Design Engineering

ISSN: 0011-9342 | Year 2021
Issue: 6 | Pages: 3357- 3369

# AN EARLY DISEASE PREDICTION MODEL FOR DIABETES PATIENTS WITH MACHINE LEARNING APPROACH

**Dr. R. Gunavathi[1]*, B. Senthil Kumar[2]**

[1]*Associate professor, School of Commerce, Finance & Accountancy, Christ University, India.

[2]Assistant Professor, Department of Computer Science, Sree Narayana Guru College.

***Abstract***

Now-a-days most of the people are affected by diabetes disease and this leads to create more problems such as cardiovascular disease and other health issues. Several methods exist to identify the diabetes disease in an earlier stage but they are failing to predict the diabetes disease in an accurate manner and also consume more time. To overcome these issues the effective diabetes detection techniques are proposed. Initially, the amount of training data is increased with sliding window concept. Then imbalance dataset are balanced using the adaptive sampling technique. Further the diabetic recognition process is improved by applying the Intensity Weighted Firefly Optimization firefly techniques (IWFO). This method selects the features based on the correlation between the features which reduces the irrelevant features involved in the diabetic recognition process. Then PCA based feature transformation technique is applied to handle the different type of feature. The selected features are classified by Hybrid random forest into two classes namely normal and diabetes. The network predicts the disease based on the relevance between each feature which helps to maximize the prediction accuracy. Further the efficiency of diabetes prediction system is enhanced by applying the present model on different dataset. At last, efficiency of the system is evaluated using Java based simulation results. From the analysis, it is clear that the proposed model predicts the diabetes disease with the minimum misclassification error rate and maximum accuracy when compared to the other methods.

**Keywords :** Machine learning, Fire fly optimization, feature selection, Random forest, Diabetes prediction.

## Introduction

The Healthcare industry has enormous volume of records and data but the maximum amount of this data is either properly extracted or analyzed to find out hidden information [1]. There are many challenges and research avenues presented by many researchers over the two decades to deal with the useful data available to make meaningful predictions. The growth of information

ISSN: 0011-9342 | Year 2021
Issue: 6 | Pages: 3357- 3369

technology creates revolution on medical industries to do automatic analysis of disease analysis and diagnosis [2]. To achieve this task, several knowledge discovery techniques have been used on medical records or data for various type of disease diagnosis.

By detecting the diabetes disease at an early stage, serious issues like eye problem and mortality rate can be reduced [3]. The seriousness of diabetes prediction process requires an expert system to diagnose diabetes in an earlier stage. The traditional diabetes disease prediction system is difficult to maintain and also creates complexity due to the high dimensionality of data. For overcoming the above issues various optimized methods are introduced in an expert system to predict the disease in an accurate manner.

In addition Traditional process is difficult to analyze the large set of diabetes data and difficult to predict the overfiting data from the collection of features. These difficulties are creating complexities and also consume more time that leads to make the confusion about clinical decision. Hence the earlier detecting diabetes expert system is created for predicting the abnormal diabetes features from the feature list. Hence the major objective of the diabetes prediction system is to improve the recognition rate in any situation using the optimized and effective feature selection and classification techniques. Based on the above analysis, the main research goal of the work is to enhance the diabetes prediction accuracy.

The present study focus on developing a early diagnostic model with efficient feature selection approach and best classifier. The Traditional firefly algorithm is enhanced for feature selection process. Further the feature transformation method is utilized in order to handle the various types of features. The hybrid random forest is adopted as classifier. The efficacy of the present model is evaluated using different dataset which is available publically.

The introduced diabetes recognition system utilizes the various mining and machine learning techniques to analyze the patients' data for predicting these diabetes diseases. The introduced technique has several processing steps such as class imbalance, feature selection, feature transformation and diabetes disease classification process. Each step has specific methods for processing their functionality.

The contribution of the present study:

- Increase the amount of training data with sliding window in order to improve the predictive performance.
- Handles the class imbalance issues with Ranked Cluster Based Adaptive Sampling.
- Different features are extracted and selected applying IWFO and PCA approach to improve the prediction factors.
- Hybrid Random forest is applied to predict the diabetes disease and minimize the misclassification rate.

ISSN: 0011-9342 | Year 2021
Issue: 6 | Pages: 3357- 3369

**Related work**

Ranganatha *et al*. (2013) reported that the medical filed predict the patient care activity based on the secondary research resource. This method stores the patients data who came for treatment and algorithms were allowed to run on that information card. The results were represented in a simple understandable words and graphs. The choosing of algorithm was based on the availability of data and if the data is very large then the models such as ID3 and naïve Bayesian algorithm models were used. The use of decision tree algorithm ID3 made the output generation easy and understandable.

Mane, T. U. (2017) used Hadoop MapReduce platform to big data approach. The classification model used for this model was ID3 decision tree and clustering of the model was done by k-means. The system was helpful in decision making based on the parameters such as cholesterol, chest pain, resting BP, age etc. As the patients could go for a second opinion, this factor too was addressed. This would also impact the process of treatment of heart disease.

Alghamdi *et al*. (2017) discussed about the significance of machine learning technology to predict incident diabetes using cardiac-respiratory system. Machine learning methods were gaining predominance in the healthcare community because of the improving performance and increasing use. The machine learning technique used was a random under-sampling technique, along with the five classification models this method showed no improvement. A significant improvement in prediction was done with the help of Synthetic Minority Oversampling Technique (SMOTE) method.

Malav, A. and Kadam, K. (2018) introduced a prediction model for heart disease called as hybrid approach. This algorithm was a combination of artificial neural network and k-means. The data was initially classified according to their pr operties and this classification was implemented by developing a combined algorithm using k-means and artificial neural network. The model focused on the data classification according to the cardiovascular disease that has a better reliable diagnosis.

Norma Latif Fitriyani et al [8] developed a system for type 2 diabetes early detection. The techniques applied in this study are outlier detection, balance data distribution and prediction with f isolation forest, synthetic minority oversampling technique tomek link and ensemble approach respectively. Four different dataset is utilized to evaluate the model and concluded that the data balancing step increase the model accuracy.

Kamrul Hasan et al [9] applied machine learning methods to develop a novel diabetes detection model. The model is initialized with outlier rejection, missing value elimination and feature selection based on correlation approaches. k-nearest Neighbour, Decision Trees, Random Forest, AdaBoost, Naive Bayes, and XGBoost) and Multilayer Perceptron were applied as classifier. This model was evaluated on pima dataset.

Design Engineering

ISSN: 0011-9342 | Year 2021
Issue: 6 | Pages: 3357- 3369

## Dataset

Four different dataset were used in this study and their details are listed in table 1 to table 4

**Table 1 Attribute description of Dataset I**

| Sno | Name | Type |
|---|---|---|
| 1 | Pregnancies | Numeric |
| 2 | Glucose | Numeric |
| 3 | BloodPressure | Numeric |
| 4 | SkinThickness | Numeric |
| 5 | Insulin | Numeric |
| 6 | BMI | Numeric |
| 7 | DiabetesPedigreeFunction | Numeric |
| 8 | Age | Numeric |
| 9 | Outcome | Numeric |

**Table 2 Attribute description of Dataset II**

| Sno | Name | Type |
|---|---|---|
| 1 | Pregnancies | Numeric |
| 2 | Glucose | Numeric |
| 3 | BloodPressure | Numeric |
| 4 | SkinThickness | Numeric |
| 5 | Insulin | Numeric |
| 6 | BMI | Numeric |
| 7 | DiabetesPedigreeFunction | Numeric |
| 8 | Age | Numeric |
| 9 | Outcome | Numeric |

**Table 3 Attribute description of Dataset III**

| Sno | Name | Type |
|---|---|---|
| 1 | Age | Numeric |
| 2 | Gender | Nominal |
| 3 | Family_Diabetes | Nominal |
| 4 | highBP | Nominal |
| 5 | PhysicallyActive | Nominal |
| 6 | BMI | Numeric |
| 7 | Smoking | Nominal |
| 8 | Alcohol | Nominal |
| 9 | Sleep | Numeric |
| 10 | SoundSleep | Numeric |

ISSN: 0011-9342 | Year 2021
Issue: 6 | Pages: 3357- 3369

| | | |
|---|---|---|
| 11 | RegularMedicine | Nominal |
| 12 | JunkFood | Nominal |
| 13 | Stress | Nominal |
| 14 | BPLevel | Nominal |
| 15 | Pregancies | Numeric |
| 16 | Pdiabetes | Numeric |
| 17 | UriationFreq | Nominal |
| 18 | Diabetic | Nominal |

**Table 4 Attribute description of Dataset III**

| Sno | Name | Type |
|---|---|---|
| 1 | Age | Numeric |
| 2 | Gender | Nominal |
| 3 | Polyuria | Nominal |
| 4 | Polydipsia | Nominal |
| 5 | sudden weight loss | Nominal |
| 6 | weakness | Nominal |
| 7 | Polyphagia | Nominal |
| 8 | Genital thrush | Nominal |
| 9 | visual blurring | Nominal |
| 10 | Itching | Nominal |
| 11 | Irritability | Nominal |
| 12 | delayed healing | Nominal |
| 13 | partial paresis | Nominal |
| 14 | muscle stiffness | Nominal |
| 15 | Alopecia | Nominal |
| 16 | Obesity | Nominal |
| 17 | class | Nominal |

## Proposed Methodology

Diabetes is one of the dangerous health issues; it is created due to the improper secretion of the insulin in body. The abnormal changes of insulin create several health problems such as nerve damage, heart disease, kidney failure, high blindness and improper blood pressure. Sometimes, diabetes symptoms are difficult to find in an exact manner that leads to all serious problems. So, the diabetes disease is identified by creating the expert system. The major difficulty in diabetes prediction is a high misclassification rate due to the high dimensionality of features. These difficulties are resolved by using an optimized learning technique. One more problem is small dataset, which is solved through sliding window concept in the proposed model. The hybrid

ISSN: 0011-9342 | Year 2021
Issue: 6 | Pages: 3357- 3369

random forest with IWFO and PCA is utilized to recognizing abnormal diabetes information with the help of the different steps such as selection and classification process. The excellence of system is analyzed using Netbeans IDE based experimental results and discussions.

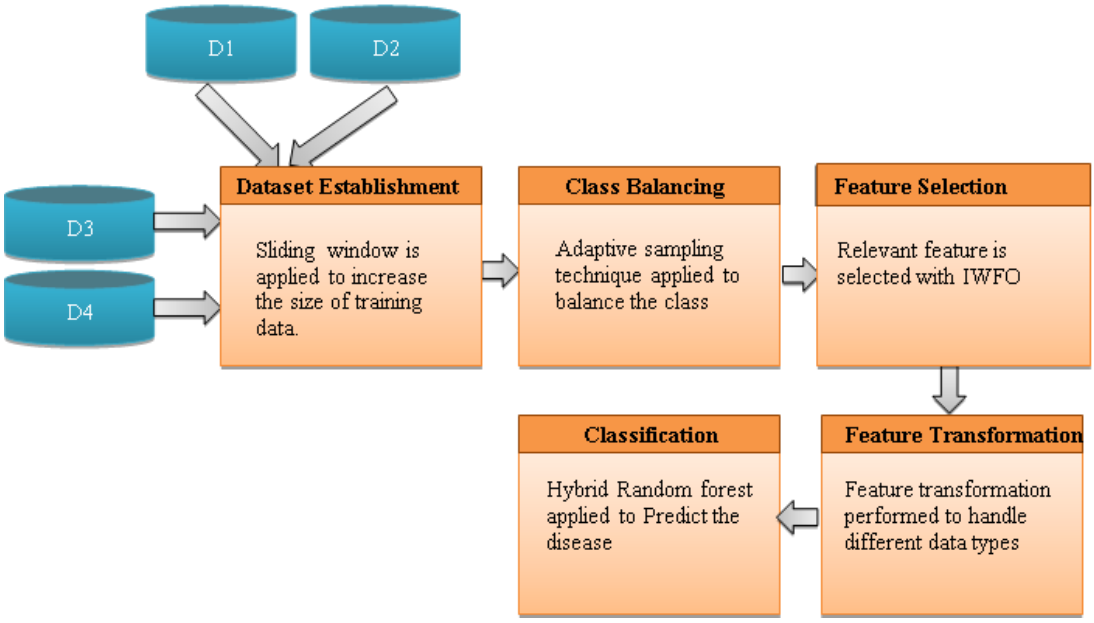



**Figure 1. Proposed framework for diabetes prediction**

## Sliding Window

The machine learning algorithms provides the efficient result for large size data. The approach require large amount of data to perfectly train the model. In online repository and other web resources contains small dataset for diabetes prediction which is further required some methods to established it. In this study the sliding window concept is utilized for dataset establishment. Zhang et al [10] applied this sliding window procedure to increase the air quality data their predictive model. As discussed in that study we applied the concept for our diabetes dataset.

The diabetes dataset contains the feature with label value by keeping this data the sliding window increase the instances by forming the high-dimensional features and labels. Through this sliding window concept the dataset can be established size of the data to ten to fifteen thousand therefore the performance of the predictive model is further improved.

## Ranked Cluster Based Adaptive Sampling

**RCBAS** technique was used to analyze the informative minority and majority set [11]. The following steps present the flow of the process. Initially the filtered minority set D may be recognized from the original minority set D. To do so, Similarity Index SI(x), for each data sample $x \in D$ was computed. After that, each x will be eliminated if its Similarity Index SI(x) carries only the majority class samples. The eliminated minority class sample known as noisy

ISSN: 0011-9342 | Year 2021
Issue: 6 | Pages: 3357- 3369

data. Hence it is sure that RCBAS removes noises effectively and also prevent the process from the noisy data. For each data sample x ∈ D, RCBAS will build a nearest majority set called N (x). The samples in N (x) can be the borderline majorities and expected to be placed near the decision boundary while the nearest majority k2 samples is small. Then all the N (x) was combined in order that it forms about the borderline majority set, D. For every value of y ∈ D RCBAS constructs N (y) and combines all such N (y) to form D.

The parameter k3 utilized in N (y) needs to be effectively adequate for including minority synthetic class samples required to generate by using RCBAS sampling approach. The disease classification using RCBAS technique is given in figure 2. The medical data is considered in which the selection of minority class samples is being done by using RCBAS Technique.

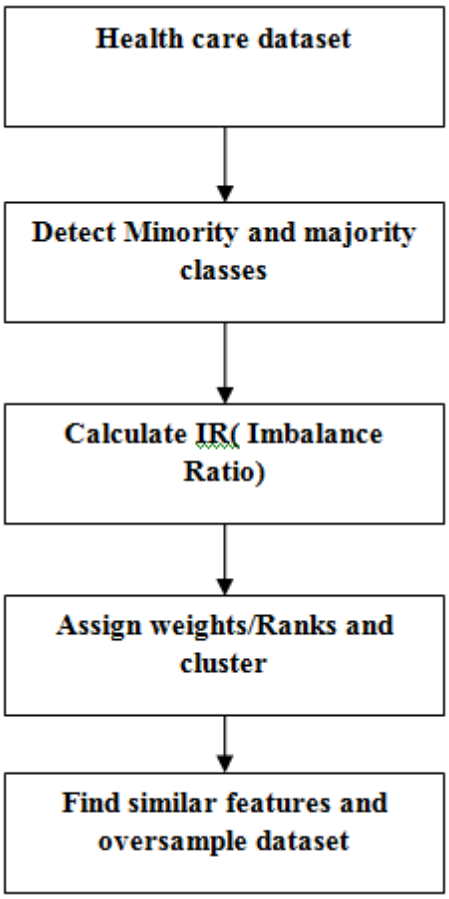



**Figure 2. Class balancing model structure**

**Feature selection**

In the second step, diabetes feature selection is done using firefly algorithm based on intensity weight. From the collected information, the optimized features are selected according to the

ISSN: 0011-9342 | Year 2021
Issue: 6 | Pages: 3357- 3369

process of firefly approach. The IWFO is one of the metaheuristic approach which select features according to the natural flashing behavior [12].

The basic firefly algorithm is enhanced with applying the intensity weight. The algorithm procedure is given below:

Step 1: load the diabetic dataset

Step 2: initialize the random variables, particles and number of population

Step 3: Define the intensity weight with random initialization of feature vector

Step 4: Update the observation coefficient

Step 5: compute the fitness value

Step 6: Repeat step 2 to 4 until the stop criteria reached

Step 7: Select the feature with best fitness value

The selected feature is taken as input to feature transformation technique which is performed using PCA algorithm discussed in next section.

**Feature transformation**

Feature transformation is performed to generate a set of features. The PCA is a most popular method which is adopted for feature transformation in the present study [13]. This method is based on mining the axes on data that displays highest variability [14]. PCA provides the great support for supervised learning which spreads out the information's in news format.

A significant problem is to decide whether a PCA-based feature transformation approach is appropriate for a certain problem or not. Meanwhile the major objective of PCA is to mine new uncorrelated features, it is logical to present some correlation-based criterion with a possibility to define a threshold value. One of such criteria is the Kaiser-Meyer-Olkin (KMO) criterion that accounts for both total and partial correlation:

$$KMO = \frac{\sum_i \sum_j r_{ij}^2}{\sum_i \sum_j r_{ij}^2 + \sum_i \sum_j a_{ij}^2} \qquad (1)$$

Where $R$ denotes the correlation matrix and $A$ denotes the partial correlation matrix. $r_{ij} = r\left(x^{(i)}, x^{(j)}\right)$ represents the elements in $R$ and $a_{ij}$ represents the elements in $A$

$$a_{ij.X^{(i,j)}}^2 = \frac{-R_{ij}}{\sqrt{R_{ii}R_{jj}}} \qquad (2)$$

ISSN: 0011-9342 | Year 2021
Issue: 6 | Pages: 3357- 3369

Where $a^2_{ij.X^{(i,j)}}$ symbolizes the partial correlation coefficient for $x^{(i)}$ $and$ $x^{(j)}$. Here $i$ and $j$ act as $X^{(i,j)}$(fixed controller) and $R_{kl}$ (algebraic complement ) in R.

When two attributes exchange a common factor with other attributes then their partial coefficient will be small that representing the distinctive variance they share. Likewise KMO value will be comes under the following factor

1. $if\ a_{ij}\ close\ to\ zero\ then$ KMO value is close to one
2. $if\ a_{ij}\ close\ to\ one\ then$ KMO value is close to zero

Popelínský [10] recommends using KMO with greater than 0.6 values for PCA. The study [10] showed that PCA with KMO> 0.5 will provide the great result

**Classification**

Classifiers works on the principle of data classification, it classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data. In classifier training, a classifier is chosen to isolate the samples in the element space. On account of parametric classifiers, the parameters are assessed in view of the samples used for training. In testing phase, the classifier identifies the input pattern by assigning a sample pattern based on the input features. The execution of any example acknowledgment framework n relies upon the change, highlight extraction, and characterization stages. In this thesis, the disease prediction accuracy is done with hybrid random forest which is discussed in the next section.

**Hybrid random forest**

The random forest is an effective classifier for disease prediction. However a single decision tree of RF will degrade the overall performance because it cannot understand a more than one logic rule during the model training. Some decision tree the may not analyze the relationship among the input and output. Hence this type of tree has to be found and removed to improve the performance of the classifier.

In this study the random forest is enhanced by removing the useless tree by pruning technique. The Back propagation neural network (BPNN) is introduced as a pruning method which effectively investigates the relationship of input on output by applying the weight. The training process of the improved RF is illustrated in figure 2.

ISSN: 0011-9342 | Year 2021
Issue: 6 | Pages: 3357- 3369

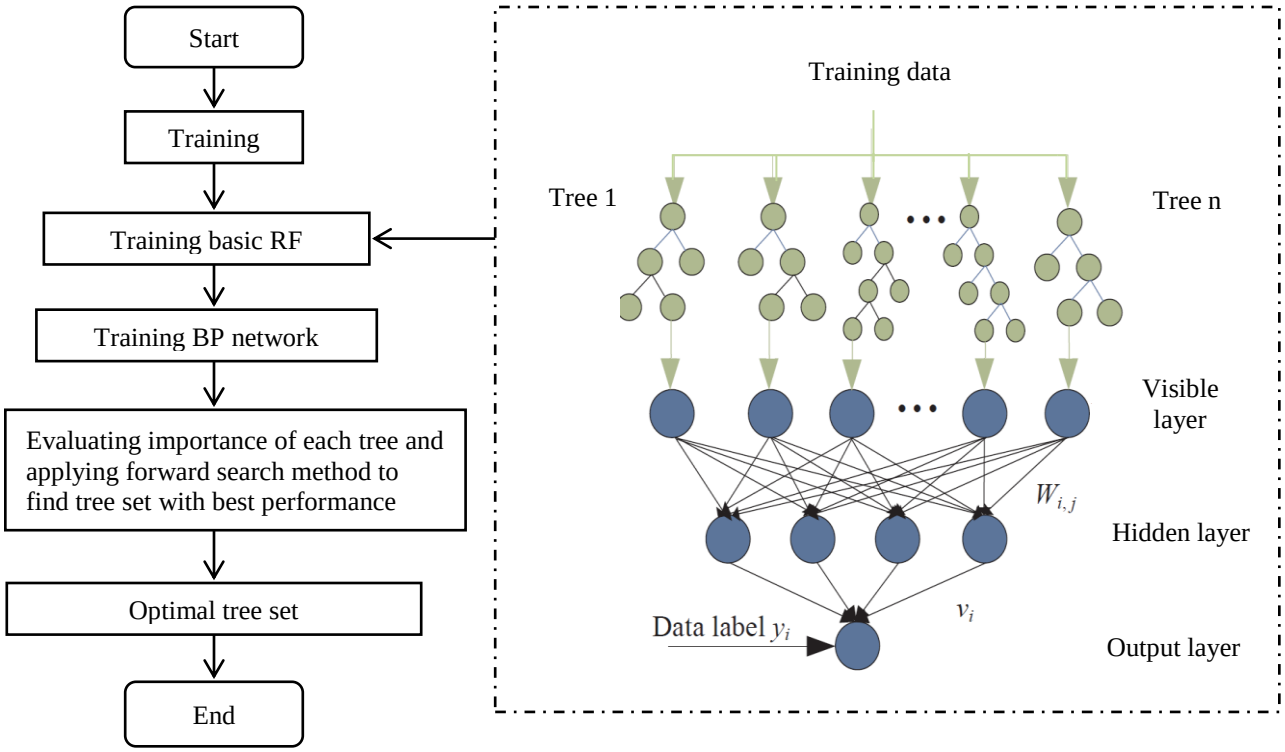



**Figure 3. BPNN based Random forest**

**Experimental Result**

The objective of this work is to discover and predict the performance of the proposed algorithm and predict the disease of affecting diabetic patients. In this section, efficiency of the proposed end to end diagnostic model is evaluated. The experiment was performed with netbeans IDE. To evaluate the present model four different dataset is used and their performance is showed in below table 5

Table 5. Performance comparison with respect to different dataset

| **Algorithm** | **Precision** | **Recall** | **F-measures** | **Accuracy** |
|---|---|---|---|---|
| Dataset I | 91% | 95% | 94% | 96.8% |
| Dataset II | 93% | 97% | 96% | 97.5% |
| Dataset III | 94.5% | 95.9% | 94.2% | 97.2% |
| Dataset IV | 96.1% | 97.5% | 97.3% | 97.9% |

ISSN: 0011-9342 | Year 2021
Issue: 6 | Pages: 3357- 3369

Table 5 provides the classification performance with necessary performance metrics. In all the Dataset, the proposed HRF approach has produced higher prediction accuracy than any other method

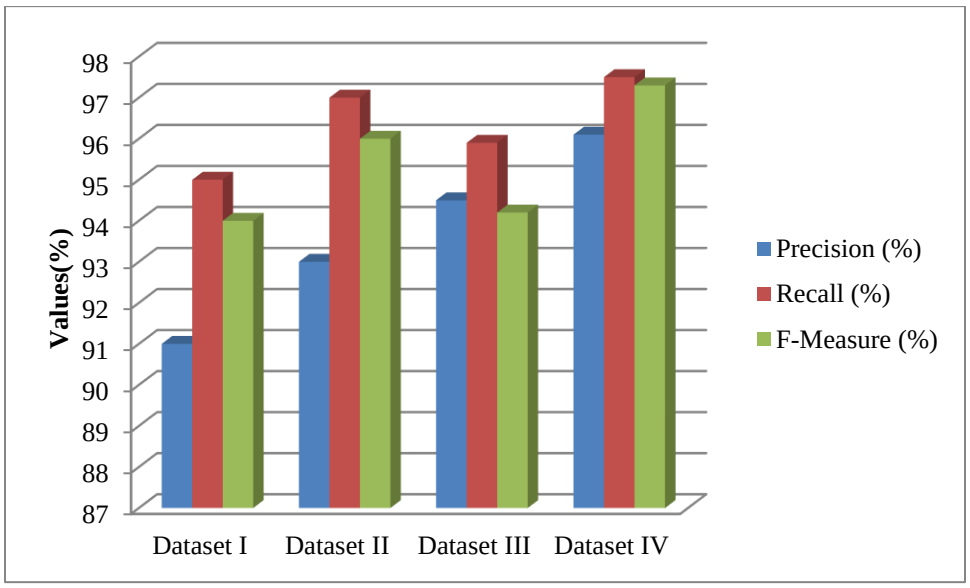
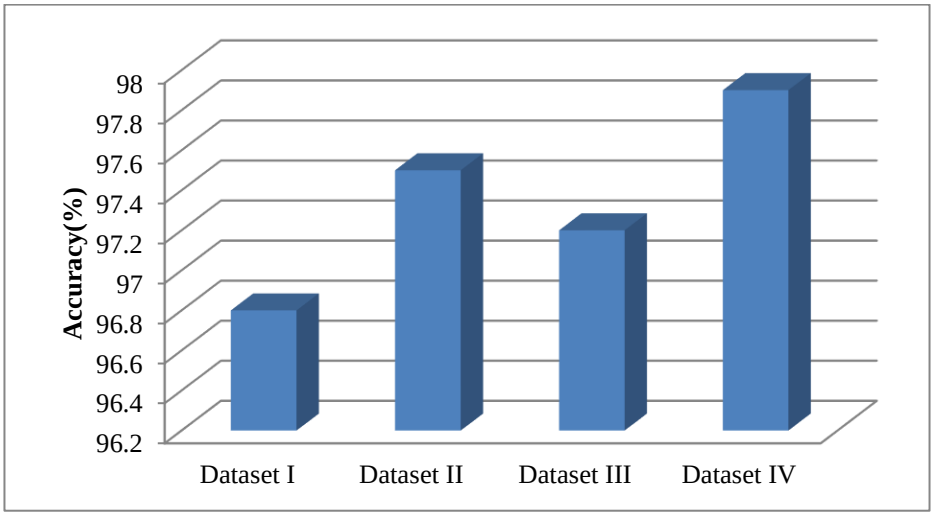



**Figure 4. Performance result of HRF**

In order to compare and validate the findings, the system is tested with the dataset that contains the features from diabetic. The important features of all dataset are considered for the implementation. Prediction accuracy for diabetes measured using hybrid random forest algorithm is shown in Table 4.6 which shows that the performance of the model is increased with increased sample size.




The performance of the proposed system is assessed using pervasive computational tests. Results of the proposed system using Hybrid random forest prediction algorithm produced an overall

ISSN: 0011-9342 | Year 2021
Issue: 6 | Pages: 3357- 3369

average accuracy of 96.8% and 97.9%. The results prove that prediction accuracy of HRF is better than any other prediction algorithm used in literature.

## Conclusion

The present work objective is to predict the diabetes in early stage to improve the quality of life of diabetes patients. From the discussion, the diabetes disease data is collected from publicly available dataset and those data are processed by discussed methods. The research focused on identifying the most essential features for predicting diabetes disease using IWFO algorithm and hybrid Random forest. Even though the optimized technologies detects diabetic disease effective manner, in the future, it will have to be improved while utilizing high dimensionality of data. The performance of the system needs to be analyzed using large volume of data with the minimum time complexity.

## References

[1]. Sivarajah, Uthayasankar, Muhammad Mustafa Kamal, Zahir Irani, and Vishanth Weerakkody. "Critical analysis of Big Data challenges and analytical methods." *Journal of Business Research* 70 (2017): 263-286.

[2]. Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future healthcare journal*, *6*(2), 94.

[3]. Deshpande, Anjali D., Marcie Harris-Hayes, and Mario Schootman. "Epidemiology of diabetes and diabetes-related complications." *Physical therapy* 88, no. 11 (2008): 1254-1264.

[4]. Ranganatha, S., Raj, H. P., Anusha, C. and Vinay, S. K. (2013). Medical data mining and analysis for heart disease dataset using classification techniques.

[5]. Mane, T. U. (2017, February). Smart heart disease prediction system using Improved K-means and ID3 on big data. In 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)., (pp. 239-245). IEEE.

[6]. Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J. and Sakr, S. (2017). Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. PloS one., 12(7).

[7]. Malav, A. and Kadam, K. (2018). A hybrid approach for heart disease prediction using artificial neural network and K-means. International Journal of Pure and Applied Mathematics., 118(8):103-110.

[8]. N. L. Fitriyani, M. Syafrudin, G. Alfian and J. Rhee, "Development of Disease Prediction Model Based on Ensemble Learning Approach for Diabetes and Hypertension," in *IEEE Access*, vol. 7, pp. 144777-144789, 2019

[9]. M. K. Hasan, M. A. Alam, D. Das, E. Hossain and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," in *IEEE Access*, vol. 8, pp. 76516-76531, 2020.

[10]. Zhang, Ying, Yanhao Wang, Minghe Gao, Qunfei Ma, Jing Zhao, Rongrong Zhang, Qingqing Wang, and Linyan Huang. "A predictive data feature exploration-based air quality prediction approach." *IEEE Access* 7 (2019): 30732-30743.

[11]. B. Senthil Kumar 1, Dr. R. Gunavathi, " Ranked Cluster Based Adaptive Sampling with Gradient Boosting Classifier for Medical Data", IOSR Journal of Computer Engineering (IOSR-JCE), Volume 20, Issue 3, Ver. III (May. - June. 2018), PP 61-67.

[12]. B. Senthil Kumar , Dr. R. Gunavathi, "An enhanced model for Diabetes prediction using Improves Firefly Feature selection and hybrid Random Forest Algorithm", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-1, October 2019.

ISSN: 0011-9342 | Year 2021
Issue: 6 | Pages: 3357- 3369

[13]. B. Senthil Kumar, R. Gunavathi, “Early prediction of diabetes using Feature Transformation and hybrid Random Forest Algorithm”, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-5, June 2020.

[14]. Pechenizkiy, Mykola, Alexey Tsymbal, and Seppo Puuronen. "PCA-based feature transformation for classification: issues in medical diagnostics." In Proceedings. 17th IEEE Symposium on Computer-Based Medical Systems, pp. 535-540. IEEE, 2004.