AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes

V. Jackins[1] · S. Vimal[1] · M. Kaliappan[2] · Mi Young Lee[3]

Abstract

Healthcare practices include collecting all kinds of patient data which would help the doctor correctly diagnose the health condition of the patient. These data could be simple symptoms observed by the subject, initial diagnosis by a physician or a detailed test result from a laboratory. Thus, these data are only utilized for analysis by a doctor who then ascertains the disease using his/her personal medical expertise. The artificial intelligence has been used with Naive Bayes classification and random forest classification algorithm to classify many disease datasets like diabetes, heart disease, and cancer to check whether the patient is affected by that disease or not. A performance analysis of the disease data for both algorithms is calculated and compared. The results of the simulations show the effectiveness of the classification techniques on a dataset, as well as the nature and complexity of the dataset used.

Keywords Artificial intelligence · Diabetes · Data mining techniques · Naïve Bayes classification · Random forest classification

✉ Mi Young Lee
  miylee@sejong.ac.kr

  V. Jackins
  jackin.silves@gmail.com

  S. Vimal
  svimalphd@gmail.com

  M. Kaliappan
  kalsrajan@yahoo.co.in

[1] Department of IT, National Engineering College, Kovilpatti, Tamil Nadu, India

[2] Department of Computer Science and Engineering, Ramco Institute of Technology, Rajapalayam, Tamilnadu, India

[3] Department of Software, Sejong University, Seoul, South Korea

1 Introduction

Due to modern lifestyle, diseases are increasing rapidly. Our lifestyle and food habit leads to create impact on our health causing heart diseases and other health issues. Data mining technique is one of the most challenging and leading research areas in healthcare due to the high importance of valuable data [1]. The recent blooming in the data mining approaches has provided a solid platform for various applications in the healthcare field. In healthcare, data mining is playing a vital role in different fields like intrusion detection, pattern recognition, cheaper medical treatments' availability for the patients, disease diagnosing and finding its procurement methods [2, 3]. An artificial intelligence makes the system more sensitive and activates the system to think. In machine learning, AI acts as a subfield to perform better prediction [4].

It also accommodates the researchers in the field of healthcare in development of effective policies, and different systems to prevent different types of disease, early detection of diseases can reduce the risk factor. The aim of our work is to predict the diseases among the trained dataset using classification algorithms. It has been trained the Naive Bayes and random forest classifier model with three different disease datasets namely—diabetes, coronary heart disease and cancer datasets and performance of each model are calculated. Over-fitting of single decision tree problem is overcome by applying the Random forest algorithm [5]. Random forest algorithm provides better prediction accuracy compared with the Naïve Bayes algorithm. In addition, it has been applied few sample test data of the three diseases to those classified models to show whether the patient data in sample test are suffering from that disease or not [6].

Artificial neural networks are the best effort classification algorithm for prediction of medical diagnosis due to its best efficiency parameter [7]. The neural network comprises of the neurons with three layers such as input layer, hidden layer and output layer for the efficiency attainment. The training data are given as the input parameter with the support of back propagation algorithm. The feed-forward neural network with support vector machine (SVM) is a best technique for prediction of cancer [8]. The ANN is used to classify the labeled images based on the determination of the true positive (TP) and false-positive (FP) detection rates. The detection mechanism is performed with the self-organized supervised learning algorithm. ANN approach gives the promising result for the detection of micro-calcifications features and biopsy detection [9]. The ANN is segmented into two approaches, initially the classifier is applied to the image data with region of interest (ROI) and second includes the ANN learn the features from pre-processed image signals. SVM is a statistical learning theory-based machine learning approach. SVM works with the ANN to map the input space to the higher-dimensional space to split the labeled images. The labeled images are determined in a marginal space forming a hyper plane which reduces the generalization error [10].

A hybrid classifier is proposed here by hybridizing support vector machine and artificial neural network [7]. A typical ANN consists of one input layer, one or
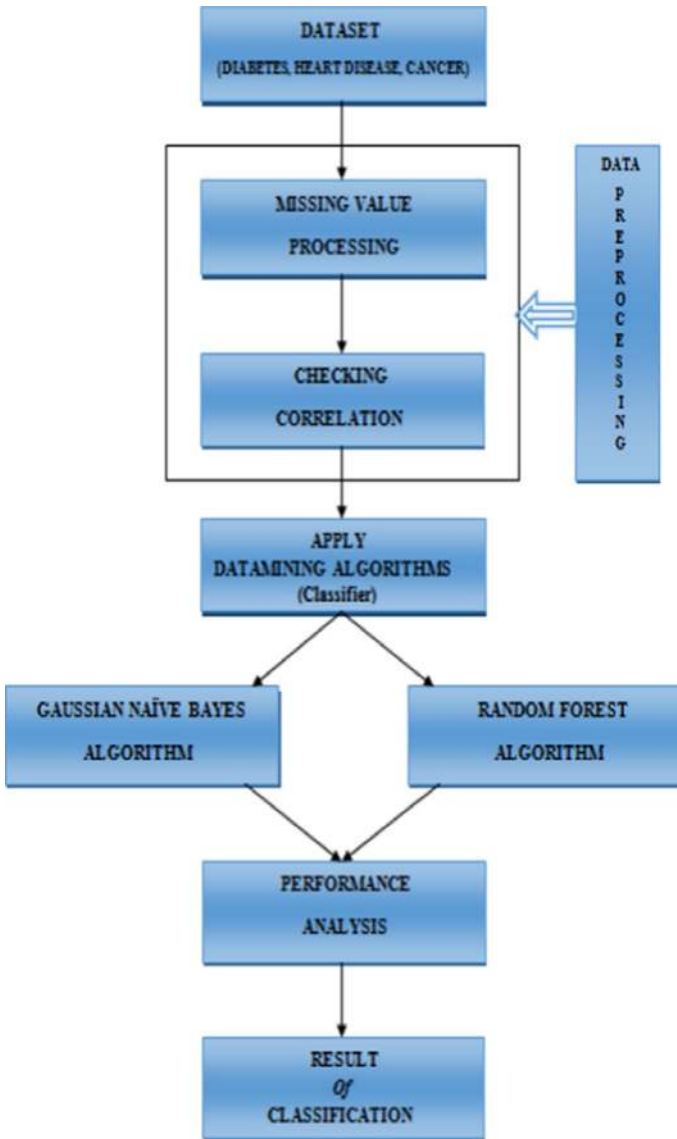
Fig. 1 Block diagram for proposed method

more hidden layers and one output layer as in Fig. 1. Each layer has several neurons, and the neurons in one layer are connected to the neurons in the adjacent layer with its own connection weights [11]. The following figure shows the artificial neural network model with one input layer, one output layer and one hidden layer. The neuron represents node in the network. The input features are fed to the neurons in the input layer.

SVM is the supervised learning model, which can perform well even with the smaller data samples [12, 13]. SVM classifier has no curse of dimensionality since it has the ability to manage sparse data in high-dimensional datasets [14]. Also, SVM classifier has better generalization than the ANN and avoid local minima problem. The logic behind the SVM method is it creates the optimal separation plane under linearly separable condition [12]. The hyper plane is optimized by increasing the margin. The margin is the distance between boundary and nearest point of each class. These points nearest to the boundary are called support vectors [15]. For healthcare data analysis, data mining techniques like classification, association rules and clustering are commonly used.

The main contribution of this paper is summarized as follows.

- Initially disease dataset is taken as a input for the system. Diabetes, heart disease and cancer datasets are taken for the analysis, so many informations are related to the patients health care and general data there in disease dataset. These are the life-threading diseases for human beings.
- Data preprocessing is applied to the input datasets, it leads to reduce the unwanted information for the further analysis. Check the missing values and checking the correlation it helps to split the training data in to 70% original data and 30% testing data, for efficient data analysis.
- Data-mining algorithms like random forest and Gaussian Naïve Bayes are applied to estimate the performance of the system against the input disease dataset. The classification results are compared with existing results, and it shows the better improvement.

The paper is organized as follows, Sect. 1 represents the introduction, Sect. 2 represents the literature survey, while Sect. 3 proposes the proposed method, Sect. 4 represents the results and discussions, Sect. 5 proposes performance evaluation, Sect. 6 identifies the proposed metrics with random forest and the final Sect. 6 proposes the conclusion with references.

2 Literature survey

Data mining is a growing field that transforms piece of data into useful information. This technique helps the authorized person make informed options and take right decisions for their betterment [16]. It used to understand, predict and guide future behavior based on the hidden patterns among huge dataset. It leads to offer tools for automated learning from the history of data and developing models to discover the outcomes of future scenarios. There are the various tools for data mining machine learning algorithms to identify and predict the various disease in terms of regression, decision tree and Bayesian network [17]. Finding of a disease, needed different test results in variety of scenarios with respect to the particular patient. By applying data mining, concept for data analysis number of tests will be reduced. It plays a vital role in data analysis to improve the performance and time saving [18].

Variety of classification and clustering algorithms plays a significant role for prediction and diagnosis of different types of diseases. Bayesian network classifier and random forest classifier are used to diagnose the risk for diabetes [10, 19, 20]. The prediction accuracy of the k-means algorithm is enhanced using both class and cluster method and making it adapt to different datasets [21]. A group of classification algorithms excluding random forest algorithm is applied on diabetes data to diagnose the risk. On comparing the performance of each method, the outcome shows that Random Forest was performed well in both accuracy and ROC curve [8, 22, 23].

In ANN hybrid classifier, each neuron of the hidden and output layer receives signals from the previous layer multiplied with the weights of the interconnection. The neuron then produces the output by passing the summed signal through the defined transfer function.

The network is trained for the given input iteratively. In each iteration, the mean square error (MSE) between the target and the achieved output is calculated. The MSE for the *j*th iteration is defined in Eq. (1) as follows,

$$F(X) = \left(t_i - a_i\right)^2 \tag{1}$$

t_j and a_j are the targeted and the achieved output, respectively. The network is trained by adjusting the weights and the bias so that the MSE get minimized. The MSE estimates the posterior probability function for the classification problem. Here in the gradient descent method, back-propagation (BP) method uses the calculated MSE at each layer to adjust the value of the interconnected weights. Though ANNs are good classifiers, they require large number of training sets to train for proper behavior. That is why it founds fine if hybridized with the classifiers which could require a smaller number of training samples to classify properly. Neural network-based cancer classifiers are used with binary and multi-class problems to identify the cancerous samples [8].

Applied Naïve Bayes algorithm developing an artificial intelligent system, based on the comparison of certain parameters used to predict whether a person is having diabetic problem or not [2, 3]. The artificial intelligent-based methods are very effective and popular one in recent years [24]. The diagnosis of diabetes and cancer prediction the adaptive neuro-fuzzy inference system shows better accuracy [12]. Also shows the accuracy information of Naive Bayes classifier and K-means algorithm. 80% accuracy is obtained from this method [12]. Modified extreme learning machine and back-propagation neural network method are addressed in prediction of diabetes mellitus [13]. The data mining techniques such as K-means algorithms, MAFIA algorithm, decision tree algorithm and other classification algorithms provide reliable performance in diagnosing the heart disease [7, 16, 25, 26]. It helps a non-specialized doctor to make the right decision about the heart disease risk level by generating original rules, pruned rules, classified rules and sorted rules [27].

3 Proposed method

The proposed method has been used with Anaconda tool (AEN 4.1 Version) for data analysis. Anaconda a package management system manages the package versions for predictive analysis and data management [28]. It has been taken three disease patient data such as diabetes, coronary heart disease and breast cancer data as input, the reason for choosing these datasets only because of life threading characteristics and find the efficiency of the proposed method in fruitful manner and some relativity is there between these datasets. These data are loaded and checked to see whether it has any missing values or not. If any missing values are found, they are replaced to a null value. Then it has been checked whether any columns in the data have any correlation with another column in the data individually. If any correlation is found between two columns, one of those column is removed. If any true and false value is found in data, it is replaced to 1 and 0, respectively. It has split the original data into training data which has 70% of original data and test data which has 30% of the original data.

To check the number of true and false cases in original, training and test data of the three-class data, it has been trained our three different class data with the Naive Bayes algorithm and calculate the results accuracy this algorithm gives to the three classes separately using confusion matrix [29]. The block diagram of the proposed method is shown in Fig. 1. The performance report shows the performance metrics of the accuracy calculation for each class data individually. Similarly, it has trained our three different class data with the random forest algorithm. It has calculated the results accuracy, and this algorithm gives the three classes separately using confusion matrix. The performance report shows the performance metrics of the accuracy calculation for each class data individually. Internal model parameters updated through epoch, every epoch contains one or more batches. The epoch can be applied until minimize the errors in datasets [30, 31].

It has taken few sample test data separately for each class data. Applying these sample data on each trained model of that disease shows us the results whether the data are identified with that disease or not. While comparing the results of both model for each class data, it can see that the model trained with Random forest gives the accurate results of classification. To find the efficiency of the proposed method, the trained data are compared separately with the proposed algorithms and also checked the performance of test data. The proposed method is also applicable for testing the real-time disease data for classification and to identify whether the patient is affected by the particular disease or not.

3.1 Proposed algorithm

3.1.1 Naive Bayes classification algorithm

Here it has used Bayes theorem for classification purpose and to assume that classification is predictor independent. It assumes that Naive Bayes classifier in the presence of a particular feature in a class is unrelated to any other feature.

Naive Bayes model is compatible for very large datasets to build and for further analysis. This model is very simple and sophisticated classification method, and it performed well even in complicated scenarios. By using Bayes theorem, calculate the posterior probability using the equation below:

$$P(a/y) = (P(y/a)P(a))/P(y) \qquad (2)$$

where $P(a/y)$ indicates the posterior probability of class, $P(a)$ represents the class prior probability, $P(y/a)$ shows the likelihood which is the probability of predictor given class and $P(y)$ indicates the predictor's prior probability.

3.1.2 Random forest algorithm

The random forest (RF) is a hierarchical collection of tree structured base classifiers. Text data usually have many number of dimensions. The dataset contains a large number of irrelevant attributes. Only few important attributes are informative for classifier model. RF algorithm uses a simple predetermined probability to select the most important relevant attribute. Breiman formulated the RF algorithm using sample data subsets and to construct multiple decision trees by mapping random sample of feature subspaces. The RF algorithm associated with a set of training documents D and N_f features can be described as follows:

(1) Initial: $D1, D2,.......D_K$ sampled by predetermined probability with replacement.
(2) For each document D_K construct a decision tree model. The training documents are randomly sampled using its subspace of m-try dimension from the available features. Calculate all possible probability based on the m-try features. The leaf node produces the best data split. The process will be continued till it reaches the saturation criterion.

Combine the K number of unpruned trees $h_1(X_1), h_2(X_2),........$ into a random forest ensemble and use the high probability value for classification decision.

Random forest pseudocode:

1. Randomly select **"n"** features from total **"k"** features.

    1. Where **n < < k**

2. Among the **"n"** features, calculate the node **"n"** using the best split point.
3. Categorize the node into **daughter nodes** using the **best split**.

4. Repeat **1 to 3** steps until "l" number of nodes has been reached.
5. Build forest by repeating steps **1 to 4** for "n" number times to create **"n" number of trees**.

4 Results and discussion

4.1 Dataset

Here, it has multiple disease data such as diabetes, coronary heart disease and breast cancer. The dataset has been collected using the wearable devices and the prediction data.

4.1.1 Diabetes dataset

This dataset is taken originally from the NIDDK. All patients' data given here are females at least 21 years old of Pima Indian heritage.

- Pregnancies: baby deliveries happened in number of times
- Glucose: the concentration test in glucose using the tolerance test for every 2
- BP: diastolic BP (mm Hg)
- Skin thickness: the thickness of skin in triceps fold (mm)
- Insulin: insulin serum for 2-h (mu U/ml)
- BMI: height/weight
- Diabetes: prediction in mm
- Age: in years
- Outcome: true class variable either (0 or 1)

4.1.2 Coronary heart disease

This dataset is used by Framingham heart study which includes several demographic risk factors:

- Age: current age of the patient;

  The dataset also includes behavioral risk factors associated with smoking:

- Smoking nature: the patient is a current smoker or not

  Medical history risk factors:

- BPMeds: whether the patient was on blood pressure medication or not.
- prevalentStroke: whether the patient had a stroke previously or not.
- prevalentHyp: hypertensive or not.
- Diabetes: patient has diabetes or not.

Risk factors from the first physical examination of the patient.

- Cholrange: total cholesterol level.
- BPs: systolic blood pressure.
- diBl: diastolic blood.
- BMI: body mass index.
- HR: heart rate.
- GL: glucose level.
- CHDRISK: CHD coronary heart disease.

4.1.3 Breast cancer

Here, the dataset taken is breast cancer Wisconsin dataset. The attributes of this dataset are.

- Regn ID
- Diagnosis (1 = true, 0 = false) and various data

4.1.4 Filling in missing values

Missing value in any data means that the data were not available or not applicable or the event did not happen. Here, it has been replaced the missing values not available into null values.

4.1.5 Correlation coefficient

Correlation coefficients are used in statistics to measure how strong a relationship is between two variables. It is the statistical measure of the linear relationship between a dependent variable and an independent variable. It is represented by a lowercase letter '*r*'.

Here, the correlation between all the columns of the datasets is calculated to measure their relationship (Table 1). The results give the correlation value of each column in a dataset against another column in that dataset. If two columns in a dataset have same correlated values, then one among them is removed to avoid repetition of values.

$$r = \frac{m\left(\sum ab\right) - \left(\sum a\right)\left(\sum b\right)}{\sqrt{\left(\left[m\sum a2 - \left(\sum a\right)2\right]\left[n\sum b2 - \left(\sum b\right)2\right]\right)}} \tag{3}$$

'*m*' represents the quantity of information, $\sum a$ indicates the total of first variable value, $\sum b$ represents the total of second variable value, $\sum ab$ indicates the sum of the product of first and second variable values, $(\sum a)$^2 indicates the sum of the square of first value, and $(\sum b)$^2 represents the sum of the square of second value.

Table 1 Diabetes data

Sl. no	Preg_num	Gluc_Conc	BP	Skin thickness	Serum insulin	BMI	diab_pred	Age	Skin	Outcome of diabetes
1	6	148	72	35	0	33.6	0.627	50	1.3790	True
2	1	85	66	29	0	26.6	0.351	31	1.1426	False
3	8	183	64	0	0	23.3	0.672	32	0.0000	True
4	1	89	66	23	94	28.1	0.167	21	0.9062	False
5	0	137	40	35	168	43.1	2.288	33	1.3790	True
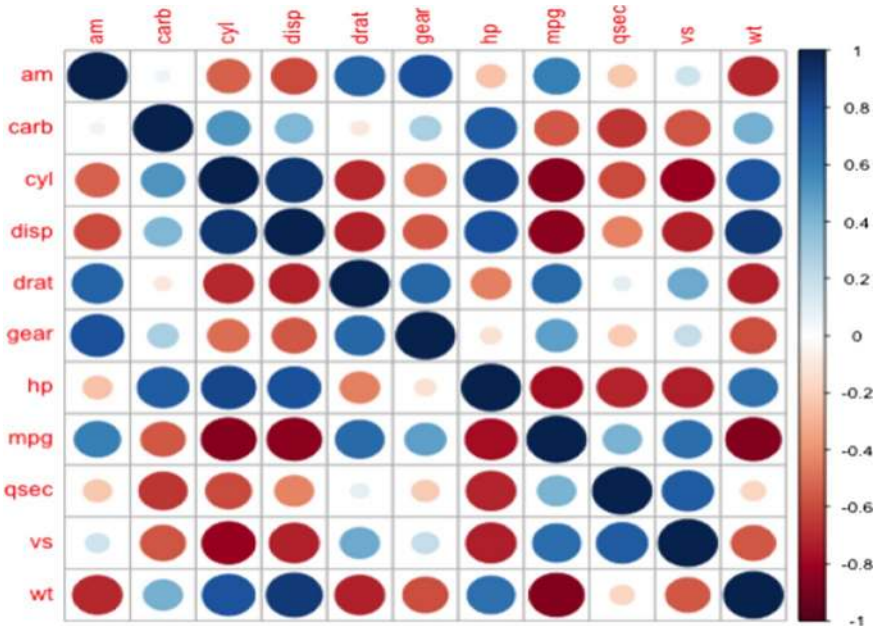
Fig. 2 The legend colors show the correlation coefficients and its corresponding colors (color figure online)

In Fig. 2, the positive coefficients are indicated with the blue color and negative coefficients in red. The color intensity is found proportional to the blue and red indicted.

All the pairs of variables and correlations represented by the correlogram. Blue color represents the positive correlations, and red color represents the negative correlations. The colour intensity and correlation coefficient are proportional (Table 2).

Finally a negative correlation of two variables implies that under consideration and changes in opposite directions, i.e., if any one of the variable increases other one is decreases and vice versa.

The correlation coefficient value is ranged from −1 to 1. If the rage is not fallen between this value means some error is there in the system. The correlation value −1 represents a exact negative correlation. The correlation value 1 shows that perfect positive correlation. wt, vs, qsec, mpg, hp, gear, drat, disp, cyl, carb, and am are representing the different parameters.

4.2 Confusion matrix

The confusion matrix is helpful to predict the classification problems. In the predicted class, the total number of exact predictions for a class goes into the expected row for

Table 2 Heart disease data

Sl. no	Age	Current smoker	BPMeds	Prevalent stroke	Prevalent Hyp	Diabetes	Chol range	BPs	diBl	BMI	HR	GL	CHD RISK
0	39	0	0.0	0	0	0	195.0	106.0	70.0	26.97	80.0	77.0	0
1	46	0	0.0	0	0	0	250.0	121.0	81.0	28.73	95.0	76.0	0
2	48	1	0.0	0	0	0	245.0	127.5	80.0	25.34	75.0	70.0	0
3	61	1	0.0	0	1	0	225.0	150.0	95.0	28.58	65.0	103.0	1
4	46	1	0.0	0	0	0	285.0	130.0	84.0	23.10	85.0	85.0	0

Fig. 3 Confusion matrix

that class value. In the same manner, the total number of incorrect class predictions go into the expected row for that class value and the class value of predicted column.

A confusion matrix having the information or data about actual and predicted classifications finished by a classification process as shown in Fig. 3. The performance is evaluated using the available data in the matrix (Table 3). The confusion matrix for a two-class classifier is shown in the table.

- Always positive (AP)—the classification model correctly finds class positively.
- Always negative (AN)—the negative class exactly labeled by the classifier.
- Always least positive (ALP)—the classification model was incorrectly predicted and labeled as positive.
- Partial least negative (PLN)—these are the positive classes that were incorrectly predicted as negative one.

4.2.1 Accuracy calculation

The prediction accuracy is calculated using the formulae

$$\text{Accuracy} = (AP + AN)/(M + N) \tag{4}$$

where, $M = AP + AN$ and $N = ALP + AN$. Or $AP + AN$ (TOTAL).

4.2.1.1 Precision (positive predictive value) Precision (PREC) is a classification technique which is used to find the items that are incorrectly labeled among the given class. The best precision result is 1.0, whereas the worst one is 0.0.

$$\text{PREC} = TP/(TP + FP) \tag{5}$$

4.2.1.2 Recall Sensitivity (SN) is calculated using the number of correct positive prediction value divided by the total number of positive predictions. It is also called as recall (REC) or true positive rate (TPR). The best value is 1.0 and the worst value is 0.0.

$$\text{SN} = TP/(TP + FN)/(TP/P) \tag{6}$$

4.2.1.3 F1-score F1-score is a weighted average of recall and calculated precision value.

Table 3 Cancer data

Sl. no	ID	Diag-nosis	Radius_mean	Tex-ture_mean	Cover-age_mean	Space_mean	Smooth-ness_mean	Com-pact_mean	Con-cave_mean	Con-cave_points_mean	Radius_worst	Fea-ture_worst	Space_worst	Area_worst	Smooth-ness_worst	Com-pact-ness_worst	Con-cav-ity_worst	Con-cave_points_worst	Sym-metry_worst	Fractal_dimen-sion_worst
1	842,302	1	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	25.38	17.33	184.60	2019.0	0.1622	0.6656	0.7119	0.2654	0.4601	0.11890
2	842,517	1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	24.99	23.41	158.80	1956.0	0.1238	0.1866	0.2416	0.1860	0.2750	0.08902
3	84,300,903	1	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	23.57	25.53	152.50	1709.0	0.1444	0.4245	0.4504	0.2430	0.3613	0.08758
4	84,348,301	1	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	14.91	26.50	98.87	567.7	0.2098	0.8663	0.6869	0.2575	0.6638	0.17300
5	84,358,402	1	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	22.54	16.67	152.20	1575.0	0.1374	0.2050	0.4000	0.1625	0.2364	0.07678

$$F1 = 2TP/(2TP + FP + FN) \tag{7}$$

4.2.1.4 Support Support is the number of occurrence of true and false values in actual class.

For diabetes data, Naive Bayes algorithm gives 76.72 and 74.46 accuracies for training and test data, respectively. Random forest algorithm gives 98.88 and 74.03 for training and test data, respectively.

For heart disease data, Naive Bayes algorithm gives 82.44 and 82.35 accuracies for training and test data, respectively. Random Forest algorithm gives 97.96 and 83.85 for training and test data, respectively.

For cancer data, Naive Bayes algorithm gives 62.06 and 63.74 accuracies for training and test data, respectively. Random forest algorithm gives 99.50 and 92.40 for training and test data, respectively, as shown in Figs. 4 and 5.

When applying sample test data of each disease to its trained model, the model trained with random forest algorithm gives accurate results when compared to the models trained with Naïve Bayes algorithm. The main reason for the performance of random forest algorithm against test data is the property of high-dimensional
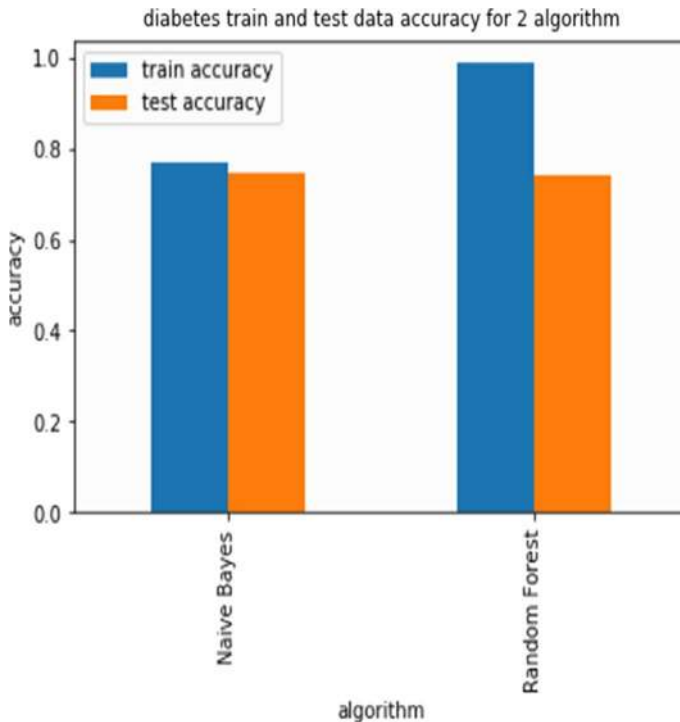
Fig. 4 Diabetes train and test data accuracy for two algorithms
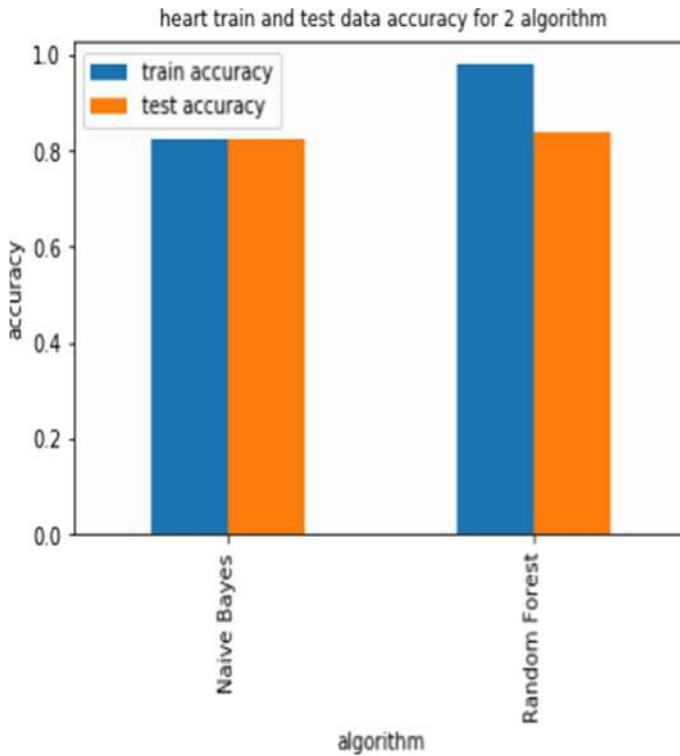
Fig. 5 Heart disease train and test data accuracy for two algorithms

feature and self-judge the essential features in the dataset. Feature interaction is also recognized by the random forest algorithm.

The results of the diagnosis are given in 1′s and 0′s where 1 indicates the sample test case has been diagnosed to have that disease. The model trained with random forest algorithm gives accurate results. The deviation is there in the results of training data and testing data results, it's only because of the number of data tested is varied in both cases [29].

5 Performance analysis

The calculated precision accuracy of testing data and trained data is compared with K-means clustering algorithm and DBSCAN (density-based spatial clustering of applications with noise) for finding the effectiveness of the proposed algorithms against diabetes dataset, coronary heart diseases and cancer dataset. The k-means clustering algorithm is one of the machine learning algorithm used to estimate the centroid of the cluster based on the calculated mean values. Some authors used the same algorithm for analysis, and we have compared the results

cancer train and test data accuracy for 2 algorithm
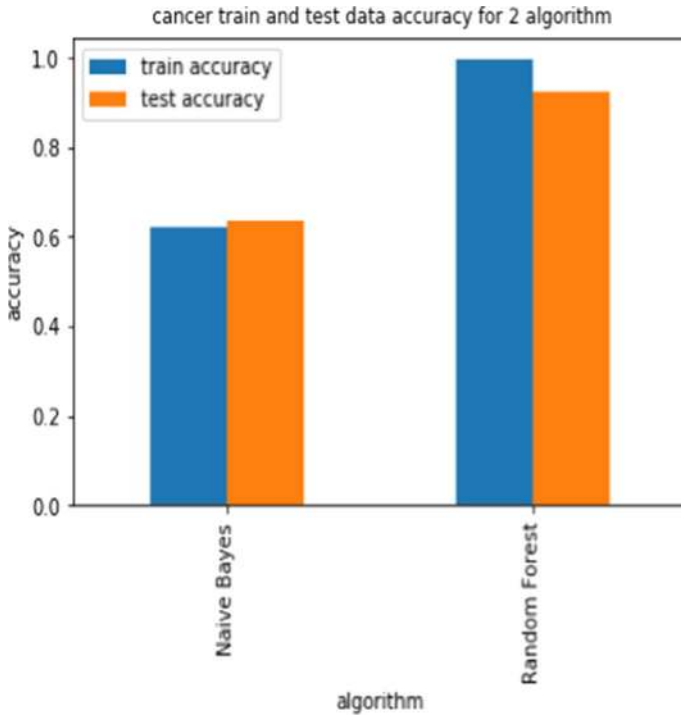
Fig. 6 Cancer train and test data accuracy for two algorithms
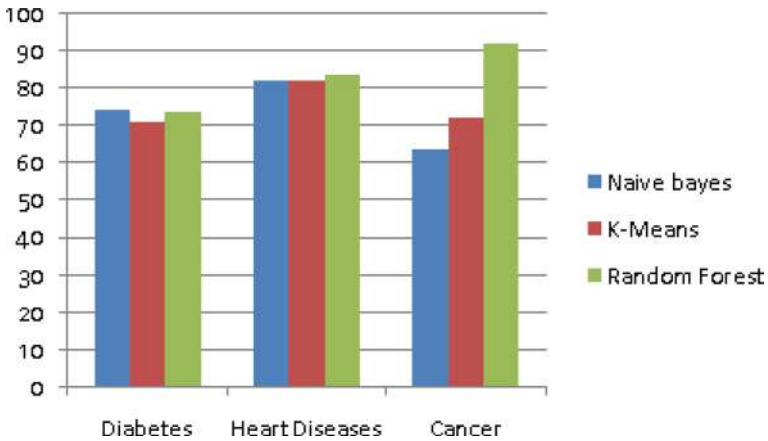
Fig. 7 Test data compared with K-Means clustering algorithm

of the k-means clustering algorithms with our proposal and the results shown in the figures (Fig. 6). The DBSCAN algorithm working is based on adding of noise content with the original data for effective detection of required data. The
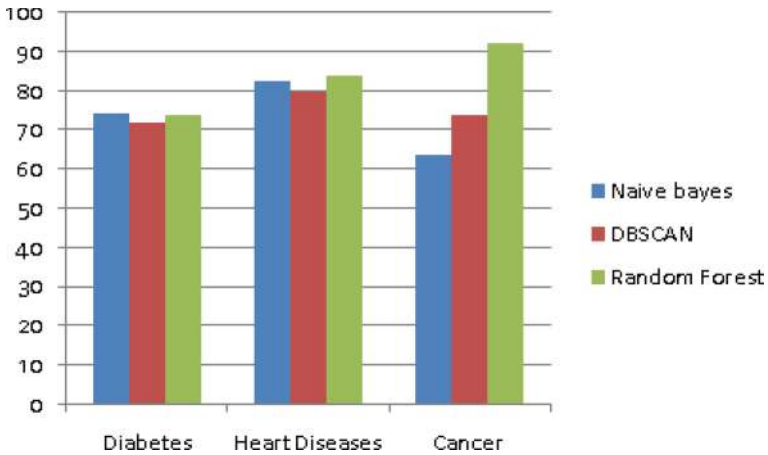
Fig. 8 Test data compared with DBSCAN Algorithm

DBSCAN algorithm results taken from the previous proposal are considered for our comparison. The same benchmark dataset is taken for the analysis.

Figure 7 shows the test data accuracy calculation results of Naïve Bayes algorithm, K-means clustering algorithm and random forest algorithm. The test data are taken for the analysis, and the performance of the k-means algorithm is somewhat better against the heart diseases dataset and cancer datasets. But the performance of proposed random forest algorithm is far better than k-means clustering.

Figure 8 shows the accuracy calculation results of Naïve Bayes algorithm, DBSCAN clustering algorithm and random forest algorithm. The test data are taken for the analysis, the results of DBSCAN algorithm are somewhat better against the
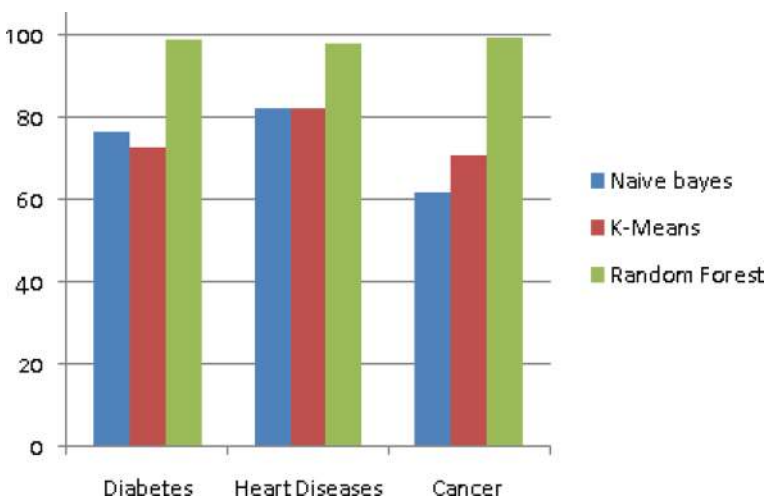
Fig. 9 Training data compared with K-Means clustering algorithm

cancer datasets. But the performance of proposed random forest algorithm is performed will than DBSCAN clustering.

Figure 9 shows the accuracy calculation results of Naïve Bayes algorithm, K-means algorithm and random forest algorithm. The training data are taken for the analysis, and the results of K-means algorithm are somewhat better against heart diseases and cancer datasets compared with Naïve Bayes algorithm. But the performance of proposed random forest algorithm shows good results than K-means clustering.

Figure 10 shows the accuracy calculation results of Naïve Bayes algorithm, K-means algorithm and random forest algorithm. The training data are taken for the analysis, the results of DBSCAN algorithm are somewhat better against heart diseases, and cancer datasets are compared with Naïve Bayes algorithm. But the performance of proposed random forest algorithm shows good results than DBSCAN clustering.

6 Conclusion

Data mining can be effectively implemented in medical domain. The aim of this study is to discover a model for the diagnosis of diabetes, coronary heart disease and cancer among the available dataset. The dataset is chosen from online repositories. The techniques of pre-processing applied are filled in missing values and removing correlated columns. Next, the classifier is applied to the preprocessed dataset, and then Bayesian and random forest models are constructed. Finally, the accuracy of the models is calculated and analyses are based on the efficiency calculations. Bayesian Classification network shows the accuracy of 74.46, 82.35 and 63.74% for diabetes, coronary heart disease and cancer data. Similarly, classification with Random forest model shows the accuracy of 74.03, 83.85 and 92.40. The accuracy outcome of Random forest model for the three diseases is greater than the accuracy values of Naïve Bayes classifier. Finally, the proposed
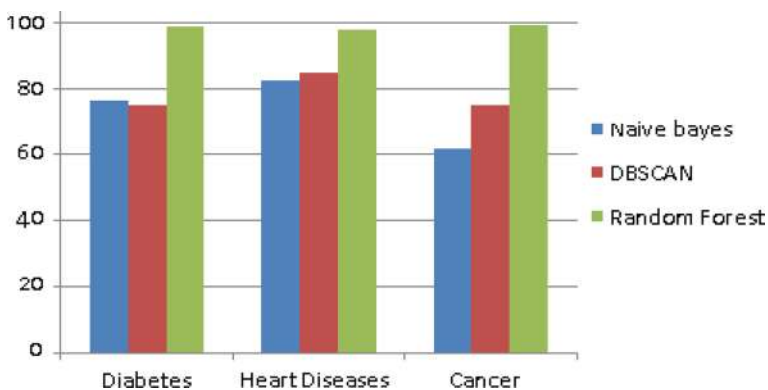

Fig. 10 Training data compared with DBSCAN Algorithm

algorithms are compared against K-means clustering algorithm and DBSCAN algorithm for identifying the effectiveness, and the result graph shows that the random forest algorithm works well compared with other two algorithms. When performing classification in the trained model by applying sample test data of each disease, the random forest model gives accurate results. The proposed model works well against train data and test data further this model will provide the better results for real-time data.

Our proposed methodology helps to improve the accuracy of diagnosis and greatly helpful for further treatment. In future enhancements, the accuracy has to be tested with different dataset and to apply other AI algorithms to check the accuracy estimation. The limitation of the proposed model is processing time, because of huge amount of data taken for estimating the performance of train data. In future, the same algorithms to be implemented with real-time data for estimating the effectiveness of the system.

Compliance with ethical standards

Conflict of interest The authors declare that there is no conflict of interest.

References

1. Renjit JA, Shunmuganathan KL (2010) Distributed and coorperative multi-agent based intrusion detection system. Indian J Sci Technol 3(10):1070–1074
2. Priyadarshini R, Dash N, Mishra R (2014) A novel approach to predict diabetes mellitus using modified extreme learning machine. In: International Conference on Electronics and Communication Systems (ICECS), 2014, pp 1–5
3. . Sankaranarayanan S, Perumal TP (2014) Diabetic prognosis through data mining methods and techniques. In: International Conference on Intelligent Computing Applications, 2014, pp 162–166
4. Dahiwade D, Patle G, Meshram E (2019) Designing disease prediction model using machine learning approach. In: Third IEEE International Conference on Computing Methodologies and Communication (ICCMC), 2019
5. Geetha R, Sivasubramanian S, Kaliappan M et al (2019) Cervical cancer identification with synthetic minority oversampling technique and PCA analysis using random forest classifier. J Med Syst 43:286. https://doi.org/10.1007/s10916-019-1402-6
6. Annamalai S, Udendhran R, Vimal S (2019) An intelligent grid network based on cloud computing infrastructures. Nov Pract Trends Grid Cloud Comput. https://doi.org/10.4018/978-1-5225-9023-1.ch005
7. Wu H, Yang S, Huang Z, He J, Wang X (2018) Type 2 diabetes mellitus prediction model based on data mining. Inform Med Unlocked 10:100–107

8. Sarwar A, Sharma V (2012) Intelligent Naïve Bayes approach to diagnose diabetes type-2. In: Special Issue of International Journal of Computer Applications on Issues and Challenges in Networking, Intelligence and Computing Technologies, November 2012

9. Pradeepa S, Manjula KR, Vimal S et al (2020) DRFS: detecting risk factor of stroke disease from social media using machine learning techniques. Neural Process Lett. https://doi.org/10.1007/s11063-020-10279-8

10. Kalaiselvi C, Nasira GM (2014) A new approach of diagnosis of diabetes and prediction of cancer using ANFIS. In: IEEE Computing and Communicating Technologies, 2014, pp 188–190

11. Robinson YH, Vimal S, Khari M, Hernández FCL, Crespo RG (2020) Tree-based convolutional neural networks for object classification in segmented satellite images. Int J High Perform Comput Appl. https://doi.org/10.1177/1094342020945026

12. Undre P, Kaur H, Patil P (2015) Improvement in prediction rate and accuracy of diabetic diagnosis system using fuzzy logic hybrid combination. In: International Conference on Pervasive Computing (ICPC), 2015, pp 1–4

13. Yi Y, Wu J, Xu W (2011) Incremental SVM based on reserved set for network intrusion detection. Elsevier J Expert Syst Appl 38(6):7698–7707

14. Ramamurthy M, Krishnamurthi I, Vimal S, Harold Y (2020) Robinson deep learning based genome analysis and NGS-RNA LL identification with a novel hybrid model. 197: 104211. https://doi.org/https://doi.org/10.1016/j.biosystems.2020.104211

15. Pradeepa S, Gayathri P, Nishmitha P, Vimal S, Oh-Young S, Usman T, Raheel N (2020) IoT based health-related topic recognition from emerging online health community: med help using machine learning technique. Electronics 9(9):1469

16. Babu S, Vivek EM, Famina KP, Fida K, AswathiP, Shanid M, Hena M (2017) Heart disease diagnosis using data mining technique. In: International Conference on Electronics, Communication, and Aerospace Technology, ICECA2017

17. Sampaul TGA, Robinson YH, Julie EG, Shanmuganathan V, Nam Y, Rho S (2020) Diabetic retinopathy diagnostics from retinal images based on deep convolutional networks. Preprints. https://doi.org/10.20944/preprints202005.0493.v1

18. Vimal S et al (2020) Deep learning-based decision-making with WoT for smart city development. In: Jain A, Crespo R, Khari M (eds) Smart innovation of web of things, CRC Press, Boca Raton, pp 51–62. https://doi.org/10.1201/9780429298462

19. Kumari M, Vohra R, Arora A (2014) Prediction of diabetes using Bayesian network. Int J Comput Sci Inf Technol (IJCSIT) 5(4):5174–5178

20. Krishnaiah V, Narsimha G, Chandra NS (2013) Diagnosis of lung cancer prediction system using data mining classification techniques. Int J Comput Sci Inf Technol 4(1):39–45

21. Long NC, Meesad P, Unger H (2015) A highly accurate firefly-based algorithm for heart disease prediction. Expert Syst Appl 42:8221–8231

22. Esteghamati A, Hafezi-Nejad N, Zandieh A, Sheikhbahaei S, Ebadi M, Nakhjavani M (2014) Homocysteine and metabolic syndrome: from clustering to additional utility in prediction of coronary heart disease. J Cardiol 64:290–296

23. Lee BJ, Kim JY (2016) Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning. IEEE J Biomed Health Inform 20(1):39–46

24. Wang Z, Srinivasan RS (2017) A review of artificial intelligence based building energy use prediction: contrasting the capabilities of single and ensemble prediction models. Elsevier J Renew Sustain Energy Rev 75:796–808

25. Lynch CM, Abdollahi B, Fuqua JD, de Carlo AR, Bartholomai JA, Balgemann RN, van Berkel VH, Frieboes HB (2017) Prediction of lung cancer patient survival via supervised machine learning classification techniques. Int J Med Inform 108:1–8

26. Veena Vijayan V, Anjali C (2015) Prediction and diagnosis of diabetes mellitus: a machine learning approach. In: 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS), December 2015

27. Ren F, Hu L, Liang H, Liu X, Ren W (2008) Using density-based incremental clustering for anomaly detection. In: International Conference on Computer and Software Engineering, IEEE, pp 986–989

28. Vimal S et al (2016) Secure data packet transmission in MANET using enhanced identity-based cryptography. Int J New Technol Sci Eng 3(12):35–42

29. Suresh A, Udendhran R, Vimal S (2020) Deep neural networks for multimodal imaging and biomedical applications. IGI Global, Hershey,. https://doi.org/10.4018/978-1-7998-3591-2

30. Nai-arna N, Moungmaia R (2015) Comparison of classifiers for the risk of diabetes prediction. In: 7th International Conference on Advances in Information Technology Procedia Computer Science, vol 69, pp 132 –142

31. Zhang Z, Shen H (2005) Application of online-training SVMs for real time intrusion detection with different considerations. Comput Commun 28(12):1428–1442