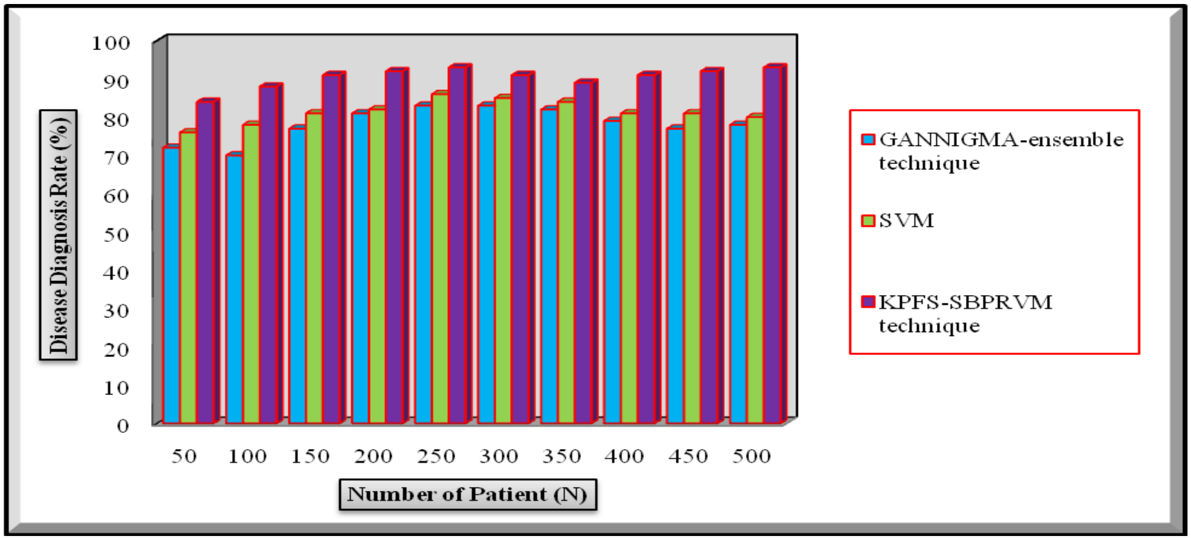


Figure 4 Experimental result of Diagnosis Rate for Diabetic Disease

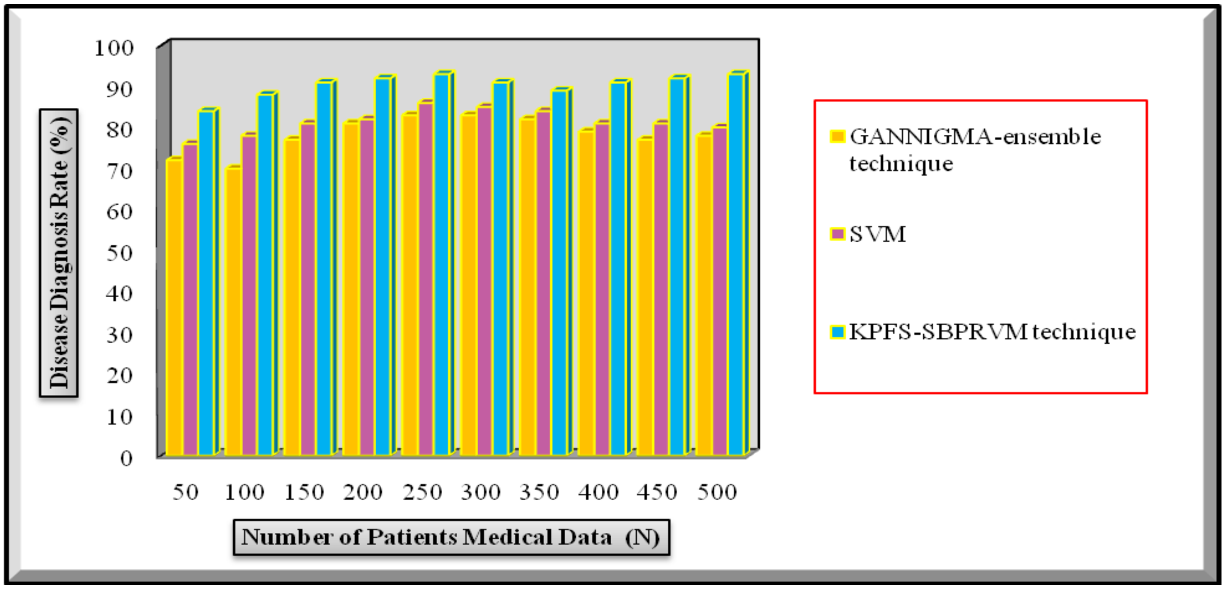


Figure 5. Experimental Result of Diagnosis Rate for Brain Tumor Disease

Figure 4 and 5 show the experimental result of diagnosis rate for both diabetic and brain tumor disease based on a varied number of patient's medical data using three methods namely GANNIGMA-ensemble technique GANNIGMA-ensemble technique [1], SVM [2] and proposed KPFS-SBPRVM technique. When acquiring 450 patient's medical data from Diabetes 130-US hospitals for years 1999-2008 Data Set to accomplish experimental evaluation, proposed KPFS-SBPRVM technique attains 92 % disease diagnosis rate whereas state-of-the-art works GANNIGMA-ensemble technique [1], SVM [2] gets 77 % and 81 % respectively. From that, disease diagnosis rate using proposed KPFS-SBPRVM technique is very higher when compared to other works [1] and [2].

This is owing to the application of SBPRVMC algorithm in KPFS-SBPRVM technique. On the contrary to state-of-the-art works, the SBPRVMC model considers a Bayesian formulation of a linear model with an appropriate prior which results in a sparse representation. This helps for KPFS-SBPRVM technique to exactly classify the patient's data based on selected medical features as normal or abnormal. This supports for KPFS-SBPRVM technique to increases the ratio of a number of patients medical data accurately classified when compared to conventional works. As a result, the proposed KPFS-SBPRVM technique improves diagnosis rate of diabetic disease by 16 % and 11 % when compared to the GANNIGMA-ensemble technique [1], SVM [2] respectively. Hence, the proposed KPFS-SBPRVM technique enhances the diagnosis rate of brain tumor disease by 17 % and 12 % as compared to existing GANNIGMA-ensemble technique [1], SVM [2] respectively.

C. Disease Diagnosis Time

In KPFS-SBPRVM technique, Disease Diagnosis Time '(DDT)' calculate the amount of time utilized to identify the presence and absence of disease through classifying patient medical data. The disease diagnosis time is mathematically evaluated in terms of milliseconds (ms) and measured mathematically as follows,

$$DDT = n * T(CSPMD) \tag{11}$$

From the above mathematical expressions (11), the diagnosis time of both diabetic and brain tumor disease is determined. Here, 'n' indicates a total number of patient's medical data whereas '$T(CSPMD)$' point outs the time taken to classify single patient medical data.

Sample Calculation for Diagnosis Time of Diabetic Disease

- **Existing GANNIGMA-ensemble technique:** the time consumed to classify one patient medical data is 0.33 ms and the total number of patient medical data is 50. Then, diseases diagnosis time is measured as follows,

$$DDT = 50 * 0.33 = 23 \ ms$$

- **Existing SVM:** the time desired to classify medical data of a single patient is 0.38 ms and the total number of patient's medical data is 50. Then disease diagnosis time is computed as follows,

$$DDT = 50 * 0.38 = 19\ ms$$

- **Proposed KPFS-SBPRVM technique:** the time used to classify single patient medical data is 0.31 ms and the total number of patients medical data is 50. Then disease diagnosis time is evaluated as follows,

$$DDT = 50 * 0.31 = 16\ ms$$

Table 3. Tabulation for Disease Diagnosis Time

Number of Patients Medical Data (n)	Disease Diagnosis Time (ms)					
	Diabetic Disease			Brain Tumor Disease		
	GANNIGMA-ensemble technique	SVM	KPFS-SBPRVM technique	GANNIGMA-ensemble technique	SVM	KPFS-SBPRVM technique
50	23	19	16	21	18	15
100	35	30	20	32	27	18
150	50	44	29	45	39	26
200	64	54	36	58	48	32
250	75	68	43	68	60	38
300	78	75	48	72	66	42
350	88	81	53	77	70	46
400	96	88	56	84	76	48
450	104	95	59	90	81	50
500	110	100	65	95	85	55

Table 3 illustrates the performance result of diagnosis time for both diabetic and brain tumor disease along with a diverse number of patient's medical data using three methods namely GANNIGMA-ensemble technique GANNIGMA-ensemble technique [1], SVM [2] and proposed KPFS-SBPRVM technique. When taking 300 patient's medical data from Epileptic Seizure Recognition Data Set to conduct experimental work, proposed KPFS-SBPRVM technique acquires 42 ms disease diagnosis time whereas existing works GANNIGMA-ensemble technique [1], SVM [2] consumes 72 ms and 66 ms respectively. Accordingly, disease diagnosis time using proposed KPFS-SBPRVM technique is minimal as compared to other works [1] and [2].

This is due to the application of and KPFS and SBPRVMC algorithms in KPFS-SBPRVM technique on the contrary to state-of-the-art works. By using the algorithmic processes of KPFS, proposed KPFS-SBPRVM technique choose the medical features that are more significant for disease diagnosis with lower time consumption. Depends on the selected features, then proposed KPFS-SBPRVM technique discovers the patient as normal or abnormal through performing classification with minimal time complexity. This aid for proposed KPFS-SBPRVM technique to decreases the amount of time consumed to identify the presence and absence of disease when compared to conventional works. Therefore, the proposed KPFS-SBPRVM technique reduces diagnosis time of diabetic disease by 24 % and 40 % when compared to the GANNIGMA-ensemble technique [1], SVM [2] respectively. Hence, the proposed KPFS-SBPRVM technique minimizes diagnosis time of brain tumor disease by 41 % and 33 % when compared to the GANNIGMA-ensemble technique [1], SVM [2] respectively.

D. False Positive Rate

In KPFS-SBPRVM technique**,** False Positive Rate '(FPR)' computes ratio of number of patients medical data mistakenly classified as normal or abnormal to the total number of patients medical data. The false positive rate is measured using below mathematical expression,

$$FPR = \frac{\tau IC}{n} * 100 \qquad (12)$$

From the above mathematical representations (12), the false-positive rate of both diabetic and tumor disease diagnosis is calculated. Here, 'τ_{IC}' designates the patient's medical data incorrectly classified. The false positive rate is estimated in terms of percentages (%).

Sample Calculation for False Positive Rate of Brain Tumor Disease

- **Existing GANNIGMA-ensemble technique:** number of patient's medical data wrongly classified is 14 and the total number of patient's medical data is 50. Then false positive rate is determined as follows,

$$FPR = \frac{14}{50} * 100 = 28\ \%$$

- **Existing SVM:** number of patient's medical data inaccurately classified is 12 and the total number of patient's medical data is 50. Then false positive rate is obtained as follows,

$$FPR = \frac{12}{50} * 100 = 24\ \%$$

- **Proposed KPFS-SBPRVM technique:** number of patient's medical data imperfectly classified is 8 and the total number of patient's medical data is 50. Then false positive rate is evaluated as follows,

$$FPR = \frac{8}{50} * 100 = 16\ \%$$

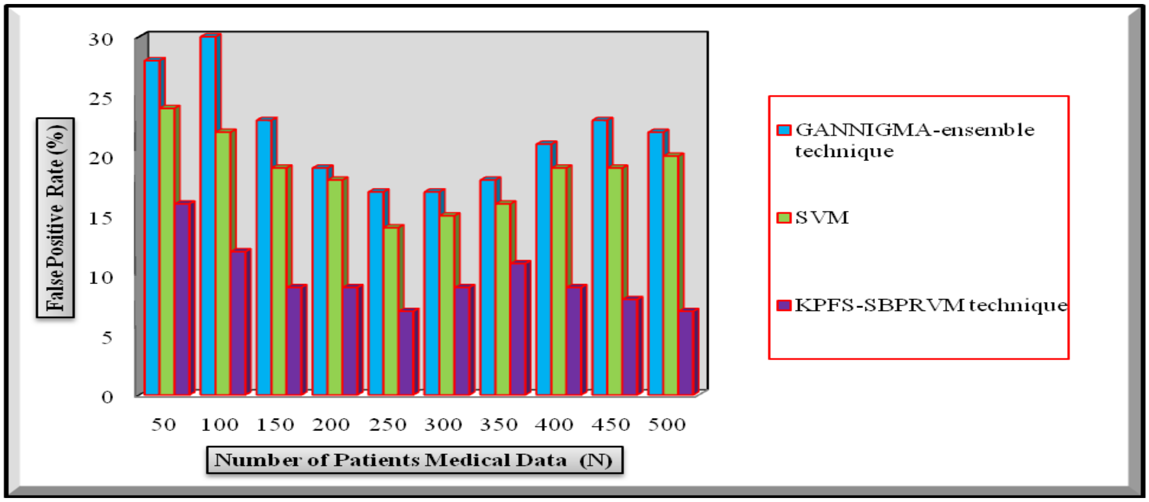


Figure 6 Experimental Result of False Positive Rate for Diabetic Disease Diagnosis

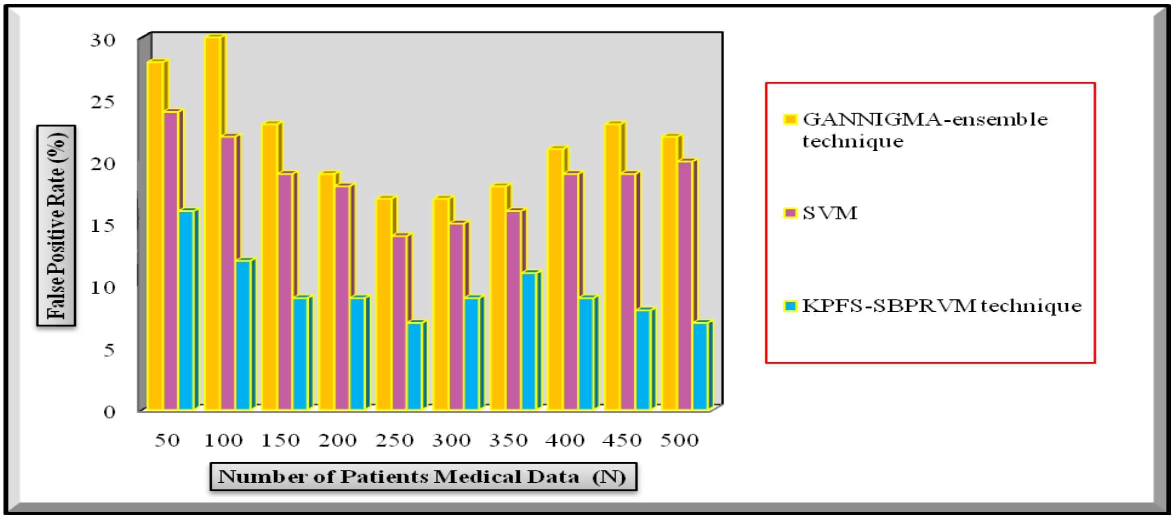


Figure 7. Experimental Result of False Positive Rate for Brain Tumor Disease Diagnosis

Figure 6 and 7 shows the comparative result of false-positive rate for both diabetic and brain tumor disease diagnosis based on a different number of patient's medical data using three methods namely GANNIGMA-ensemble technique GANNIGMA-ensemble technique [1], SVM [2] and proposed KPFS-SBPRVM technique. When getting 400 patient's medical data from Diabetes 130-US hospitals for years 1999-2008 Data Set to carry out the experimental evaluation, proposed KPFS-SBPRVM technique attains 9 % false-positive rate whereas existing works GANNIGMA-ensemble technique [1], SVM [2] obtains 21 % and 19 % respectively. Hence, false-positive rate using the proposed KPFS-SBPRVM technique is minimal when compared to other works [1] and [2].

This is because of the application of KPFS and SBPRVMC algorithms in KPFS-SBPRVM technique on the contrary to conventional works. With the application of KPFS algorithmic processes, proposed KPFS-SBPRVM technique finds the medical features that are more relevant for accurately performing disease diagnosis. Followed by, proposed KPFS-SBPRVM technique exactly find outs the patient as normal or abnormal via classification. This assist for proposed KPFS-SBPRVM technique to decreases the amount of time consumed to identify the presence and absence of disease when compared to conventional works. Therefore, the proposed KPFS-SBPRVM technique decreases the false positive rate of diabetic disease by 55 % and 48 % as compared to GANNIGMA-ensemble technique [1], SVM [2] respectively. As a result, the proposed KPFS-SBPRVM technique lessens the false positive rate of brain tumor disease by 61 % and 54 % when compared to the GANNIGMA-ensemble technique [1], SVM [2] respectively.

V. Literature Survey

Ensembles of Neuro-Fuzzy Inference System was presented in [11] for discovering the hepatitis disease with higher accuracy through classification. Ensembles of simple convolutional neural networks (CNNs) and softmax cross-entropy classifier was employed in [12] for diagnosis of Alzheimer diseases. A survey of various Feature selection and classification systems developed for chronic disease prediction was analyzed in [13] to identify the disease at an early stage. Machine learning and Map Reduce algorithm were utilized in [14] for effective identification of various disease presence in disease-frequent societies.

Multilayered Probabilistic Neural Network classifier was introduced in [15] for automatic classification of an abnormal heartbeat from ECG big data. Convolutional neural network (CNN)-based multimodal disease risk prediction algorithm was designed in [16] using structured and unstructured data from the hospital.

Computer-aided diagnosis (CAD) was presented in [17] to find the neurological abnormalities using the medical big data and thereby enhances consistency of diagnosis, the success of treatment. The greedy deep weighted dictionary learning was developed in [18] to get better classification performance for medical diseases analysis.

A hybrid associative classifier (Clasificador Híbrido Asociativo con Traslacion) *CHAT* was applied in [19] for performing medical data classification process with a minimal error. An Effective Fuzzy Rule Classifier (EFRC)–based decision support system was designed in [20] for identification of disease of patients. Machine Learning Techniques developed for classification of health care data was presented in [21].

VI. Conclusion

An effective KPFS-SBPRVM technique is developed with the goal of improving disease diagnosis performance at an early stage through classification. The objective of KPFS-SBPRVM technique is achieved with the help of KPFS and SBPRVMC algorithms. The proposed KPFS-SBPRVM technique enhance the ratio of a number of patients medical data accurately classified with assists of SBPRVMC algorithm as compared to conventional works. Moreover, the proposed KPFS-SBPRVM technique increases the ratio of a number of medical features precisely chosen when compared to state-of-the-art works. Also, the proposed KPFS-SBPRVM technique minimizes the number of patient medical data wrongly classified and time required for efficient disease diagnosis as compared to conventional works. The efficiency of KPFS-SBPRVM technique is determined in term of feature selection accuracy, disease diagnosis rate, disease diagnosis time and false positive rate with respect to different numbers of features and patient medical data and compared against two existing methods. The experimental results show that the proposed KPFS-SBPRVM technique gives better disease diagnosis performance in terms of feature selection accuracy and disease diagnosis accuracy, disease diagnosis time and false positive rate when compared to state-of-the-art works.

References

[1] Shamsul Huda; John Yearwood; Herbert F. Jelinek; Mohammad Mehedi Hassan; Giancarlo Fortino; Michael Buckland, "*A Hybrid Feature Selection With Ensemble Classification for Imbalanced Healthcare Data: A Case Study for Brain Tumor Diagnosis*", IEEE Access, Volume 4 , Pages: 9145 – 9154, 2016.

[2] N. Sneha, Tarun Gangil, "*Analysis of diabetes mellitus for early prediction using optimal features selection*", Journal of Big Data, Springer, Volume 6, Issue 13, December 2019.

[3] Yichuan Wang, LeeAnn Kung, William Yu Chung Wang, Casey G. Cegielski, "*An integrated big data analytics-enabled transformation model: Application to health care*", Information & Management, Elsevier, Volume 55, Issue 1, January 2018, Pages 64-79.

[4] Min Chen, Jun Yang, Jiehan Zhou, Yixue Hao, Jing Zhang, and Chan-Hyun Youn, "*5G-Smart Diabetes: Toward Personalized Diabetes Diagnosis with Healthcare Big Data Clouds",* IEEE Communications Magazine, Volume 56, Issue 4, April 2018, Pages 16 – 23.

[5] Jie Xu, Li Wang, Yunfeng Shen, Kaifen Yuan, Yue Nie, Yingxuan Tian, Xiangdong Jian, Xing Ma, and Jinhong Guo, "*Family-based Big Medical-Level Data Acquisition System*", IEEE Transactions on Industrial Informatics, Volume 15, Issue 4, April 2019, Pages 2321 – 2329.

[6] C. B. Sivaparthipan, N. Karthikeyan and S. Karthik, "*Designing statistical assessment healthcare information system for diabetics analysis using big data*", Multimedia Tools and Applications, Springer, November 2018, Pages 1–14.

[7] Dimitrios Kollias, Athanasios Tagaris, Andreas Stafylopatis, Stefanos Kollias, Georgios Tagaris, *"Deep neural architectures for prediction in healthcare"*, Complex & Intelligent Systems, Springer, Volume 4, Issue 2, Pages 119–131, June 2018.

[8] JafarA.ALzubi, Balasubramaniyan Bharathikannan, Sudeep Tanwar, Ramachandran Manikandan, Ashish Khanna, Chandrasekar Thaventhiran, "*Boosted neural network ensemble classification for lung cancer disease diagnosis*", Applied Soft Computing, Elsevier, Volume 80, Pages 579-591, July 2019.

[9] Ankita Sharma, Deepika Shukla, Tripti Goel, and Pravat Kumar Mandal, "*BHARAT: An Integrated Big Data Analytic Model for Early Diagnostic Biomarker of Alzheimer's disease*", Frontiers in Neurology, Volume 10, Article 9, Pages 1-7, February 2019.

[10] Diellza Nagavci, Mentor Hamiti, Besnik Selimi, "*Review of Prediction of Disease Trends using Big Data Analytics*", International Journal of Advanced Computer Science and Applications, Volume 9, Issue 8, Pages 46-50, 2018.

[11] Mehrbakhsh Nilashi, Hossein Ahmadi, Leila Shahmoradi, Othman Ibrahim, Elnaz Akbari, "*A predictive method for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique"*, Journal of Infection and Public Health, Elsevier, Volume 12, Issue 1, Pages 13-20, January–February 2019.

[12] Samsuddin Ahmed, Kyu Yeong Choi, Jang Jae Lee, Byeong C. Kim, Goo-Rak Kwon, "*Ensembles of Patch-Based Classifiers for Diagnosis of Alzheimer Diseases*", IEEE Access, Volume 7, Pages 73373 – 73383, May 2019.

[13] Divya Jain, Vijendra Sing, "*Feature selection and classification systems for chronic disease prediction: A review"*, Egyptian Informatics Journal, Elsevier, Volume 19, Issue 3, Pages 179-189, November 2018.

[14] Vinitha S, Sweetlin S, Vinusha H and Sajini S, "*Disease Prediction Using Machine Learning Over Big Data*", Computer Science & Engineering: An International Journal (CSEIJ), Volume 8, Issue 1, Pages 1-8, February 2018.

[15] Hari Mohan, RaiKalyan Chatterjee, "*A unique feature extraction using MRDWT for automatic classification of abnormal heartbeat from ECG big data with Multilayered Probabilistic Neural Network classifier*", Applied Soft Computing, Volume 72, Pages 596-608, November 2018.

[16] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, Lin Wang, "*Disease Prediction by Machine Learning Over Big Data from Healthcare Communities*", IEEE Access, Volume 5, Pages 8869 – 8879, April 2017.

[17] Siuly Siuly, "*Medical Big Data: Neurological Diseases Diagnosis through Medical Data Analysis*", Data Science and Engineering, Springer, Volume 1, Issue 2, Pages 54–64, June 2016.

[18] Chunxue Wu, Chong Luo, Naixue Xiong, Wei Zhang, Tai-Hoon Kim, "*A Greedy Deep Learning Method for Medical Disease Analysis*", IEEE Access, Volume 6, Pages 20021 – 20030, April 2018.

[19] Abril Valeria Uriarte-Arcia ,Itzamá López-Yáñez, Cornelio Yáñez-Márquez, *"One-Hot Vector Hybrid Associative Classifier for Medical Data Classification*", PLoS ONE, Volume 9, Issue 4, Pages 1-13, April 2014.

[20] Mallikarjun M. Kodabagi, Ahelam Tikotikar, "*Clustering-based approach for medical data classification*", Concurrency and Computation Practice and Experience, Wiley Online Library, Volume 31, Issue 14, Pages 1-14, 2018.

[21] H.S. Hota, Seema Dewangan, "*Classification of Health Care Data Using Machine Learning Technique*", International journal of Engineering Science Invention, Volume 5, Issue 9, Pages 2319 – 6734, September 2016.

[22] Diabetes 130-US hospitals for years 1999-2008 Data Set: https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008.

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

[24] Murugesan C and Marimuthu C.N "*Cost optimization of PV-diesel systems in Nano grid using LJ cuckoo search and its application in mobile towers*", Mobile Networks and Applications Volume 24, Issue 2, Pages 340-349, April 2019.

[25] Deepa A and Marimuthu C.N "*Design of a high speed Vedic multiplier and square architecture based on Yavadunam Sutra*", Sādhanā 44 (9), 197 .