International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

A Hybrid System to Improve the Performance of Diabetes Disease Prediction using Genetic Algorithm

Emrana Kabir Hashi, Md. Shahid Uz Zaman

Abstract*: Currently, data mining is playing a significant role in the healthcare system. It helps to extract the hidden pattern from the clinical dataset for further analysis. Also, it can be used to build a tool to manage the medical management system. Among the life-threatening diseases, diabetes mellitus is treated as a serious disease worldwide. Due to its mortality rate, early prediction and diagnosis are very important. Several research works are going on the mentioned issues to reduce the complications caused by diabetes as well as the mortality rate. The medical science needs to analyze an enormous quantity of clinical data for diagnosis purposes using machine learning techniques. In recent approaches, the disease datasets may contain insignificant and digressive features causing less accurate results. The aim of this paper is to analyze the existing prediction systems and hence develop a hybrid disease prediction model using the Genetic Algorithm for Naïve Bayes, Decision Tree and Support Vector Machine classifiers for better accuracy. This proposed diabetes prediction model produces the accuracies of 0.8182, 0.8052, and 0.8312 when Naïve Bayes, Decision Tree, and Support Vector Machine classifiers are used respectively. From the experimental results, it can be demonstrated that for all cases Support Vector Machine provides higher accuracy comparing to the other classifiers. In the analysis, the Pima Indian diabetes dataset is used to construct the proposed model.

***Keywords* : Machine Learning, Feature Selection, Genetic Algorithm, Decision Tree, Naïve Bayes, Support Vector Machine.**

I. INTRODUCTION

Every year the healthcare industry produces a massive amount of clinical data which store the diagnosis results of patients. So the challenge is to provide quality services and effective treatments by developing a computer-based decision support system. With the blessings of modern technology, the applications of machine learning and data science play an important role in healthcare to predict disease which might offer higher choices to physicians and value-effective treatment to patients [1]–[10].

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author
Emrana Kabir Hashi*, Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh. Email: emranakabir@gmail.com
Md. Shahid Uz Zaman, Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh. Email: szaman22@gmail.com

Diabetes has become a serious disease and it is important to diagnosis it at an early stage due to its mortality rate. Diabetes is a disease where the blood glucose levels get increase which is due to the defects in secretion of insulin, or its action, or both. In diabetes, the cells of a person produce an insufficient amount of insulin or unable to use insulin properly and efficiently that further leads to hyperglycemia and type-2 diabetes. A lot of rehearses have been done on this diabetes disease to predict this disease in early-stage to overcome its complications such as strokes, blindness, kidney failure, damage and failure of several organs. The Pima Indians Diabetes Database of the National Institute of Diabetes and Digestive and Kidney Diseases has used as diabetes dataset [2], [10] and [11].

There are several classification algorithms such as Decision Tree, K nearest neighbors (KNN), Naïve Bayes, and Support Vector Machine (SVM), etc. help to develop a prediction model. But less significant features are responsible to decrease the performance of these classifiers. Better analysis can be obtained using Feature Selection which is the process of identifying and removing the irrelevant and redundant data from a dataset in order to improve the performance of the machine learning algorithms [9] and [10]. So it is important to develop an efficient feature selection model which will increase the performance of classification techniques. This work presents a hybrid approach to implement a clinical decision support system utilizing feature selection and classification techniques.

The research objective is to develop a hybrid prediction model to predict diabetes diseases using Genetic Algorithm (GA) based feature selection for the Decision tree, Naïve Bayes and SVM classifiers.

II. RELATED WORKS

This section summarizes a number of researches that employed various data mining and machine learning algorithms in the design of intelligent healthcare applications.

Huang and Wang [1] implemented a prediction model using a GA-based feature selection method for SVM classifiers. This paper obtained an 81.5% average overall hit ratio for the Pima Indians Diabetes Database. GA-SVM model was used for predicting several real-world datasets and the results showed that this model provided significant improvement in the performance of classification in comparison with Grid search.

The paper [2] in 2018 by Patil and Tamane worked for upgrading the performance of KNN (accuracy- 83.12%) and Naïve Bayes (accuracy- 81.12%) in diabetes prediction using a genetic algorithm for the feature section.

The paper [3] in 2018 by D. Choubey, S. Paul, S. Kumar, and S. Kumar showed 78.69% accuracy for Pima Indians Diabetes Database which is implemented by GA as an Attribute Selection and Naïve Bayes for Classification.

For predicting diabetes disease (Pima Indians Diabetes Database), paper [4] presented a comparison between Decision Tree (accuracy- 76.96%) and Naïve Bayes algorithm (accuracy- 79.56%) with the help of WEKA tool.

The paper [5] proposed a feature selection model using symmetrical uncertainty attribute set evaluator and fast correlation-based filter and showed that the LIBSVM classifier selected 4 features and obtained accuracy- 77.99% for Pima Indians Diabetes Database.

The research paper [6] developed a method using combined dataset of Diabetes disease. Here Fselect (accuracy- 63.54%, specificity- 43.00%, and sensitivity- 99.80%), wrapper (accuracy- 70.69%, specificity- 38.36% and Sensitivity- 89.95) and Ranker (accuracy- 72.61%, specificity- 41.04%, and sensitivity- 90.76%) methods are used for feature selection and LIBSVM for classification feature.

The paper [7] in 2018 by D Jain and V. Singh presented the merits and demerits of various feature selection approaches and classification algorithms for disease prediction. This study implied that there are more opportunities for developing new approaches to increase the success rate.

In the [8] and [9] research paper, the authors conferred a model wherever feature selection approaches area unit used for classification algorithms to realize the most effective attainable performances.

III. RESEARCH METHODOLOGIES

A. Wrapper Approach for Feature Selection

Feature selection is additionally called variable selection, attribute selection, variable subset selection. It is the process of eliminating irrelevant or redundant attributes in a given dataset. In the medical domain, immense amounts of knowledge square measure generated. Mining on the reduced set of features, it is more beneficial to extract and perceive the unknown pattern [8]–[10]. Additional in supervised learning, it is going to increase the performance of the classifiers with the selected feature subset. Three types of traditional feature selection approach for machine learning [2] i.e. Filter, Wrapper, and Embedded approaches are broadly used to extract the feature set and this paper, the wrapper approach is utilized to select the optimum feature subset.

The wrapper approach uses the classifier as an induction algorithm to measure the good subset. Wrapper methods typically achieve better accuracy rate and use cross validation to avoid over-fitting and sometimes they are too expensive for the large dimensional database in terms of computational complexity and time complexity, since each feature set considered must be evaluated with the classification algorithm. Fig. 1 illustrates the wrapper approach to select the feature subset, wherever the induction algorithm used as an evaluation function to search the optimal feature subset [8] and [9]. The accuracy of the induction algorithm is calculated using an accuracy estimation technique. Some common algorithms like Random Generation plus Sequential Selection, Simulated Annealing and Genetic Algorithm are commonly used for the wrapper approach.

GA is the most advanced algorithm for feature selection. It is a heuristic combinatorial optimization method. GA selects the foremost relevant feature to enhance the performance of the predictive model.

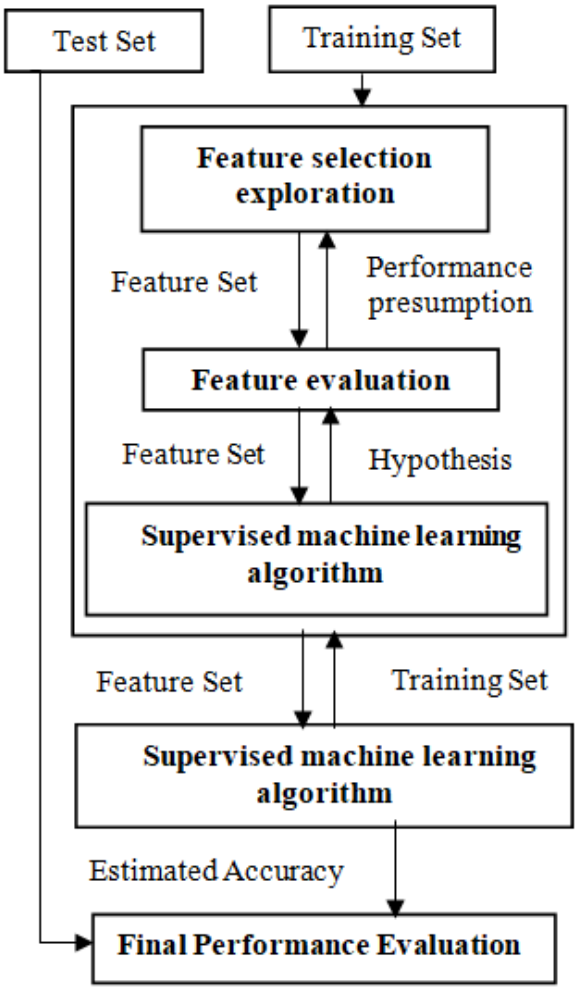

Fig. 1.The block diagram of wrapper approach.

B. Decision Tree Classifier

Classification is a data mining function that is the organization of data by assigning items in a given or target class. The decision tree classifier generates a tree structure model where the leaf node known as a decision node and an internal node represents a test node. The most useful attribute can be selected by calculating entropy and information gain formula. The entropy is the degree of components calculated with the assistance of the probability of the attribute item. The decision tree (C4.5) selects an attribute with the most information gain in each stage of iteration.

The highest information gain from all attributes employed to found the decision nodes. To classify unknown records, the entropy, info (p,T) and information gain are calculated by equation (1), (2) and (3) respectively [10]–[12].

$$\text{Gain(p)} = \text{F(Info (T)} - \text{ Info (p,T))} \quad (1)$$

Where,

$$\text{Info(T)} = \text{Entropie(p)} = -\sum_{i=1}^{n} p_i \times \log(p_i) \quad (2)$$

$$\text{Info(p,T)} = \sum_{j=1}^{n} \left(p_j \times \text{Entropie}(p_i)\right) \quad (3)$$

Here, F = number of known sample/total number of sample in the dataset for a given attribute, p_i = the set of probability distribution, T= Test, p_j= the set of all possible values for attribute T.

C. Naïve Bayes Classifier

Naïve Bayes classifier generates Bayes rules of conditional probability where all attributes are independent. It estimates the likelihood to create a model with predictive capabilities. This classifier can be viewed as both descriptive and predictive type where probabilities are descriptive and this is used to predict the target class [11] and [13].

Let X be a data sample and Ci be a hypothesis that X belongs to a class, the posterior probability [13] of a class is calculated by equation (4):

$$P(Ci|x) = \frac{\text{P(X|Ci)} * P(c)}{\text{P(x)}} \quad (4)$$

Where, P(ci |x) is the posterior probability of class(target) given predictor(attribute), P(c) is the prior probability of class, P(x|ci) is the likelihood which is the probability of predictor given class, P(x) is the prior probability of predictor.

As diabetes dataset contains continuous valued attributes so Gaussian distribution has applied in the proposed model. For example, x is the continuous valued attribute in the training set. The mean and variance of x in each class are computed after segmenting the data by class. The probability of distribution [13] of v has given a class c, P(x= v|c) can be computed by equation (5):

$$\text{P(x = v|c)} = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{\frac{(v-\mu_c)^2}{2\sigma_c^2}} \quad (5)$$

Where,

µc be the mean of the values in x associated with class c, σ_c^2 be the variance of the values in x associated with class c

D. Support Vector Machine Classifier

SVM is a supervised learning algorithm that categorizes data by finding a model of the optimal hyperplane. This can separate different dimensional data with a maximum interclass margin. It is one of the most popular machine learning and statistical algorithms which uses structural risk minimization principle [14]. It is used for both linear (linear kernel) and non-linear (RBF, sigmoid, and polynomial kernel) datasets. The selection of good kernel functions is also a research issue. There are some popular kernel functions used for general purposes [14] and [15].

• Linear kernel:

$$K(X_i, X) = X_i^T Xj$$

• Polynomial kernel:

$$K(X_i, X) = (\gamma\, X_i^T Xj + r)^d, \gamma > 0$$

• RBF kernel:

$$K(X_i, X) = exp(-\gamma \,||\, X_i - Xj \,||^2, \gamma > 0$$

• Sigmoid kernel:

$$K(X_i, X) = \tan h(\gamma\, X_i^T Xj + r)$$

The reason behind using a linear kernel is that the diabetes dataset has two classes, so it is linearly separable by the linear kernel function. It is less prone to overfitting than nonlinear kernel SVM. Sometimes it works faster than others in the training and testing phase.

IV. PROPOSED SYSTEM

This research proposes a hybrid disease prediction model using a wrapper feature selection method for Naïve Bayes, Decision tree and SVM classifiers. Here, the Genetic Algorithm is applied to extract the optimal feature subsets. Fig. 2 represents the overview of this proposed model. The main steps and explanations are described in the following sections.

A. Preprocessing

Data preprocessing is applied to the original dataset to identify and process the blatant, incomplete, irreverent and inconsistent value of attributes. In healthcare sectors generally, information assortment and storage are quite troublesome. Missing values can be replaced by a constant value, randomly estimated value from other predictive models, a mean, median or mode value for that corresponding column. Then scaling total dataset like 90% data are selected as a training set and the rest 10% are selected as a testing set.

B. GA based Wrapper Feature Selection

GA is a randomized search algorithm. It starts with generating a population or individual randomly. In the binary string, 1 represents the presence of attribute and 0 represents the absence of the attribute. The fitness function of this model is measured by the accuracy of classifiers. Here, 10 fold cross validation is applied to the training set to evaluate the fitness value. Then the crossover is applied and randomly chosen two individuals and combines their features to get the offspring. Then mutation is applied to randomly change the feature value of one individual. During this work, single bit crossover is chosen with crossover point=0.5, multi-bit flip is chosen with mutation point=0.2 and tournament selection size is 3 for 21 generations and 100 populations.

Finally, remove the repetitive individuals from new generations. Here, Naïve Bayes, C4.5 and SVM classifiers are used as induction algorithm to evaluate the fitness value of new individuals. The search space is 2^N where N is the number of the attribute. For diabetes disease dataset the number of search spaces is 2^8=256 and chromosome length is 8. Therefore stopping criteria is an important supported analysis on evaluation function. Some commonly used criteria are search completed or next iteration failed to produce a better subset.

In this system, 21 iterations are considered and checked that it provides higher accuracy during which number of iteration and note down the optimal feature subset.

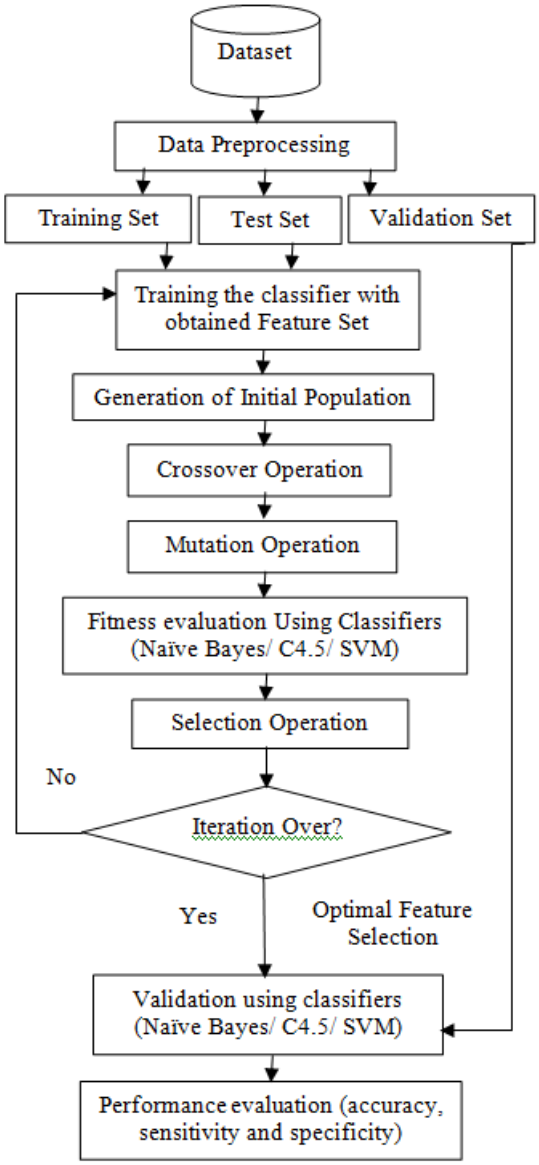

Fig. 2.The overview of proposed system.

C. Classification Models Construction

In this step, Naïve Bayes, C4.5 and SVM classifiers are applied to construct the classification models. The validation set is classified using this system with resultant optimal feature subset.

D. Performance Evaluation and Comparison

This step predicts the class of validation dataset and calculates the value of true positive, true negative, false positive and false negative. The performance of all classifiers are evaluated and compared.

V. EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

The attribute selection process is applied in the training set to select the optimal feature subset. Using GA, optimal features are selected based on the classifier models to check the stopping criteria. 10 fold cross validation is performed to evaluate the performance. Here, the training model of 0 to 20 iterations with average accuracy, standard deviation, minimum accuracy, and maximum accuracy are calculated using Naïve Bayes, decision tree and SVM classifiers.

A. Performance Evaluation Matrices

In this system, Accuracy, Sensitivity, Specificity, and ROC curve were used to evaluate the performance of the classifiers and also used 10 fold cross validation criteria to validate the classification results.

- In 10 fold cross validation criteria, first randomly divide the whole dataset into 10 equal sized subsets and trained nine subsets and testes the rest one, whole model is trained and tested for 10 times, finally average ten correctly classified accuracies [10].

  $$CV\ A = 1/10 \sum_{i=1}^{10} Ai \quad (6)$$

  Where CV A is the cross validated accuracy and Ai is the i^{th} correctly classified accuracy.
- A confusion matrix is a representation of classification results. In a general confusion matrix, TP, FN, FP, and TN represent True Positive, False Negative, False Positive and True Negative respectively.
- Specificity is calculated by dividing the true negative (TN) samples to the sum of true negative (TN) and false positive (FP) samples.

  $$\text{Specificity} = \frac{TN}{TN+FP} \quad (7)$$
- Sensitivity is calculated by dividing the true positive (TP) samples to the sum of true positive (TP) and false negative (FN) samples.

  $$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (8)$$
- Calculation of accuracy is performed by taking the ratio of truly classified samples (true negative, true positive) to the total number of samples.

  $$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$
- Receiver Operating Curve (ROC) is a fundamental tool for diagnostic test evaluation of the binary classifier system. In the ROC curve, the true positive rate (Sensitivity) is plotted in function of the false positive rate (1- Specificity) for various threshold settings.

B. Diabetes Dataset

In the proposed system, the feature selection and classification have been applied to the Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases. It contains 9 numerical valued attributes named pregnancy, plasma, pres, skin, insulin, mass, pedi, age, and class, the first 8 are independent and the last one is dependent. There is a total of 768 instances with 500 instances are tested negative and 268 instances are tested positive [4], [10] and [11].

C. Diabetes Prediction Model Using Naïve Bayes + GA

Table I represents the training performance of the Naïve Bayes + GA model and the graphical view of this model is shown in Fig. 3.

After applying Naïve Bayes classifier with the GA approach, 5 attributes were selected. In the 18th number of iteration, the training model reached the highest cross validated accuracy and that point 'plasma', 'skin', 'insulin', 'pedi' and 'age' attributes were selected as optimal feature subset. Then the test samples were classified by this model and obtained accuracy was 0.8182 and total runtime for both training and testing was required 6929.99ms.

Table- I: Performance of Naïve Bayes + GA model

Sl. no.	Average Accuracy	Standard deviation	Min	Max
0	0.707143	0.070407	0.442857	0.814286
1	0.760857	0.0386834	0.6	0.8
2	0.787	0.0225963	0.657143	0.857143
3	0.798143	0.021914	0.728571	0.857143
4	0.804	0.028	0.671429	0.857143
5	0.821571	0.0339565	0.685714	0.857143
6	0.838143	0.0329536	0.628571	0.857143
7	0.849714	0.0225407	0.742857	0.857143
8	0.846714	0.0363574	0.628571	0.857143
9	0.853	0.0184385	0.742857	0.857143
10	0.849286	0.0281668	0.671429	0.857143
11	0.848571	0.0321349	0.671429	0.857143
12	0.850714	0.0226216	0.742857	0.857143
13	0.854143	0.0215988	0.671429	0.857143
14	0.849714	0.0289602	0.671429	0.857143
15	0.85	0.025274	0.742857	0.857143
16	0.853	0.0206324	0.742857	0.857143
17	0.850286	0.0251915	0.742857	0.857143
18	**0.855857**	**0.00905651**	**0.785714**	**0.857143**
19	0.851857	0.0198623	0.742857	0.857143
20	0.852857	0.0176705	0.742857	0.857143

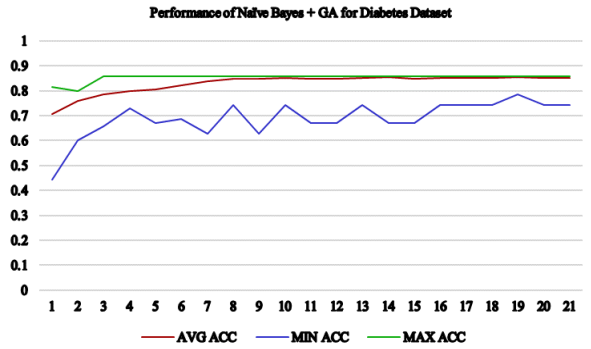

Fig. 3.The performance analysis of Naïve Bayes + GA.

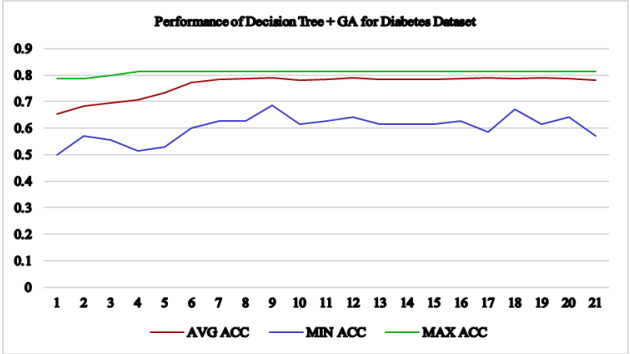

Fig. 4.The performance analysis of Decision Tree + GA.

D. Diabetes Prediction Model Using Decision Tree + GA

Table II and Fig. 4 are the representation of the diabetes prediction model using Decision Tree + GA. In the 16th number of iteration, the training model reached the highest cross validated accuracy and 5 attributes namely 'pregnancy', 'plasma', 'pres', 'insulin' and 'mass' were selected and the test accuracy was 0.8052 and total runtime was 7961.0ms.

Table- II: Performance of Decision Tree + GA model

Sl. no.	Average Accuracy	Standard deviation	Min	Max
0	0.653143	0.0522974	0.5	0.785714
1	0.683429	0.0449816	0.571429	0.785714
2	0.695143	0.0470441	0.557143	0.8
3	0.706	0.0593908	0.514286	0.814286
4	0.732286	0.0662592	0.528571	0.814286
5	0.771571	0.0426899	0.6	0.814286
6	0.783857	0.0308158	0.628571	0.814286
7	0.785857	0.0260686	0.628571	0.814286
8	0.789286	0.0237654	0.685714	0.814286
9	0.780429	0.0397264	0.614286	0.814286
10	0.783286	0.0296273	0.628571	0.814286
11	0.788857	0.0265407	0.642857	0.814286
12	0.784714	0.0301192	0.614286	0.814286
13	0.785	0.0326625	0.614286	0.814286
14	0.783857	0.0331758	0.614286	0.814286
15	0.788286	0.0284555	0.628571	0.814286
16	**0.790286**	**0.0286342**	**0.585714**	**0.814286**
17	0.787857	0.0278755	0.671429	0.814286
18	0.789857	0.0299861	0.614286	0.814286
19	0.786429	0.0331586	0.642857	0.814286
20	0.779714	0.0424351	0.571429	0.814286

E. Diabetes Prediction Model Using SVM + GA

The training performance of the model is presented in Table III and Fig. 5. In the 12th number of iteration, 4 attributes namely 'pres', 'mass', 'pedi' and 'age' were selected as the optimal feature subset. Finally, the obtained test accuracy was 0.8312 and runtime was 1090795.0ms.

Table- III: Performance of SVM + GA model

Sl. no.	Average Accuracy	Standard deviation	Min	Max
0	0.68	0.0678233	0.571429	0.8
1	0.735429	0.0498766	0.585714	0.8
2	0.766714	0.0237379	0.7	0.8
3	0.780571	0.0302047	0.628571	0.8
4	0.796143	0.0077867	0.757143	0.8
5	0.795429	0.0163981	0.671429	0.814286
6	0.797143	0.015253	0.671429	0.814286
7	0.798571	0.0173205	0.657143	0.814286
8	0.799	0.0285182	0.642857	0.814286
9	0.801143	0.0369047	0.571429	0.814286
10	0.808286	0.0262779	0.571429	0.814286
11	0.810571	0.0223716	0.6	0.814286
12	**0.812857**	**0.00914732**	**0.728571**	**0.814286**
13	0.809	0.0258453	0.585714	0.814286
14	0.806429	0.04157	0.571429	0.814286
15	0.808429	0.0281465	0.571429	0.814286
16	0.812	0.010246	0.742857	0.814286
17	0.811429	0.0127775	0.742857	0.814286
18	0.81	0.0261472	0.571429	0.814286
19	0.812286	0.0157817	0.657143	0.814286
20	0.810286	0.0149939	0.728571	0.814286

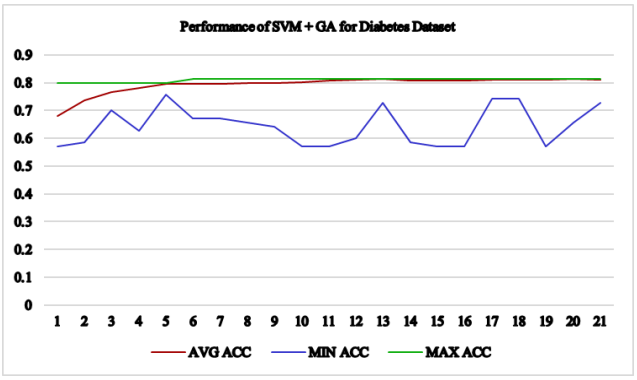

Fig. 5.The performance analysis of SVM + GA.

F. Performance Comparison of Proposed System

The performance of different classifiers for diabetes disease prediction has been compared. The summary results are presented in Table IV.

For Naïve Bayes+GA approach, its sensitivity is 0.90, specificity is 0.729, accuracy is 0.8182, and runtime is 6929.9998ms. For decision tree+GA approach, its sensitivity is 0.8333, specificity is 0.7586, accuracy is 0.8052, and runtime is 7961.0ms. For SVM+GA approach, its sensitivity is 0.8542, specificity is 0.7931, accuracy is 0.8312, and runtime is 1090795.0ms. Note that, both Naïve Bayes+GA and Decision tree+GA approach selected 5 attributes but SVM+GA approach selected small feature subset with 4 attributes. Here, the SVM + GA model has provided better accuracy performance compared with the other models.

Table- IV: Performance comparison of proposed system

Models	Sensitivity	Specificity	Accuracy	Time (ms)
Naïve Bayes+GA	0.90	0.729	0.8182	6929.99
Decision tree+GA	0.8333	0.7586	0.8052	7961.0
SVM+ GA	0.8542	0.7931	0.8312	1090795.0

In Fig. 6, Fig. 7, and Fig. 8, the ROC curves of different models are depicted below.

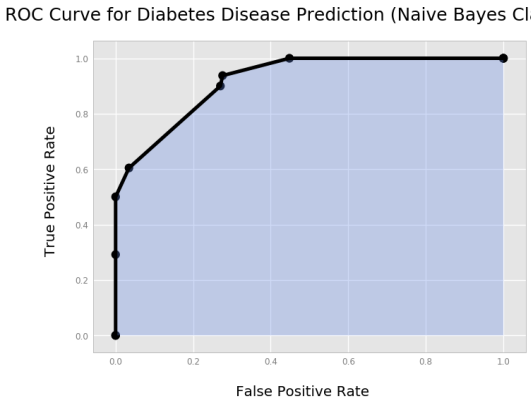

Fig. 6.The ROC for Naïve Bayes + GA model.

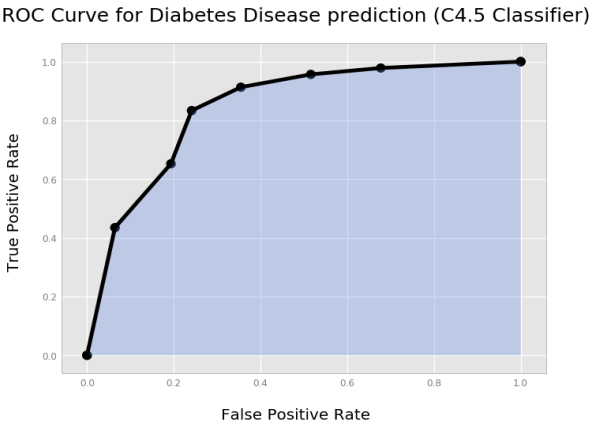

Fig. 7.The ROC for Decision Tree + GA model.

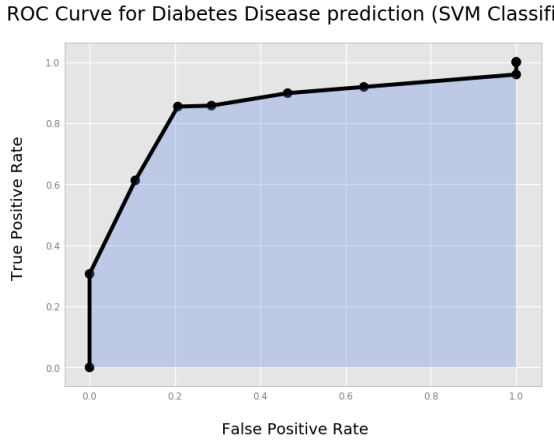

Fig. 8.The ROC for SVM + GA model.

The two populations with a true positive rate and false positive rate are discriminated against for every possible cut-off point. For the proposed three approaches, ROC spaces can demonstrate that the curve rising up closer to the left-hand border and then follow the right side of the top border. For different models, the ROC curve can be plotted for the different numbers of folds with the pair of sensitivity and specificity.

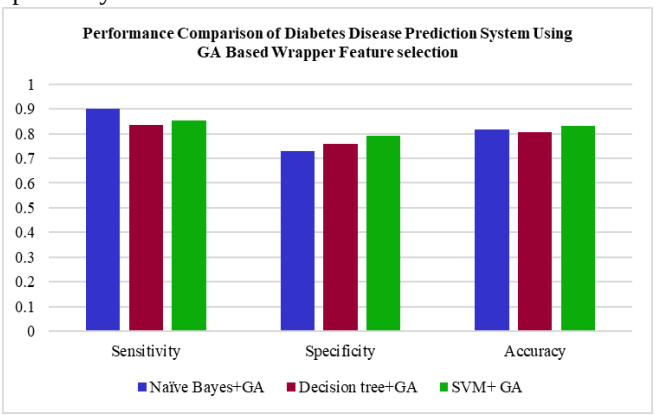

Fig. 9.The bar chart comparison of proposed system.

Finally, in Fig. 9 sensitivity, specificity, and accuracy are plotted in the bar chart comparison of the proposed system. The graphical view of sensitivity, specificity, and accuracy are plotted against three proposed models with three different colors.

G. Performance Comparison of the Baseline system and the Proposed System

The performance of the baseline prediction system and proposed prediction system have been compared. The comparison results are given in Table V.

Table- V: Performance comparison of the baseline system and the proposed system

Classifier	Accuracy (Baseline)	Accuracy (Proposed)	Time (ms) (Baseline)	Time (ms) (Proposed)
Naïve Bayes	0.7012	0.8182	16.00	6929.99
Decision tree	0.7142	0.8052	24.9967	7961.0
SVM	0.7402	0.8312	2595.0002	1090795.0

From this comparison, the baseline system accuracies are 0.7012, 0.7142, and 0.7402, and the proposed system accuracies are 0.8182, 0.8052, and 0.8312 for Naïve Bayes, Decision tree, and SVM respectively. It can be demonstrated that the performance of the proposed prediction model is better than the baseline model. But the runtime for the proposed system is inferior to that of the baseline prediction system. In this research, the implementation of both systems is performed under Python. Here, different classifiers are wrapped with GA and selected the optimal attribute subset. With fewer attributes, the proposed model provides better accuracy than the baseline model.

VI. CONCLUSION

The working process started with developing a disease prediction system using Naïve Bayes, Decision tree and SVM classifiers without feature selection approach. Finally, GA is used for feature selection with Naïve Bayes, decision tree and SVM classifiers. It can escape from the local minima problem and work well for both small and large numbers of feature sets. It is observed from the experimental results for all cases, the proposed system provides better performance compared with the baseline system. Also, SVM provides better accuracy than other classifiers as it is associated with an expert knowledge model. The different parameters and adaptive threshold values of classifiers ensure better performance. Although the SVM is slower compared to others the linear kernel SVM is used for training and testing purposes. Finally, considering the accuracy, SVM provides the best performance.

In the future, this system assumed to be experiment with different kernel functions like RBF, polynomial and sigmoid for SVM classifiers. Also, the embedded feature selection approaches may be used to select the optimal feature subset.

REFERENCES

1. C. Huang, and C. Wang, "A GA-based feature selection and parameters optimizationfor support vector machines." *Expert Systems with applications*, 31.2, 2006, pp. 231-240.
2. R. Patil Nitin, and S. Tamane, "Upgrading the performance of KNN and naïve bayes in diabetes detection with genetic algorithm for feature selection," *International Journal of Scientific Research in Computer Science* 3.1, 2018, pp. 1371-1381.
3. D. Choubey, S. Paul, S. Kumar, and S. Kumar, "Classification of Pima Indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection," *Proceedings of the International Conference on Communication and Computing System (ICCCS 2016)*, 2017, pp. 451-455.
4. A. Iyer, S. Jeyalatha and R. Sumbaly, "Diagnosis of diabetes using classification mining techniques," *Int. J. of Data M. & Know. Manag. Process, IJDKP, United Arab Emirates*, vol. 5, January 2015, pp. 1-14.
5. S. Balakrishnan, and R. Narayanaswamy, "Feature selection using FCBI in type II diabetes databases," *International Journal of the Computer, the Internet and the Management* 17, no. 1, 2009, pp. 50.1-5.8.
6. A.Negi, and V. Jaiswal, "A first attempt to develop a diabetes prediction method based on different global datasets," *In Parallel, Distributed and Grid Computing (PDGC), 2016 Fourth International Conference on*, 2016, pp. 237-241.
7. D Jain, and V. Singh, "Feature selection and classification systems for chronic disease prediction: A review," *Egyptian Informatics Journal*, 19.3, 2018, pp. 179-189.
8. H. Frohlich, O. Chapelle, and B. Scholkopf, "Feature selection for support vector machines by means of genetic algorithm," *Proceedings. 15th IEEE International Conference on Tools with Artificial Intelligence*. IEEE, 2003.
9. R. Kohavi, and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence* 97.1-2, 1997, pp. 273-324.
10. E. Kabir, M. Shahid, and M. Rokibul, "Developing Diabetes Disease Classification Model using Sequential Forward Selection Algorithm," *International Journal of Computer Applications*. 180 no 5, 2017, pp. 1-6.
11. A. Jakka, and V. Rani, "Performance Evaluation of Machine Learning Models for Diabetes Prediction," *10.35940/ijitee.K2155.0981119*, 2019, pp. 1976-1980.
12. Q. Dai, C. Zhang, and H. Wu, "Research of decision tree classification algorithm in data mining," *International Journal of Database Theory and Application*, 9.5,2016, pp. 1-8.
13. I. Rish, "An empirical study of the naive Bayes classifier," *IJCAI 2001 workshop on empirical methods in artificial intelligence*, Vol. 3. No. 22, 2001, pp. 41-46.
14. D. Srivastava, and L. Bhambhu, "Data classification using support vector machine," *Journal of Theoretical and Applied Information Technology*, 12, 2010, pp. 1-7.
15. J. Nayak, B. Naik, and H. Behera, "A comprehensive survey on support vector machine in data mining tasks: applications & challenges," *International Journal of Database Theory and Application*, 8.1, 2015, pp.169-186.

AUTHORS PROFILE

Emrana Kabir Hashi is working as an Assistant Professor in Department of Computer Science & Engineering at Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh. She completed B. Sc and M. Sc. in CSE from RUET. Her research interests include Data Mining, Machine Learning, Artificial Intelligence, GIS and VRPs.

Dr. Md. Shahid Uz Zaman is working as a Professor in Department of Computer Science & Engineering at Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh. He completed his B. Sc from RUET, M. Sc. From Shanghai University, China and Ph.D from University of the Ryukyus, Japan. His areas of specialization include Machine Learning, GIS-based mapping, VRPs, Satellite imaging, Database Management System and Algorithms.