 doi:10.20944/preprints202002.0462.v1

Article

A New Intelligent Approach for Effective Recognition of Diabetes in the IoT E-HealthCare Environment

Amin Ul Haq[1*], Jian Ping Li[1], Jalaluddin khan[1], Muhammad Hammad Memon[1], Shah Nazir[2], Sultan Ahmad[3], Ghufran Ahmad khan[4], Amjad Ali[5]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China

[2]Department of Computer Science, University of Swabi, Pakistan

[3]Department of Computer Science College of Computer Engineering & Sciences Prince Sattam bin Abdulaziz University, Alkharj, Saudi Arabia

[4]School of information science and Technology, Southwest Jiaotong University, Chengdu, China

[5] Department of Computer Science and Software Technology, University of Swat, Pakistan

* Correspondence: Amin Ul Haq (khan.amin50@yahoo.com)

Abstract: A significant attention has been made to the accurate detection of diabetes which is a big challenge for the research community to develop a diagnosis system to detect diabetes in a successful way in the IoT e-healthcare environment. Internet of Things (IOT) has emerging role in healthcare services which delivers a system to analyze the medical data for diagnosis of diseases applied data mining methods. The existing diagnosis systems have some drawbacks, such as high computation time, and low prediction accuracy. To handle these issues, we have proposed a IOT based diagnosis system using machine learning methods, such as preprocessing of data, feature selection, and classification for the detection of diabetes disease in e- healthcare environment. Model validation and performance evaluation metrics have been used to check the validity of the proposed system. We have proposed a filter method based on the Decision Tree (Iterative Dichotomiser 3) algorithm for highly important feature selection. Two ensemble learning Decision Tree algorithms, such as Ada Boost and Random Forest are also used for feature selection and compared the classifier performance with wrapper based feature selection algorithms also. Machine learning classifier Decision Tree has been used for the classification of healthy and diabetic subjects. The experimental results show that the Decision Tree algorithm based on selected features improves the classification performance of the predictive model and achieved optimal accuracy. Additionally, the proposed system performance is high as compared to the previous state-of-the-art methods. High performance of the proposed method is due to the different combinations of selected features set and GL, DPF, and BMI are more significantly important features in the dataset for prediction of diabetes disease. Furthermore, the experimental results statistical analysis demonstrated that the proposed method would be effectively detected diabetes disease and can easily be deployed in IOT wireless sensor technologies based e-healthcare environment.

Keywords: Diabetes disease, Feature selection, E-Healthcare, Decision tree, Performance, Machine learning, Internet of things, Medical data.

 doi:10.20944/preprints202002.0462.v1

1. Introduction

Diabetes disease (DBD) is a big health issue from which many people are suffered around the world. The primary cause of this disease is associated with glucose level increase in the blood [1]. One major cause of DBD (hyper-glycemia) is the deficiency of insulin and beta cells in the pancreas produced insufficient insulin which is called type-1 DB. In type-2 DBD, the body cannot use the produced insulin accordingly [2]. The DBD is the leading cause of different other critical complications, such as kidney disease, heart disease, neurological damages, damages to the retina and damage to feet and legs [3]. In 2014, about 422 million adults were suffered from DB as compared to 108 million in 1980. The diabetes disease increased from 4.7% to 8.5% in the adult population. The DBD was the direct reason for the death of 1.6 million in 2015 and 2012, 2.2 million deaths caused by high blood glucose [4]. In 2030 DBD will the 7th major cause of death [5]. The early detection of DBD is extremely important for effective treatments but all people with DBD are unaware of their condition and even unaware until complications appear [6]. The complication of type-2 DBD can be prevented or delayed by detection in early-stage and intervention in people at risk [2], [6].

Thus, the early-stage detection of DBD is extremely necessary for effective and on-time recovery from DBD. To diagnosis the DBD, various techniques have been adopted but all these techniques have some major drawbacks to detect the DBD in its initial stages, such as computational complexity, high computation time, and less accuracy. To overcome these drawback various researchers designed different diagnostic systems to detect the DBD accurately and efficiently. Thus, the intelligent analysis of medical data including data-mining and machine learning methods which are effective approaches for the detection of DBD people. However, there are various factors to analyze for diagnosis of DBD and this complicates the job of the physicians. The medical data of the DBD and expert decision system to detect the DBD are the most important factors in diagnosis. Data-mining and machine learning-based techniques have been proposed for the detection and control of DBD by various researchers. Here, the related expert decision support systems for the diagnosis of DBD proposed by various researchers are briefly discussed.

In [7], the authors have proposed a diabetes diagnosis system used different Artificial Neural Networks, Radial Basis Function and general regression neural network. The performance of GRNN was high as compared to the Multilayer perceptron (MLP) and RBF. The GRNN achieved 80.21% accuracy. In [8], the authors designed DBD, diagnosis system and used a Multilayer neural network structure by deploying the Levenberg-Marquardt (ML) algorithm and Probabilistic neural network architecture for classification of diabetes and healthy people. They used a 10-fold cross-validation method. Kemal et al. [9] designed a two-stage diagnosis system and achieved 89.47% accuracy. In stage one input features were reduced by applied principal component analysis algorithm and the second stage adaptive neuro-fuzzy inference system was deployed for DBD diagnosis. Abdu et al. [10] proposed an intelligent system for the diagnosis of diabetes using an adaptive network-based fuzzy inference system with Modified Levenberg Marquardt algorithm. The diagnosis system achieved 82.3% accuracy. Rohollah et al. [11] developed a Logistic Adaptive Network-Based Fuzzy Inference Diagnosis system applied samples with miss values and obtained 88.05% accuracy. Humar et al. [12] proposed a hybrid Neural Network System that was developed using Artificial Neural Network and Fuzzy Neural Network for diagnosis of DBD and obtained accuracy 79.16 %. Kemal et al. [13] developed a cascade learning system based on Generalization Discriminant Analysis (GDA) and Least Square Support Vectors machine(LS-SVM) for diabetes detection. Bankat et al. [14] designed a diagnosis system that used a K-mean Clustering algorithm to eliminate incorrectly classified samples from the data set. The C4.5 algorithm achieved a high accuracy of 92.38%. Yang et al. [15] developed a diagnosis system using the Bayes network and obtained 72.3% accuracy. Muhammad et al. [16] designed a three-stage system by using genetic programming with comparative partner selection for DB detection.

Few methods have been proposed to generate rule-based classification system. Wiphada et al. [17] designed a two stages rule generated system and confirmed on many UCI datasets. If the first step neural networks nodes were pruned and analyzing the maximum weight and linguistic rules

were created utilizing frequency interval data representation. The proposed method obtained 74% accuracy. Mostafa et al. [18] proposed a framework of learning rule from the dataset and achieved 79.48% accuracy. They have designed the new update rule and give attention to the cooperation concept to generate strong rules. Fayssal et al. [19] developed a fuzzy classifier integrating with mutation operator to an Artificial Bee Colony algorithm for the creation of decision rule and obtained 84.21% accuracy. In [20] developed sampling the recursive rule extraction(Re-RX) integrated with the J48 graft algorithm for creation decision rules of the data set and achieved 83.83% accuracy. In this study [21] developed two stage hybrid model of classification and decision rule extraction(TSHDE). They used a fuzzy ARTMAP classifier with Q learning known as QFAM in the first stage and used a genetic algorithm (GA) for rule extraction from QFAM in the second stage. The proposed method obtained 91.91% accuracy. Wei et al. [43] used the point process to treat the fMRI datasets of healthy controls and patients of diabetes, and then the functional brain network of subjects is designed using two sets of BLOD signals. The proposed method performances were good.

Currently, optimization algorithms are using by researchers for decision rule generation. Binu et al. [22] developed an adaptive genetic fuzzy system (AGFS) for optimizing the rule and function of membership for the classification of medical data. Ramalingaswamy et al. [23] proposed a spider monkey optimization based rule miner (SM-RuleMiner) for diagnosis of diabetes and 89.87% accuracy achieved. They have developed a novel fitness to calculate the fitness value of each candidate rule. Mohammad et al. [42], proposed hybrid method SVR using NSGA-II method for diabetes disease detection and achieved 86.13% accuracy. Ani R et al.[61] designed IOT based E-healthcare system using ensemble classifier and the method attained 93% accuracy. Zhe Yang et al. [62] developed an IOT cloud founded wearable ECG detecting method for smart e-healthcare. Khan et al.[63] proposed IOT based smart secure health care system to facilitate the best probable patient monitoring, efficient diagnosis, and timely diagnosis of patients.

In the above review of literature of the proposed diabetes techniques drawbacks and advantages have been summarized in Table 1 for good understanding the significant of our suggested technique. All these prior recommended approaches used numerous methods to diagnosis the diabetes. Though, all these approaches have deficiency of prediction accuracy and more execution time. According to Table 1 the prediction accuracy of diabetes identification technique want more enhancements for efficient and accurate detection at early stages for better treatment and recovery. Thus, the key problems in these current methods are low accuracy and high computation time and these might be due the use of non-suitable features in dataset. To tackle these issues new approaches are required to detect diabetes properly. The enhancement in prediction accuracy is a big challenge and research gap.

In this research study, we have been designed an IOT intelligent decision system based on machine-learning algorithms to successfully detect diabetes and to ensure a treatment in the early stages. Machine learning classifier DT has been used for classification. The Filter based DT (ID3) algorithm has been proposed for suitable features selection and its performances are high as compared to other feature selection techniques, such as DT ensemble Ada Boost [50], Random forest [51], [22], [23] and wrapper based feature selection method. Different validation methods, such as Train/Test, K-Fold and Leave-One-Subject-Out (LOSO) have been used to select the best hyper parameters for the predictive model. Performance measuring metrics, such as classification Accuracy, Sensitivity, and Specificity, MCC, ROC-AUC, Precision, Recall, F1-score and Execution time are used to check the performance of the proposed system. The proposed system has been tested on the diabetes data set. Additionally, the performances of the proposed method have been compared with the state of the art methods, such as LANFIS [39], SM-Rule-Miner [23], TSHDE [21], C4.5 algorithm [14], Intelligent SVM [38], Modified K-Means Clustering +SVM (10-FC) [40] and BN [41]. The experimental results demonstrated that the proposed method Filter based (DT-(ID3) +DT) achieved high classification accuracy as compared with previous methods. All the experimental results are statistically analyzed using statistical procedures. The Generic framework of machine learning process described in Figure 1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 February 2020 doi:10.20944/preprints202002.0462.v1

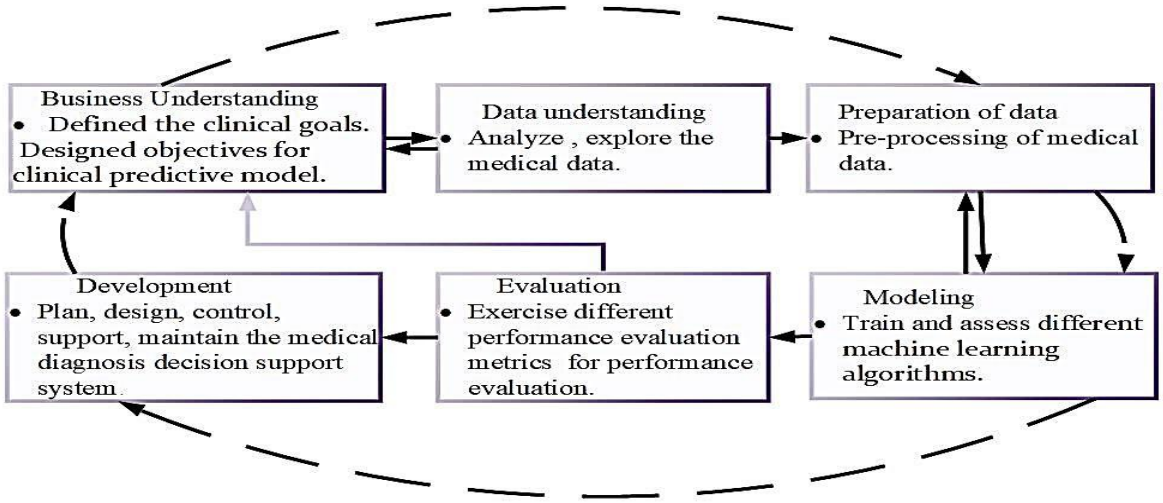


Figure 1. The generic framework of machine learning data mining process

The proposed research work is summarized in the following contributions/novelty:

I. To design IOT wireless sensor based e-health care diagnosis system for diabetes detection.

II. To propose Filter based DT-(ID3) algorithm for features selection and the proposed algorithm select more appropriate features from the dataset. Also, two DT ensembles algorithms, such as Ada Boost and Random Forest are used for feature selection and compared the performance of DT on the proposed feature selection algorithm with these two FS algorithms and wrapper based feature selection methods.

III. To use the classifier DT and the performance have been checked on original features set and on selected features set along with cross validation methods, such as Training/ testing set, K-fold, and LOSO. The LOSO is more suitable then train/test and k-folds validations. The classifier performance with LOSO validation method in terms of accuracy on selected features is high as compared to other validation methods, such as train/test and k-folds. Additional other performance evaluation metrics results are very high with LOSO validation.

IV. To recommend that the proposed method can be used to effectively detect the diabetes disease and the system can be easily incorporated in healthcare. The performance of the proposed method in terms of accuracy is high as compared to other states of the art methods and we analyzed it statistically.

The remaining parts of the paper are organized as follows. Section 2 includes the proposed method to diagnosis diabetes, a brief explanation of the data preprocessing, features selection algorithm, and theoretical and mathematical background of machine learning classifiers. The validation procedures of classifiers, such as K-fold, LOSO, Training and testing and statistical methods for comparing models are discussed in this section. The experimental setup and results are analyzed and discussed in section 3. Finally, section 4 shows the conclusion of the paper.

Table 1. Synthesized summary of the proposed methods for diabetes detection

Ref	Method	Limitations	Advantages	Accuracy (%)
[7]	Artificial Neural Networks, Radial Basis Function and general regression neural	The performances of ANN and RBF techniques are satisfactory and have high computation time and low	The GRNN obtained high performance in term of computation and accuracy.	80.21

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 February 2020 doi:10.20944/preprints202002.0462.v1

	network	prediction accuracy.		
[8]	Two stage approach	The proposed method computation time is more due the for feature searching in the dataset.	The accuracy of the proposed method is good.	89.47
[10]	Adaptive network-based fuzzy inference system with Modified Levenberg Marquardt algorithm	The method computational complex.	High accuracy due more appropriate features selection.	82.3
[9]	Hybrid system based on ANN and Fuzzy Neural Network	Computationally complex	The proposed method accuracy is high.	79.16
[14]	Diabetes diagnosis method	Computationally complex.	Low accuracy as compared others methods.	92.38
[21]	TSHDE	Low prediction accuracy.	Low time complexity.	91.91
[23]	SMRuleMiner	Low accuracy.	Low computation Time.	89.87
[42]	Hybrid method SVR using NSGA-II	High computation time.	High accuracy.	86.13
[41]	BN	Computationally complex.	Prediction performances are good.	99.51
[56]	DNN	Data set is small performance is not good	No need feature selection	95.60
[55]	DPM	The feature selection technique is complex for large data set.	Accuracy is good.	96.74
[53]	Artificial Neural Network (ANN)	Less prediction accuracy.	Computationally low complex.	82.35
[52]	SVM	High computation time.	High accuracy.	97.14
[40]	Modified K-Means Clustering +SVM (10-FC)	Low accuracy.	Low complex computationally.	96.71
[54]	SBNN+PSO+ALR	Prediction accuracy is low.	Low complex computationally.	88.75

2. Materials and Method of Research

The following sub-sections contain the explanation of the materials and methods used in this paper.

Mathematical notations used in the paper are summarized in Table 2.

Table 2. Mathematically symbols and notations used in the paper

Symbol	Description
H	Data set
S	Subset
F	Feature set
n	Number of instances in dataset
X	Input features in dataset
Y	Predicted output classes label
b	Bais is offset value from the origin
w	d-dimensional coefficient vector

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 February 2020 doi:10.20944/preprints202002.0462.v1

i	i is i^{th} sample in data set
x_i	i^{th} instance of dataset sample X
y_i	Target labels to x
R	Training set
T	Test set
t	finite set
IG(F)	information gain
P-value	Test probability value
α	Degree of freedom
f	Feature in dataset
MI	mutual information
F_i	i^{th} feature in dataset
ϕ	Empty sect
T	Transpose of matrix
p	probability
H_o	Null hypothesis
H_1	Alternate hypothesis

2.1. Dataset

In this study, the diabetes dataset was used for modeling and testing the proposed method which is available on kaggle machine learning repository [24]. Various preprocessing techniques have been applied before the feature selection process, such as min-max, variance, deviation, standardization, mean scaling and removal of missing values on the dataset [46], and [47].

2.2. Problem Statement of Feature Selection

The binary feature selection problem is described as follows:
Let us consider diabetes disease dataset that have sample set $X = \{x_1, x_2, x_3, \ldots . x_n\}$ and a finite set of "t" target label $Y = \{y_0, y_1\}$ with "r" features

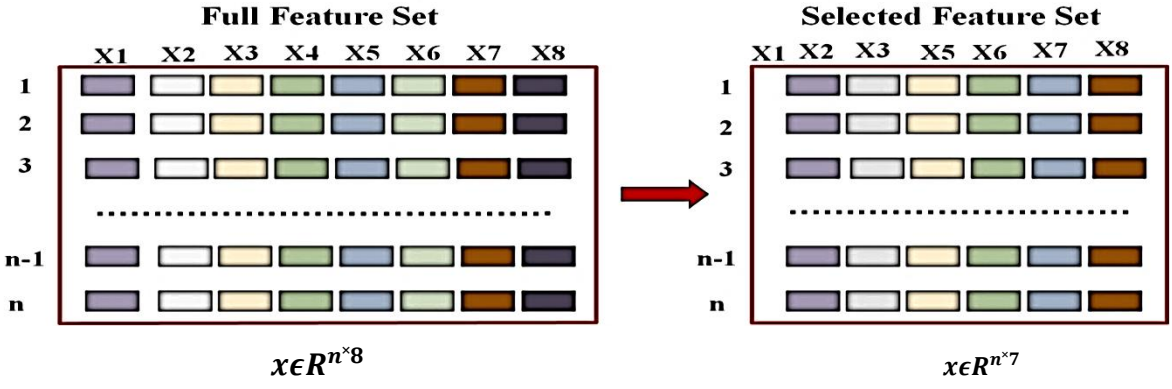
$$H = \{f_1, f_2, f_3, \ldots f_r\}.$$

The data set is expressed in equation (1) as below:

$$F(X,Y) = \{(X_i, Y_i) | X_i \in R^n, Y_i \in \{y_0, y_1\}\}_{i=1}^{k} \quad (1)$$

Where $X_i = \{x_1, x_2, x_3, \ldots x_n\} \in R^n$, are instances in the dataset and $Y_i \in \{y_0 = 0, y_1 = 1\}^t$ are output target classes' labels in the dataset.

In this equation (1) if x_i has the target label y_j then y_{ij}=1 otherwise y_{ij}=0
Additionally, $X = \{x_1, x_2, x_3, \ldots . x_n\}^T \in R^n$ is the instances matrix and $Y = \{y_o, y_1\}^T \in \{0,1\}^{n \times 1}$ is output label matrix. Figure 2 demonstrated the feature selection process.



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 February 2020 doi:10.20944/preprints202002.0462.v1

Figure 2. Feature selection process

2.2.1. Proposed Filter Based Decision Tree Approach for Feature Selection

The relevant feature selection makes our approach more effective. The feature selection process is necessary for avoiding over fitting, increase prediction performance and reduce the execution time of the classifier. Therefore, the major goal to create small subset $S = \{f_1, f_2, f_3, \dots, f_n\} (p \leq r)$ containing enough representative information. To ensure that S can achieve optimal performance, it must possess Max-relevance and Minimum redundancy properties. The filter-based method measures the relevance of a feature by correlation with the dependent variable while the wrapper feature selection algorithm measures the usefulness of a subset of feature by actually training the classifier on it. The filter method is less computationally complex than the wrapper method. The feature set selected by filter is general and can be applied to any model and it is independent of a specific model. In feature selection global relevance is of greater importance. To achieve these goals, we proposed a filter-based strategy using decision tree (DT) ID3 (Iterative Dichotomiser 3), Ada boost and Random forest algorithms for important features selection. The theoretical and mathematical background of these features selection algorithms is presented in the below sections.

2.2.1.1. Filter Based Decision Tree Iterative Dichotomiser 3 (DT-ID3) Feature Selection Algorithm

The ID3 algorithm begins with the actual data set F as the root node. In each iteration, it iterates through non used feature of the dataset F and computes the entropy $H(F)$ or information gain $IG(F)$ of that feature. Then ID3 selects feature which has the smallest entropy or largest information gain value. The Set F in then divide by the selected feature to generate subset S. ID3 uses two metrics for measuring the feature importance, such as entropy and information gain [25]-[26]. The entropy (F) is a measure of the amount of uncertainty in the dataset F which expressed in equation (2):

$$H(F) = \sum_{x \epsilon X} -p(x) log_2\, p(x) \quad (2)$$

Where F is the original data set for which entropy is being calculated, X is the features in the dataset F, and p(x) is the proportion of the number of elements in class x to the number of elements in the set F. When $H(F)$=0, the set F is perfectly classified.

The information gain $IG(F)$ is the measure of the difference in the entropy from before to after the Set F is split on feature A. It means how much uncertainty in set F was reduced after splitting set F on attribute A. Mathematically it is expressed in equation (3).

$$IG(F, A) = H(F) - \sum_{t \epsilon T} p(t)\, H(t) = H(F) - H(F|A) \quad (3)$$

Where $H(F)$ is entropy set F, T is the subsets generated from splitting set F by feature A such that $F = \bigcup_{t \epsilon T} t$, P(t) is the proportion of the number of elements in t to the elements in F, and $H(t)$ is the entropy of the subset t. The ID3 algorithm information gain can be computed for each remaining feature. The feature with high information gain is used to divide the set F on this iteration.

We summarize the pseudo-code of feature selection for diabetes disease data set in Algorithm 1.

Algorithm 1. Filter Based DT-ID3 Approach for Feature Selection
Input: Feature set F, Samples set $x_i \epsilon X$, label set $y_i \in Y$, target feature f.
Output: Selected feature subset S.
Begin
Initialization: $S = \phi, k = 1$;
1: while $F \neq \phi$ do;

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 February 2020 doi:10.20944/preprints202002.0462.v1

2: $f = \mathrm{ID3Tree}$ classifier (n_estimate);
3: $f = f.fit(X, Y)$;
4: model = select from model (f);
5:f.feature_importance
6:print (f.feature_importance);
7: Find $f \in F$;
8: S= model.tramsform (X);
9: $S_k = f$;
10: $F = F - \{f\}$;
11: $k = k+1$;
12: End while;
13: Return S;
14. Finish;

2.2.1.2. Ada Boost Feature Selection Algorithm

The AdaBoost (adaptive boosting) is ensemble decision tree algorithm [50]. It is also used for feature selection. The pseudo-Code of Ada Boost feature selection is given in Algorithm 2.

Algorithm 2. Ensemble Decision Tree Ada Boost FS algorithm

1. Initialize weights $w_{1,\ i} = \frac{1}{2m}$, $\frac{1}{2l}$ for $y_i = 0, 1$, where m and L are the number -ve and +ve instances;
2. For t=1, to T:
Normalized the weight: $w_{t,\ i} = \frac{w_{t,\ i}}{\sum_{j=1}^{n} w_{t.\ j}}$ (4)
For each feature, $j, Train\ the\ classifier\ h_j$ wich is control to using a single feature
3. The error is computed w. r. t.
w_t, $\xi_t = \sum_i w_i \left| h_j\left(x_i - y_i\right)\right|$; (5)
4. Select the classifier h_t, with the lowest error ξ_t
Modify the weights:
$w_{t+1,\ i} = w_{t,\ i}\ \beta_t^{1-e_i}$

2.2.1.3. Random Forest Feature Selection Algorithm

Random Forests (RF) is an ensemble algorithm [51]. RF is also used for feature selection and the algorithm work as follows: at each node of the tree, it randomly selects some subsets of features $f \subseteq F$. where f is the set of features. The node divides the feature into subsets f instead of F and f is smaller than F. The procedures of features selection of RF features selection algorithm are given in algorithm 3.

Algorithm 3. Ensemble Random Forest FS algorithm

1. Randomly select f features from F features et where $f \subseteq F$;
2. The node d is computed using the best split point in features f;
3. Divide the nodes into sub nodes by using the best splits;
4. repeat the steps 1 to 3 until I number of nodes is reached;

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 February 2020 doi:10.20944/preprints202002.0462.v1

5. Create forest by repeating steps 1 to 4 for n number times to generate N number of trees.

2.2.2. Wrapper Based Feature Selection Using Sequential Backward Selection Algorithm

Wrapper methods are based on greedy search algorithms as they evaluate all probable arrangements of the features. A wrapper based sequential backward selection (SBS) is a standard feature selection algorithm, which comprehended the feature space into subspace feature with the lowest latency in classifier performance and reduces the model execution time. In some cases, SBS can increase the analytical ability of the model if a model facing over fitting problem [57]. SBS sequentially removes features from the full feature space until the new feature subspace has sufficient features. To determine which feature should be removed from feature space at each phase essential to define a function of criterion J to minimize. The criterion is calculated by the criterion that is simply being the variance in the performance of the classifier before and after the elimination of a specific feature. The feature that removed at each phase can be defined as the feature that maximizes the criterion [58],[59]. The pseudo-code of the SBS algorithm is given in Algorithm 4.

Algorithm 4. Sequential Backward Selection of Feature
1: Algorithm starting with $k = d$, the d is dimensional of feature full space X_d 2: Eliminate feature x^-, that maximizes the criterion X^- = arg max $J(X_k - x)$ Where xϵ X_k 3: Eliminate feature x^- from feature space: $X_k - 1 = X_k - x^-; k = k - 1$ 4: Finish if k reached the required features, if not then repeat step 2.

2.3. Classification Algorithm

To classify diabetes and healthy people, we used the decision tree classifier in this study. A DT [27], [28] is a supervised machine learning classifier, $h: X \rightarrow Y$, that predicts the target labels related to sample x by traveling from root node of the tree to a leaf. A DT mostly applied for classification problems [29], [34]. DT structure like a tree. Every node on the root to leaf path, the successor child is selected on the basis of a splitting of the input feature. Generally, the splitting is based on one of the features of x or the predefined set of dividing rules. The leaf node possesses specific information.

2.4. Clross Validation Methods

In this study, we applied three cross validation measuring methods, such as Train/test, K-fold, and leave one subject out.

2.4.1. Training/Testing Splits

In this validation method, the samples in the data set are split for training and testing of the classifier [35]. The 70% instances are used for training and 30% are used for validation of the classifier.

2.4.1. K-Folds Cross Validation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 February 2020 doi:10.20944/preprints202002.0462.v1

In K- Folds [36] process data is split into K equal parts. The dataset split in K-1 and K-10 in each iteration for training and testing respectively. K times the process of validation executed. Average K calculation performed to achieve the classifier performance. Here we use k=10 in k Fold process. In 10-Folds validation dataset 90% used for training and 10% for testing. Finally, at the end of the 10 folds process, averages value is calculated [37]. The average estimated performance is given is calculated through equation (6).

$$E = \frac{1}{10}\sum_{i=1}^{10} E_i \quad (6)$$

2.4.2. Leave One Subject Out Cross Validation

LOSO is a cross validation special method in which samples of the data splits the an equal number of samples. One sample is used for validation and remaining all are used for training of the classifier. This method is useful for the data set of small size.

2.5. Performance Evalution Matrix

To measure classification performance of the classifier, we use different metrics in this study, such as accuracy, specificity, sensitivity, Recall, precision, MCC, F1-score and ROC curve [29], [35-36], [44 - 45] and [48]. The binary confusion matrix has been used to computes these matrices. The confusion matrix of binary classification is given in Table 3.

Table 3. Confusion matrix

	Predicted diabetes patient (+)	**Predicted healthy people (-)**
Actual diabetes patient (+)	TP	FN
Actual healthy people (-)	FP	TN

The predicted output as True positive (TP) when diabetes subject is classified as diabetes, True negative (TN) when the healthy subject is classified as healthy. False positive (FP) if a healthy subject is considered a diabetes subject, similarly False negative (FN) if diabetes subject is considered a healthy subject. With the help of these four confusion matrices, performance evaluation matrices are computed.

Accuracy (Acc): Accuracy describes the overall performance of the classifier and mathematically accuracy expressed as below in equation (7):

$$Acc = \frac{TN+TP}{TP+TN+FP+FN} \times 100\ \% \quad (7)$$

Sensitivity/Recall: Sensitive show that the diagnostic test is positive and the person has diabetes disease and it also called True Positive Rate (TPR). Mathematically written in equation (8):

Sensitivity (Sn) /Recall/True Positive Rate (TPR):

$$Sn = \frac{TP}{Tp+FN} \times 100\ \% \quad (8)$$

Specificity (Sp): Specificity describes that a predictive test is negative and the person is healthy. The specificity is expressed in equation (9):

$$Sp = \frac{TN}{TN+FP} \times 100\ \% \quad (9)$$

$$\text{Precision} = p = \frac{Tp}{TP+FP} \times 100\ \% \quad (10)$$

Major Complication or Comorbidity (MCC): MCC show the classifier predictability with value between [-1, +1].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 February 2020 doi:10.20944/preprints202002.0462.v1

If MCC is +1, it means the classifier predictions are ideal. If MCC is -1 which shows that classifier generates wrong predictions. If MCC is 0 it means that the classifier produces random predictions. The MCC is mathematically expressed in equation (11):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \times 100\ \% \quad (11)$$

F1- score: F1 score is the harmonic mean of precision and recall and mathematically expressed in equation (12):

$$F1 - score = 2 . \frac{\text{precision.recall}}{\text{precision+recall}} \quad (12)$$

ROC-AUC

The ROC is a graphical tool for model performance analysis which compares the "True Positive Rate" and "False Positive Rate" in the classification results ML classifiers. AUC characterizes the ROC of the model. A high value of AUC shows a high performance of the model.

2.6. Statistical Test for Methods Comparison

To compare the performance of machine learning supervised Models, statistical tests are necessary. In this study, we used McNamara's test [49][60] to compare the predictive performance/accuracy of the proposed method and other methods for diabetes detection. The description of McNamara's test is described below. To use McNamara's test, the instances of data set S split into a training set R and test set T. We train our method and other methods with training data and evaluating on test data set. For each instance $x \epsilon T$ of the test set we determine how classified by two models. The test is applied to a 2×2 contingency table that reports the results of two tests on an instance of n subjects as shown in Table 4.

Table 4. Contingency Table

	Test model 2 positive	**Test model 2 negative**
Test Model 1 positive	Number of instances Miss classified by Mode 1, and Model 2 or n_{00}	Number of instances misclassified by Model 1, not by Model 2 or n_{01}
Test model 1 negative	Number of instances misclassified by model 2, not by model 1or n_{10}	Number of instances misclassified by neither model 1, nor model 2 or n_{11}

Where the total number of instances in the test set is n and mathematically written in equation (13):

$$n = n_{00} + n_{01} + n_{10} + n_{11} \quad (13)$$

Two tails hypothesis, Under the null hypothesis, the two models should have the same accuracy or error rate, which can be written mathematically in equation (14):

$$H_0 : n_{01} = n_{10} \quad (14)$$

The alternate hypothesis, the two models should have different accuracy or error rate which can be expressed mathematically in equation (15):

$$H_1 : n_{01} \neq n_{10} \quad (15)$$

McNamara's test statistic is computed in equation (16):

$$p - value = teststatistic = \frac{(n_{01} - n_{10})^2}{n_{01} + n_{10}} \quad (16)$$

The selection of significance level, the test statistic or p-value interpreted as below: AS we know that the test statistic is chi square distribution with 1 degree of freedom. Additionally the confidence level and alpha are the complements of each other. The alpha is the level of significant and if the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 February 2020 doi:10.20944/preprints202002.0462.v1

value of alpha is low then the confidence level will be high and the model will be more significant. Similarly, if the alpha value is high then the confidence level will be low and the model is less significant.

If $p\ f\ \alpha$: then H0 is fails to reject, the models have no difference,

If $p \leq \alpha$: then H0 is rejected and alternate H1 is accepted the models have different performances when trained on the particular training set R.

2.7. Methodology of the Proposed Technique for Diabetes Disease Detection

The major aim of the proposed research is to detect diabetes disease effectively. In the designing of the proposed technique Decision Tree algorithm has been used for suitable feature selection. The classifier Decision tree has been used for the classification of diabetes and healthy people. Cross validation methods, such as train/test, K-fold and LOSO are used for best hyper parameters tuning of the predictive model. Additionally, different evaluation metrics are used for model performance evaluation. The diabetes data set has been used for testing of the proposed method. Data preprocessing techniques are applied before feature selection. The overall procedures of the proposed method are given in Algorithm 5 and graphically shown in the flow chart in Figure 3. The following is the procedure of the proposed method of diabetes and healthy people detection.

Algorithm 5. Proposed method

Begin
Step1: Preprocessing of the dataset using different statistical techniques;
Step2: Feature selection using DT (ID3) algorithm;
Step3: Using train/test, k folds and LOSO cross validation techniques for tuning hyper parameters and best model selection;
Step4: Classification of diabetes and healthy people using DT classifier;
Step5: Computes different performance evaluation metrics for model evaluation;
Step 6: Finish

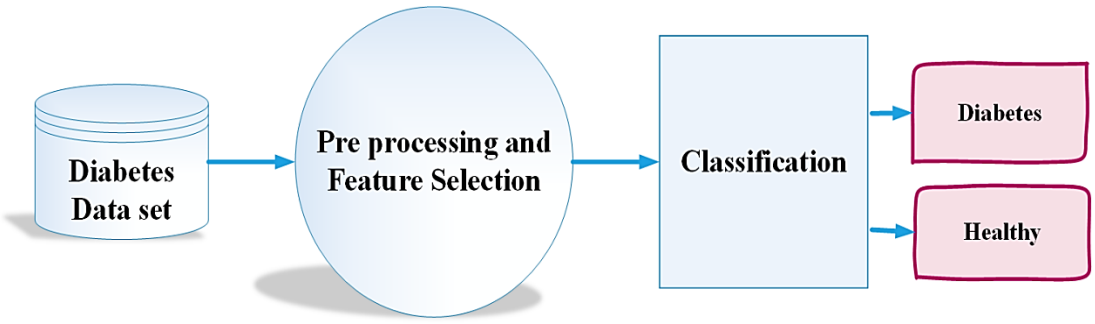


Figure 3. Flow chart of the proposed method IOT based Diabetes Detection

3. Experiments and Results Discussion

The experimental setup and results are briefly discussed in the following sub-sections.

3.1. Experimental Setup

In this study, different experiments have been performed to identify diabetes disease. In these experiments, we performed data pre-processing using different statistical techniques. Then the processed dataset has been used for feature selection. The proposed ID3 algorithm has been used for

 doi:10.20944/preprints202002.0462.v1

feature selection. Classifier DT has been trained and tested on full and on selected feature sets to evaluate the performance of DT on full and on selected features. The performance of DT classifier was good as compared to other classifiers and therefore, we only report the performance of DT in this paper. In these different validation methods, such as train/test splits, k- fold and LOSO have been used for tuning hyper parameters and best model selection. Additionally, various model performance evaluation metrics have been computed automatically for model performance evaluation, such as accuracy, specificity, sensitivity, precision, recall, f1-score, MCC and ROC-AUC cure. The experimental results are tabulated and analyzed based on full and on selected feature sets. The result of the proposed method has been compared with the state of the art methods and drawn different graphs for better presentation. Furthermore, different tools have been used for simulation of these experiments, such as Visio, Origin pro, and python on Intel@R Core TM i5, 2400 CPU, 4GB RAM with window 10.

3.2. Experimental Results

All the experimental results are reported and discussed in the below sub-sections.

3.2.1. Results of Pre-Processing Operations on the Dataset

The diabetes dataset has 2000 instances and 9 columns. The binary outcome column has two classes which take values '0' or '1' where '0' for negative case means the absence of diabetes and '1' for positive case means the presence of diabetes disease. The remaining 8 columns are real value attributes. Thus, the dataset is of 2000×8 features matrix. Furthermore, in data set 1316 are healthy subject and 684 are diabetes subjects. The dataset was generated from Type 1(DM1) diabetes patients. DM1 generally occurs in children but it can appear in aged people also. In type 1 diabetes don't produce insulin and type 2 have not enough insulin in patients

The diabetes dataset instances and attributes along with some statistical information described in Table 5. Furthermore, the visual representation of data set features shown in Figure 4 and co-relation among the features of data set visualized in Figure 5 using a heat map. The ratio of diabetes and healthy subjects in the data set graphically shown in Figure 6.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 February 2020 doi:10.20944/preprints202002.0462.v1

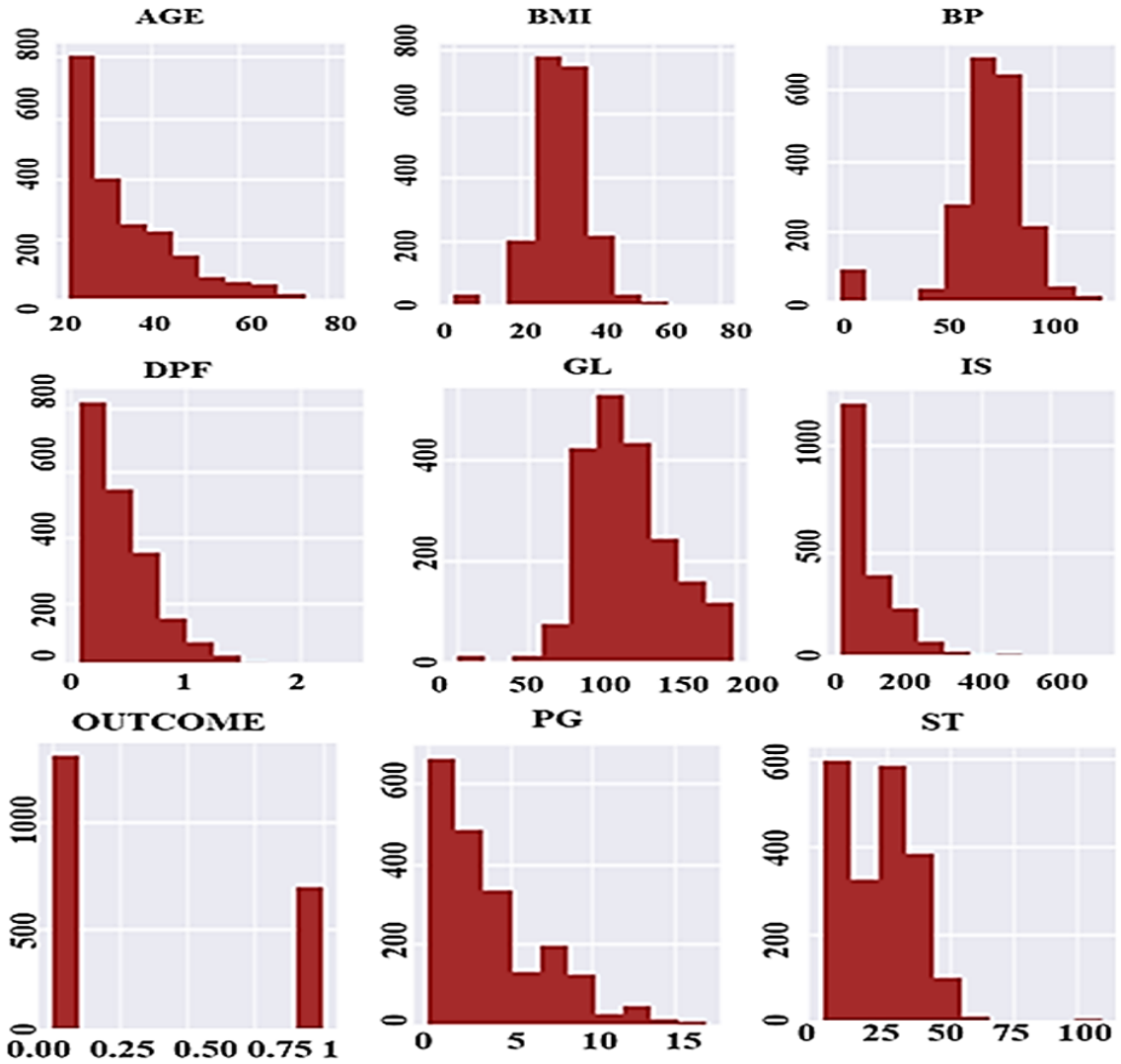


Figure 4. Histograms for the visual representation of features

Table 5. The diabetes dataset description along with some statistical operations

No	Feature Name	Feature Code	Description	Min-max	Mean, (±) STD
1	Pregnancies	PG	Number of period pregnant	0.000000-17.000000	3.703500, (±) 3.306063
2	Glucose	GL	Plasma glucose concentrations	0.000000-199.000000	121.182500, (±) 32.068636
3	Blood Pressure	BP	Blood pressures (mm Hg)	0.000000-122.000000	69.145500, (±)19.188315
4	Skin Thickness	ST	Triceps skin fold thickness(mm)	0.000000-110.000000	20.935000, (±) 16.103243
5	Insulin	IS	Serum insulin concentration	0.000000-744.000000	80.254000, (±)111.180534
6	BMI	BMI	Blood mass index	0.000000-80.600000	32.193000, (±) 8.149901
7	Diabetes Pedigree Function	DPF	Diabetes pedigree function	0.078000-2.420000	0.470930, (±) 0.323553

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 February 2020 doi:10.20944/preprints202002.0462.v1

8	Age	AGE	Age in years	21.000000-81.000000	33.090500, (±)11.786423
9	Outcome	1 = yes 0 = no	Diabetes=1 Healthy=0	0.000000-1.000000	0.342000, (±) 0.474498

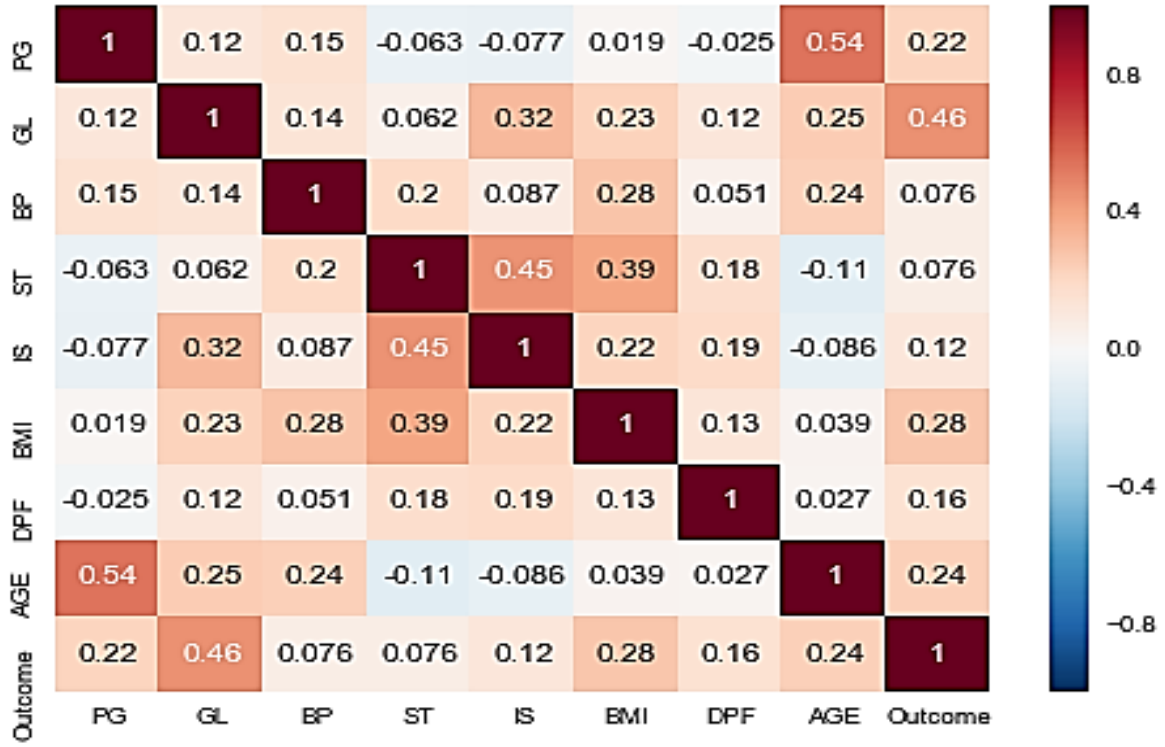


Figure 5. Heat map of the dataset

3.2.2. Experimental Results of Feature Selection Algorithm Filter Based DT(ID3)

The proposed algorithm DT (ID3) has been used in order to select more appropriate features for correct and efficient classification of diabetes and healthy people. The proposed algorithm generates a subset of features and on these selected features set the classifier shows good performance instead of the whole features set. The proposed algorithm ranked all the features as shown in Table 6. Then DT (ID3) algorithm selected importance features from whole features space. The selected features set contained features, such as GL, AGE, IS, DPE, PG, BMI, and BP. The selected features by DT (ID3) are given in Table 7. These selected features are important for the detection of diabetes disease. The selected features are graphically shown in Figure 7 for better understanding.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 February 2020 doi:10.20944/preprints202002.0462.v1

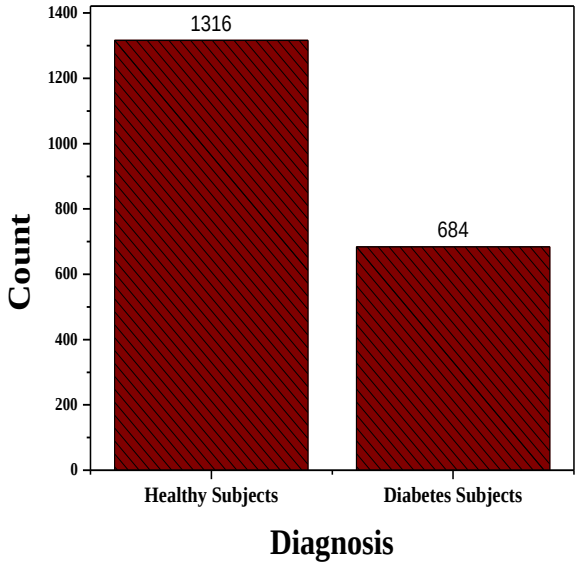


Figure 6. Ratio of diabetes and healthy subjects in the data set.

Table 6. The feature ranking and importance by DT (ID3) algorithm

S.No	Feature Label	Ranking	Score
1	PG	IS	0.07605
2	GL	ST	0.07947
3	BP	BP	0.10179
4	ST	PG	0.11071
5	IS	DPF	0.11491
6	BMI	BMI	0.13829
7	DPF	AGE	0.14366
8	AGE	GL	0.23511

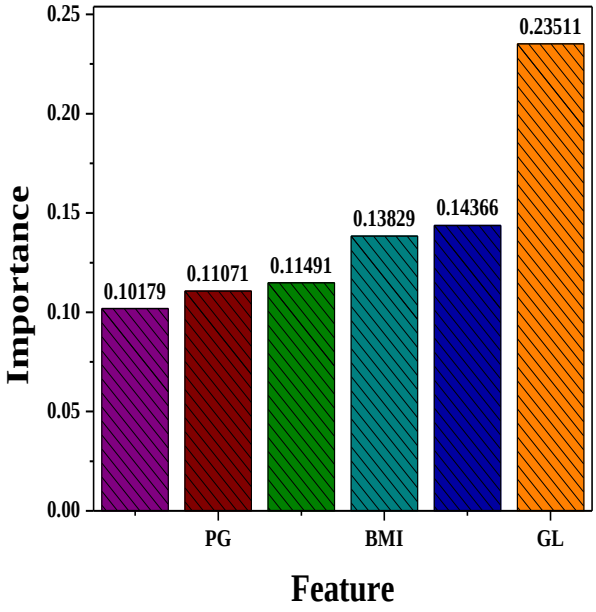


Figure 7. Feature selected by DT (ID3) algorithm

Table 7. The Rank and score of feature selected by DT (ID3), Ada Boost and Random Forest algorithm

S.NO	Feature Set	Feature selection algorithm		
		DT(ID3)	Ada Boost	Random Forest

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 February 2020 doi:10.20944/preprints202002.0462.v1

1	PG	GL	GL	BP
2	GL	AGE	BMI	GL
3	BP	IS	DPF	AGE
4	ST	DPF	BP	ST
5	IS	BMI	AGE	IS
6	BMI	BP	IS	BMI
7	DPF	PG		DPE
8	AGE			

3.2.3. Experimental Results of Ensemble Ada Boost FS Algorithm

The Ada boost is an ensemble learning algorithm. It generates a small decision tree with few features with the low computational process. The algorithm randomly selects some subset of the feature on the basis of feature weights. The features selected by Ensemble Ada boost are GL, BMI, DPF, IS, BP and AGE which are five features and these features reported in Table 7.

3.2.4. Experimental Results of Ensemble Random Forest FS Algorithm

The features selected by the Random Forest Algorithm are BP, GL, AGE, ST, IS, DPE, and BMI, which are important according to this algorithm. The features have been reported in Table 7.

3.2.5. Experimental Results of Wrapper Based Sequential Backward Selection of Feature

A wrapper based algorithm discovers the feature space to score feature subsets according to their predictive power and optimizing the subsequent induction algorithm that uses the respective subset for classification. The feature subset selected by the wrapper based sequential backward selection algorithm has been reported in Table 8. According to this algorithm, these are important features for the diagnosis of diabetes disease. The feature ST and BP are not included in the selected feature sub set. Therefore, these features have a low impact in the diagnosis of diabetes disease.

Table 8. The Feature selected by wrapper based Sequential backward selection algorithm

No of feature in set	Feature set
6	{GL, AGE, BMI, DPF, PG, IS}

3.2.6. Classification Performance of Classifier DT with Individual Feature

In this section, the classifier DT performance has been checked with the individual feature in order to identify the individual importance of each feature of the data set in the prediction of diabetes disease. The individual prediction performance on each feature has been reported in Table 9. According to Table 8 the most features are DPF, GL, BMI, IS and AGE and Classifier Achieved high accuracy on these features. The Feature DPF achieved 84 % test accuracy, 84% 10 folds average accuracy and 83 % accuracy with LOSO validation method. Similarly, the second most important feature is GL and classifier DT achieved 75% accuracy only on this feature,10 folds and LOSO based validation methods achieved 77% and 76 % accuracy respectively. The third important feature in the dataset is BMI and on this feature, the classification obtained 74% test accuracy, 73% accuracy with k-folds where k is 10, and with LOSO based method achieved accuracy is 72%. Similarly, other important features in data set are IS, AGE, PG, BP, and ST on which classifier achieved good performance respectively. Thus according to classifier performance on individual features, we reached on the conclusion that in this data set feature DPF and GL are the most highly important features and these two features have great significances in the prediction of diabetes disease. The important these features also indicated from Figure 6 because these feature score values are high and GL have 0.23511 and DPF have score value is 0.14366.The features, such as GL, DPF, BMI have a low percentage of missing values and higly corelated features. The other features in the data set are low important and have loosely correlated to the target output variable. Further, these features have a low impact on the prediction of diabetes disease. Figure 8demonstrated the accuracy of individual

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 February 2020 doi:10.20944/preprints202002.0462.v1

feature and classifier DT achieved different accuracies with different validation methods. Further, more individual importance of each feature also demonstrated in Figure 8 in which GL, and DPF ROC curve and AUC values high as compared to other feature values. Thus, from Table 9, Figure 9(a), Figure 9(b) and 10 we concluded that the feature GL and DPF are most important features in diabetes disease diagnosis and have great significant importance in data set. If the features such GL and DPF are not considered in the prediction of diabetes disease then the predictive performance of DT definitely will be effected and give less accurate results. Additionally, according to Table 9, the feature selection algorithms also select these features for the effective detection of diabetes disease. However, the other features in the data along with these important features also have a great impact on the prediction performance of the classifier, DT for the diagnosis of diabetes disease. In Table 9 classification performance of the classifier, DT has been checked on full features set and feature set without GL. Thus according to Tables 9 and 10, the feature GL is critically important in the prediction of diabetes disease. The classifier achieved 97% test accuracy without GL and with GL achieved 98.2%.

The fasting blood sugar level less than 100mg/L is normal. If fasting blood sugar level from 100 to 125 mg/dl is considered normal and if its 126 mg/dl or higher the person has diabetes. Thus the level of fasting blood sugar level value is used for the classification of diabetes and healthy people. Although in this work we used machine learning classifiers to classify diabetes and healthy subjects. The classifier prediction accuracy shows the overall performance of the system and the system accurately classify healthy and diabetes subjects. The feature selection algorithm chooses suitable features for target classification. Therefore the main aim of this work to classify the healthy and diabetes subjects using diabetes data set important features. The Feature GL, DPF, and BMI are selected by all feature selection algorithms. The feature ST according to Table 9 is a low significant feature in the prediction of diabetes disease.

Table 9. Classification performance on individual features

Classifier	Feature	Acc (%)	Sn (%)	Sp (%)	MCC (%)	ROC-AUC (%)	K-fold (%)	LOSO (%)	Execution Time(s)
DT	GL	75	45	88	67	67	77	76	0.001
	BP	68	8	74	52	53	67	66	0.005
	BMI	74	45	88	66	66	73	72	0.005
	DPF	84	66	87	78	78	84	83	0.002
	IS	73	34	92	64	63	73	73	0.001
	ST	68	14	95	54	54	65	66	0.001
	PG	69	27	90	59	58	69	70	0.0009
	AGE	70	40	85	62	63	70	71	0.0018

Table 10. Classification performance on full features and features set without GL

Classifier	Feature	Acc (%)	Sn (%)	Sp (%)	MCC (%)	ROC-AUC (%)	K-fold (%)	LOSO (%)	Execution Time(s)
DT	Full with GL	98.2	100	97	99	99.0	99.0	99.8	0.006
	Without GL	97	75	82	97	97	99.5	99.7	0.005

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 February 2020 doi:10.20944/preprints202002.0462.v1

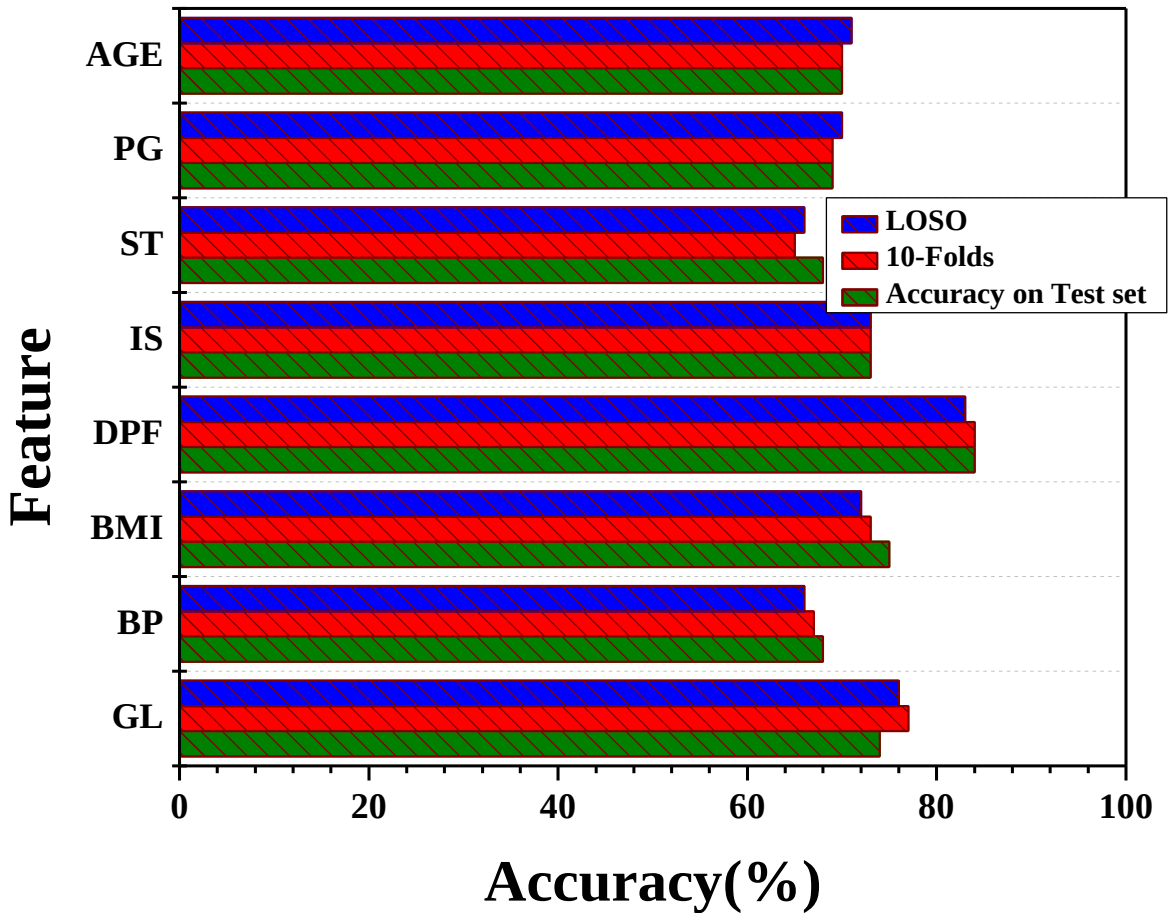


Figure 8. Classification Performance on Individual feature

3.2.7. Classification Performance on full features set and on selected features sets selected by filter-based DT (ID3), Ada boost and random forest

In these experiments, the DT classifier has been used for the classification of diabetes and healthy people. The performance of DT has been evaluated on full and on selected features set along with different cross-validation methods, such as train/test splits, k-folds and LOSO for best hyperperameters tuning and for best model selection. In train/test split method 70% instances used for training and 30% instances were used for testing. Similarly in k- fold the value of k=10 was used. The model performance evaluation metrics have been computed and shown in Table 11.

According to Table 11, DT classifier on full features set achieved 98.2% test accuracy while on selected features set selected by ID3 algorithm achieved 99% test accuracy. The specificity, sensitivity, and MCC on full features set were 97%, 100%, and 99% respectively while on the selected features set these were 99%, 100%, and 99% which are high as compared to full features set. The precision, recall and F1-score results on full features set were 99.8%, 100% and 100% on the other hand on selected features set by (ID3) the precision 100%, recall 100% and F1-score 100% which is good as compared to full features set. The ROC-AUC value of DT on full features set was 99% while on selected features set (ID3) it was 99.8% which demonstrated that on selected features set the ROC-AUC value is good and covered more area instead ROC-AUC value on full features set.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 February 2020 doi:10.20944/preprints202002.0462.v1

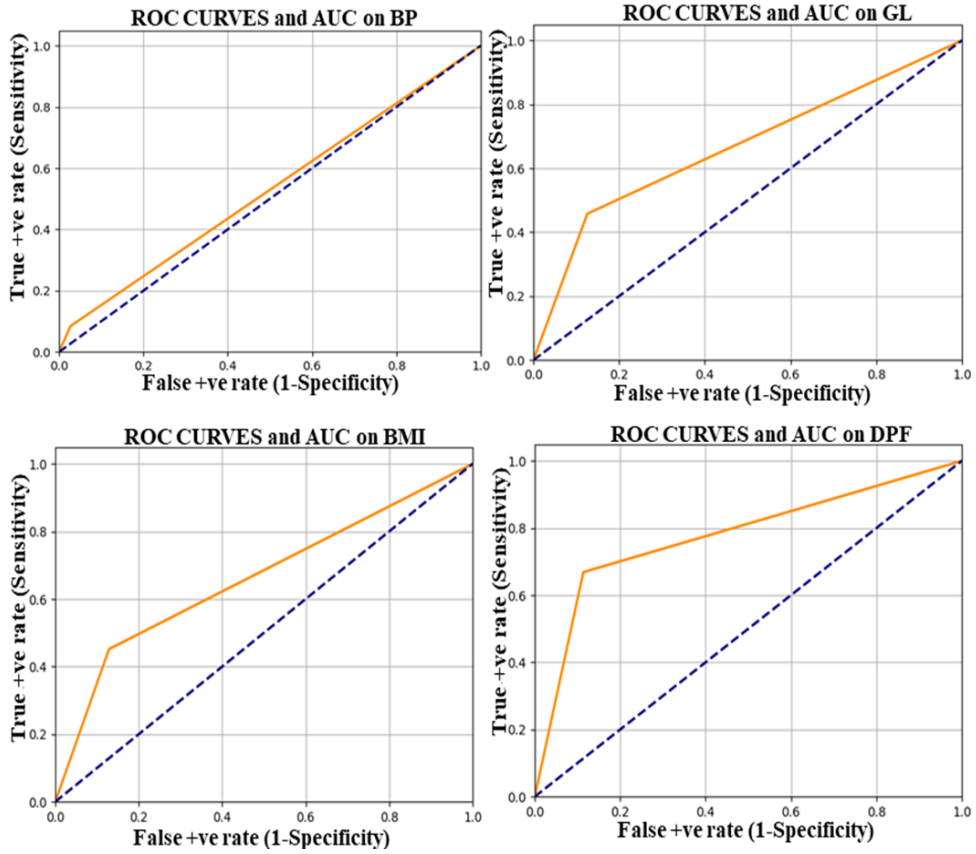


Figure 9 (a). ROC-AUC on individual feature

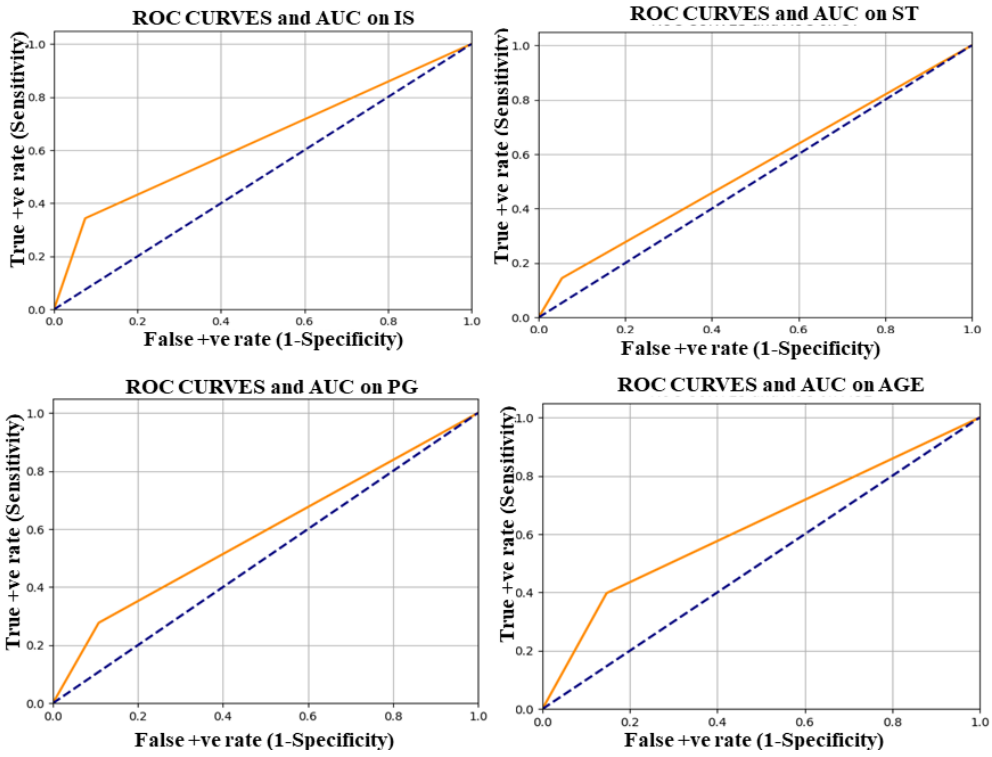


Figure 9 (b). ROC-AUC on individual feature

The 10-folds results of DT on full features set were 99.2% while on selected features set by (ID3) the 10-folds accuracy was 99.8% which is very good as compared to 10-folds value on full features

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 February 2020 doi:10.20944/preprints202002.0462.v1

set. The LOSO validation accuracy on full features set was 99.6% while on selected features set by (ID3) it was 99.9% which demonstrated that the LOSO result is good on selected features set as compared to the LOSO results on full features set. The execution time of DT on selected features set by (ID3) was 0.005 seconds while on full features set the execution time was 0.006 seconds. The confusion matrix of selected features set as selected by ID3 is shown in Figure 10. Thus, the execution time of DT decreases on selected features. The classification accuracy of DT on selected features set by FS ID3 with cross validation methods train/test, 10-folds, and LOSO are graphically shown in Figure 11 for better understanding which demonstrates that LOSO validation performance is good as compared to the performances of train/test and k fold validation. The LOSO validation achieved 100% accuracy. The F1-score value of DT on selected features set by ID3 is 100% while on full features set is 99.9% and is shown in Figure 12.

Another feature selection algorithm ADA BOOST selects important feature the data set which is reported in Table 7. The classifier performance has been checked on these selected features and reported in Table 10. The classifier DT achieved 98.5 % test accuracy, 99.3 % average accuracy of 10 folds and 99.6% accuracy with LOSO validation. Similarly, feature selection algorithm RANOM FORET selected 7 important features from the data set as we reported in Table 7. On these selected features set the classifier performances have been checked and tabulated in Table 11. According to Table 11, the classifier DT performance in term accuracy with three validation methods are 98.3% test accuracy, 99.4%, 10 folds average accuracy and with LOSO validation the accuracy is 99.7%.

The ROC curve and AUC metric performance of classifier DT with three features selection algorithms are reported in Table 11 and graphically shown in Figure 13 for better understanding. According to Figure 13 the ROC curve of DT-ID3 bigger than the ROC curves of DT-ADA-BOOST and DT-RANDOM FOREST which demonstrated that DT-ID3 performance of classification of diabetes disease and healthy are good as compared to DT-ADA-BOOST and DT-RANDOM FOREST. Additionally, the AUC value of DT-ID3 is 99.8 %, while DT-ADA-BOOST and DT-RANDOM FOREST are 98.6% and 97.7% respectively which are smaller than DT-ID3, Thus, according to these metrics DT-ID3 more suitable for the classification of healthy and diabetes disease subjects. Similarly, Classification accuracies of DT-ID3, DT-ADA-BOOST and DT-RANDOM FOREST have been reported in Table 11 and graphically demonstrated in Figure 14. The classification test accucary of DT-ID3 is 99% and DT-ADA-BOOST and RANDOM FOREST are 98.5% and 98.3% respectively. So among these methods, the DT-ID3 accuracy is high. Therefore, on the basis of accuracy metric DT-ID3 is a suitable method for the classification of healthy and diabetes disease subjects.

Furthermore, the computation times of these methods are tabulated in Table 11 and visually presented in Figure 15. According to Table 11 and Figure 15, DT-ID3 computation time is 0.005 seconds, DT-ADA-BOOST computation time is 0.004 seconds and DT-RANDOM FOREST computation Time is 0.006 seconds. Thus the computation of DT-ADA BOOST is relatively low as compared to other methods.

According to experimental results on full features, the classifier DT with different validation, such as Training/testing, k-folds, and LOSO achieved 98.2%, 99.2% and 99.6% respectively which more high as compared to the state of the art methods. Thus proposed DT classifier more suitable for this dataset as compared to other ML classifiers. Furthermore, the data preprocessing and feature selection mechanism improve the classification accuracy of DT with different validations, such as Training/testing, k-folds and LOSO achieved 99.0%, 99.8%, and 99.9% respectively. The improvement in classification accuracy due to the selection of important features by DT-ID3 FS algorithm. The ST feature according to DT-ID3 algorithm has a low impact in the prediction of diabetes disease. Thus, we think that preprocessing and feature selection is critically important for significant improvement in the accuracy of the classifier. Due to the successful detection of diabetes disease by the proposed method (DT-ID3) we recommend the proposed method for efficient and accurate detection of DB in healthcare.

Table 11. Classification performance with and without selected feature set by filter fs algorithms

Classifier	Feature set	Acc (%)	Sn (%)	Sp (%)	MCC (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)	K-folds (%)	LOSO (%)	Execution Time(s)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 February 2020 doi:10.20944/preprints202002.0462.v1

	Selected by FS algorithm										
DT	Full features set	98.2	98	97	97	99.8	98	98.6	98.0	99.2	99.6	0.006
	ID3	99.0	100	98	99	100	100	100	99.8	99.8	99.9	0.005
	Ada Boost	98.5	98	99	98	98	98	99	98.6	99.3	99.6	0.004
	Random Forest	98.3	98	98	98	95	98	99	98.7	99.4	99.7	0.006

Figure 10. Confusion matrix on selected features set by DT-ID3

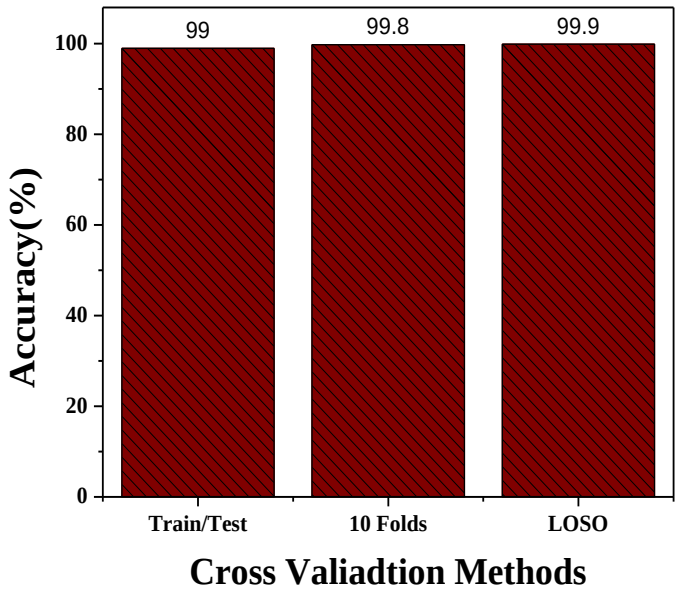


Figure 11. Accuracy on selected features set by DT-ID3 with different validation methods

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 February 2020 doi:10.20944/preprints202002.0462.v1

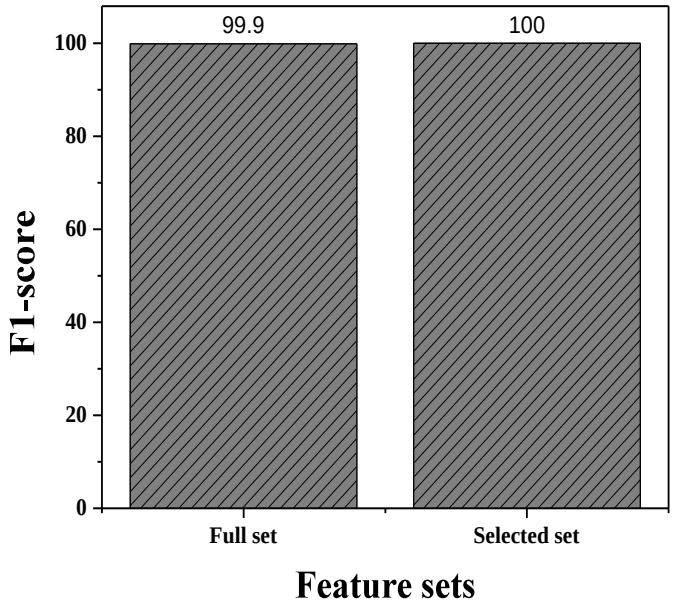


Figure 12. F1-score on full and on selected features set DT-ID3

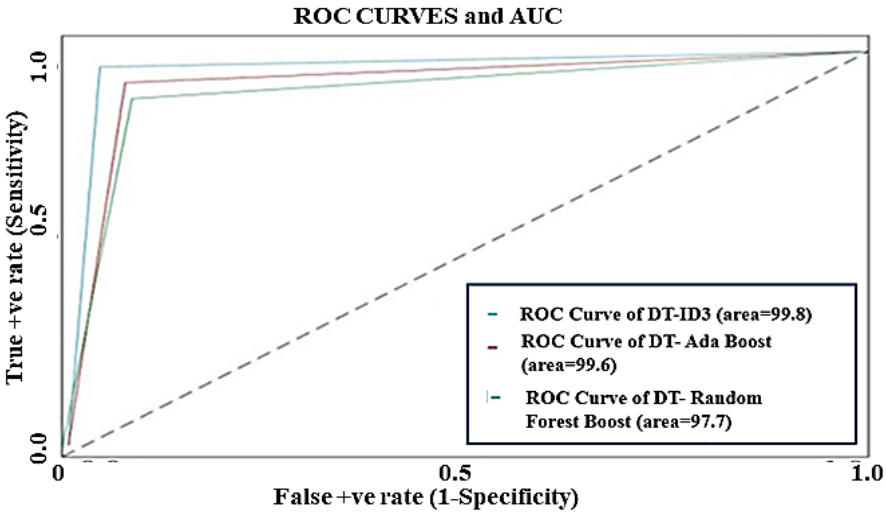


Figure 13. ROC-AUC OF DT on different selected features sets

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 February 2020 doi:10.20944/preprints202002.0462.v1

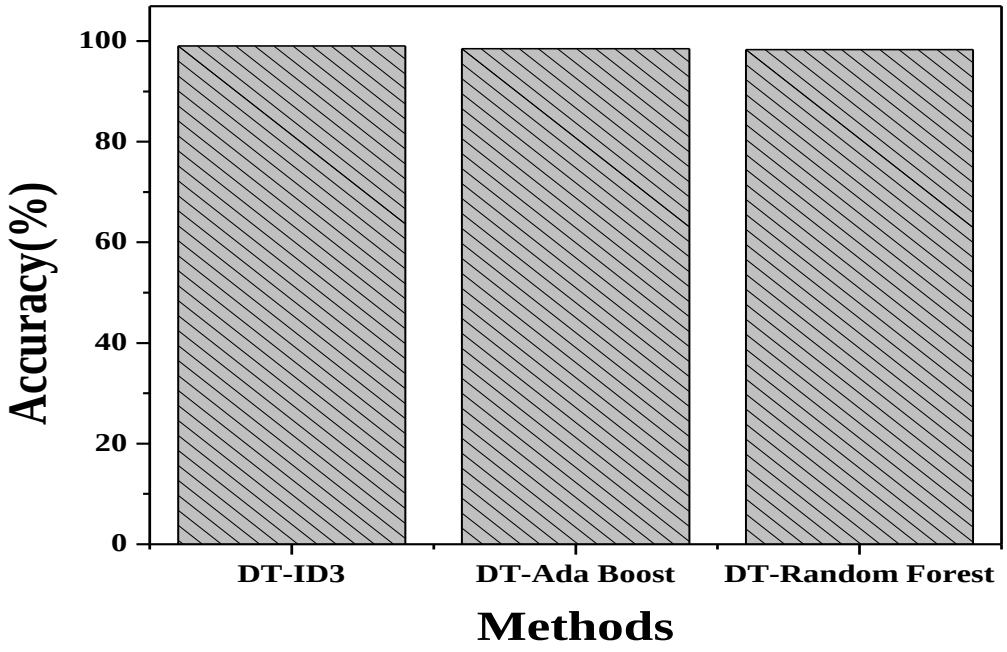


Figure 14. Accuracy of DT with different feature selection algorithms

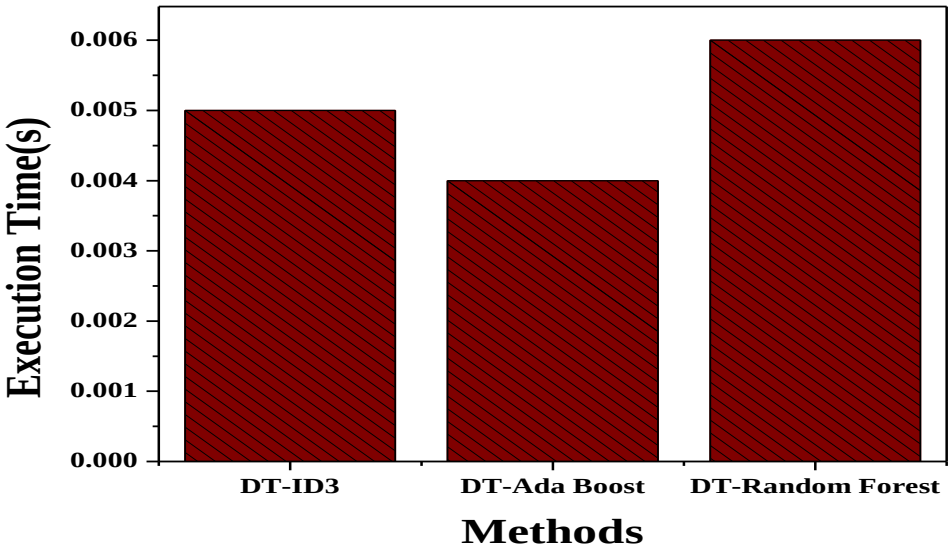


Figure 15. DT execution with different feature selection algorithms

3.2.8. Performance of Classifier on selected features set selected by Wrapper based Sequential Backward Selection algorithm

In this section, we embed the features selected by the wrapper based SBS FS algorithm in classifier DT in order to check the performance of the classifier. The experimental results have been reported in Table 11. According to Tables 11, the classifier DT achieved 98% test accuracy, 98.5% average accuracy with 10-folds and 98.9% accuracy with LOSO validation methods. Thus, we reach on the conclusion on the basis of Table 11 and 12 that the performance of Filter-based feature selection method with classifier DT is high as compared to the Wrapper based feature selection method. Furthermore, the filter based methods are computationally less complex as compared to wrapper methods and over fitting problems of filter based methods are less comparatively to the wrapper. Therefore, the propose Filter based DT-ID3 FS algorithm is more suitable for feature selection from the said dataset because the number of features in the data set is small.

Table 12. Classification performance with and without selected feature set by filter fs algorithms

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 February 2020 doi:10.20944/preprints202002.0462.v1

Classifier	Selected Feature set by FS algorithm	Acc (%)	Sn (%)	Sp (%)	MCC (%)	Precision (%)	F1-Score (%)	ROC-AUC (%)	K-fold (%)	LOSO (%)	Execution Time(s)
DT	SBS	98	99	98	98	99	98	97.6	98.5	98.9	0.007

3.2.9. Statistical Test for Comparison of DT-(ID3DT), Ada Boost and Random Forest

Statistically, we used McNamara's test in order to compare the performance of the DT-(ID3-DT), ADA BOOST AND RANDOM FOREST. The experimental hypothesis setting in such a way that $H_0 : n_{01} = n_{10}$, if the performance of DT (ID3-DT) and with other methods (ADA BOOST AND RANDOM FOREST) have equal accuracy. And $H_1 : n_{01} \neq n_{10}$

In the alternate hypothesis, the two models have accuracy differences. To test the null and alternate hypothesis we calculated the test statistic or p-value. The value of alpha for all experiments is 0.05 and the confidence level 95%. Thus on the basis of p-value and alpha, we accept or reject the null hypothesis on the following conditions.

If $p \text{ f } \alpha$: then H_0 is fail to reject, the models have no different.

If $p \leq \alpha$: then H0 is rejected and alternate H1 is accepted the models have difference performance when trained on the particular training set R.

Test-statistic or p-value is calculated for each method and reported in Table 13. The significant level is 0.05. The DT-(ID3-DT) p-value is 0.04 and DT-ADA BOOST p-value is 0.05 and DT-(RANDOM FOREST) p-value is 0.06. Since DT-(ID3-DT) p-value is less the alpha and DT-ADA BOOST p-value is equal to alpha and DT-(RANDOM FOREST) p-value is great then alpha. Thus it means that the null hypothesis is rejected and two methods have significant different in term of accuracy. It is mean that the two models are significant differences in terms of accuracy. The p-value of DT (ID3-DT) is less the alpha, therefore, DT-(ID3-DT) is more significant than the other two approaches.

Table 13. The p-value for comparison OF THREE methods

Method	P-value
DT-(ID3-DT)	0.04
DT-ADA BOOST	0.05
DT-RANDOM FOREST	0.06.

3.2.10. Performance Comparison of our method with previous methods for Diabetes Diseases detection

The performance of the proposed method (DT (ID3)-DT) was compared with the existing methods in the literature in terms of accuracy for diabetes disease detection. The proposed method obtained good results in terms of accuracy. The accuracies of the proposed method with previous methods are given in Table 14. The proposed method achieved good performance in terms of accuracy and achieved 99% test accuracy, 99.8 % k- fold average accuracy and 99.9% accuracy with LOSO validation. Hence, the proposed method could effectively diagnosis diabetes disease. Furthermore, it can be easily incorporated in the smart health care system.

Statistically, to compare the performance of the proposed method with previously proposed methods in this study we used McNamara's test. Our hypothesis that $H_0 : n_{01} = n_{10}$, if the performance of DT(ID3-DT) and with other methods have the same accuracy.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 February 2020 doi:10.20944/preprints202002.0462.v1

$H_1 : n_{01} \neq n_{10}$ The alternate hypothesis, the two models have accuracy different. To test the null and alternate hypothesis we calculated the test statistic or p-value. The value of alpha for all experiments is 0.05 and the confidence level 95%. Thus on the basis of p-value and alpha, we accept or reject the null hypothesis on the following conditions

If $p \text{ f } \alpha$: then H0 is fail to reject, the models have no difference.

If $p \leq \alpha$: then H0 is rejected and alternate H1 is accepted the models have different performance when trained on the particular training set R.

Test-statistic or p-value is calculated for each method and reported in Table 13. The significant level is 0.05. The DT-(ID3-DT) p-value is 0.04 and it is less than alpha and other methods p-values are greater the proposed method p-value. Thus it means that the null hypothesis is rejected and the methods have significant differences in terms of accuracy. The smaller p-value of DT (ID3-DT) then alpha demonstrated that DT-(ID3-DT) is more significant than previous approaches.

Table 14. Performances comparison of the proposed method with previous methods on diabetes dataset

Reference	Method	Accuracy (%)	P-value
[39]	LANFIS	88.05	0.87
[23]	SM-Rule-Miner	89.87	0.92
[21]	TSHDE	91.91	0.21
[14]	C4.5 algorithm	92.38	0.69
[38]	Intelligent SVM	94	0.48
[40]	Modified K-Means Clustering +SVM (10-FC)	96.71	0.07
[52]	Support Vector Machine	97.14	0.06
[53]	Artificial Neural Network (ANN)	82.35	1.23
[54]	SBNN+PSO+ALR	88.75	0.31
[55]	DPM	96.74	0.08
[56]	DNN	95.60	0.09
[41]	BN	99.51	0.06
Our study	DT(ID3)+DT	99(Test accuracy)	0.04
	DT(ID3)+DT	99.8(k-fold)	
	DT(ID3)+DT	99.9(LOSO)	

4. Conclusion

Internet of Things has emerging role in healthcare services which delivers a system to analyze the medical data for diagnosis of diseases applied data mining methods. In the successful detection of diabetes disease is a critical medical issue for medical experts and researchers. To tackle this problem, we have proposed a IOT based e-healthcare system for the detection of diabetes using ML datamining techniques. In the proposed method, we have used DT (ID3) algorithm for features selection as features selection is necessary for effective training and testing of the classifier.

Additionally, ensemble learning DT Feature selection algorithms ADA-BOOST and RANDOM FOREST are also used for feature selection. DT machine learning classifier has been used for the detection of diabetes. The DT has no need for extra parameters during the training and testing process. Additionally, we used different cross-validation techniques to validate the predictive model, such as train/Test splits, k-fold, and LOSO. To check the model classification performances various performance evaluation metrics have been used in this study, such as accuracy, specificity,

sensitivity, MCC, ROC-AUC, precision, recall, F1-score and execution time. The diabetes disease dataset used to check the proposed method.

The experimental results analysis demonstrated that the proposed feature selection algorithm Filter Based DT (ID3) selects more suitable features and classifier DT achieved good performances on these selected features as compared to feature sets selected by ADA-BOOST and RANDOM FOREST algorithms. The Features GL, DPF and BMI are more significantly important features in dataset and have great influence in the detection of diabetes disease and all features selection algorithms select these features. The feature ST has an impact in the detection of diabetes disease and two FS algorithms not selected it. The proposed method DT (ID3)+DT achieved 99% test accuracy, 99.8% accuracy with k-flods and 99.9% accuracy with LOSO validation. Furthermore, the classifier DT performance with Filter based feature selection method is high as compared to the wrapper based feature selection method in terms of accuracy and computation time. The experimental results of matrics used in this research are enough good. Statistical analyzed that the performance of the proposed method in terms of accuracy is good as compared to the previously proposed methods. Thus, the results of the proposed research suggest that the proposed IOT based method is more suitable for the detection of diabetes disease in healthcare. In the future, we will use an embedded based feature selection method in order to select an important feature from the data set. The proposed method will also be applied for other data sets, such as Parkinson, heart, breast cancer for efficient and accurate diagnosis of these diseases.

Acknowledgments:

This work was supported by the National Natural Science Foundation of China (Grant No. 61370073), the National High Technology Research and Development Program of China (Grant No. 2007AA01Z423), the project of Science and Technology Department of Sichuan Province.

Conflicts of Interest:

The authors declare that they have no competing interests.

References

1. N. H. Barakat, et al., "Intelligible support vector machines for diagnosis of diabetes mellitus," IEEE transactions on information technology in biomedicine, vol. 14, pp. 1114-1120, July 2010.
2. D. A. T. Edition, "International Diabetes Federation 2007," IDF (http://www. idf. org/home/index.cfm, 2014.
3. [3] A. D. Association, "Diagnosis and classification of diabetes mellitus," Diabetes care, vol. 33, pp. S62-S69, 2010.
4. W. H. Organization, World health statistics: monitoring health for the SDGs sustainable development goals: World Health Organization, 2016.
5. C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," PLoS medicine, vol. 3, p. e442, 2006.
6. M. Franciosi, et al., "Use of the diabetes risk score for opportunistic screening of undiagnosed diabetes and impaired glucose tolerance: the IGLOO (Impaired Glucose Tolerance and Long-Term Outcomes Observational) study," Diabetes care, vol. 28, pp. 1187-1194, 2005.
7. K. Kayaer and T. Yıldırım, "Medical diagnosis on Pima Indian diabetes using general regression neural networks," in Proceedings of the international conference on artificial neural networks and neural information processing (ICANN/ICONIP), 2003, pp. 181-184.
8. HasanTemurtas, et al., "A comparative study on diabetes disease diagnosis using neural networks," Expert Systems with applications, vol. 36, pp. 8610-8615, 2009.
9. K. Polat and S. Güneş, "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease," Digital Signal Processing, vol. 17, pp. 702-710, 20, october 2007.
10. A. M. Sagir and S. Sathasivam, "Design of a modified adaptive neuro fuzzy inference system classifier for medical diagnosis of Pima Indians Diabetes," in AIP Conference Proceedings, 2017, p. 040048.

11. R. Ramezani, et al., "A novel hybrid intelligent system with missing value imputation for diabetes diagnosis," Alexandria engineering journal, vol. 57, pp. 1883-1891, March, 2018.
12. H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases," Expert Systems with applications, vol. 35, pp. 82-89, 2008.
13. K. Polat, et al., "A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine," Expert Systems with applications, vol. 34, pp. 482-487, 2008.
14. B. M. Patil, et al., "Hybrid prediction model for type-2 diabetic patients," Expert Systems with applications, vol. 37, pp. 8102-8108, 2010.
15. Y. Guo, et al., "Using bayes network for prediction of type-2 diabetes," in 2012 International Conference for Internet Technology and Secured Transactions, 2012, pp. 471-472.
16. M. W. Aslam, et al., "Feature generation using genetic programming with comparative partner selection for diabetes classification," Expert Systems with applications, vol. 40, pp. 5402-5412, 2013.
17. W. Wettayaprasit and U. Sangket, "Linguistic knowledge extraction from neural networks using maximum weight and frequency data representation," in 2006 IEEE Conference on Cybernetics and Intelligent Systems, 2006, pp. 1-6.
18. M. F. Ganji and M. S. Abadeh, "Using fuzzy ant colony optimization for diagnosis of diabetes disease," in 2010 18th Iranian Conference on Electrical Engineering, 2010, pp. 501-505.
19. F. Beloufa and M. A. Chikh, "Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm," Computer methods and programs in biomedicine, vol. 112, pp. 92-103, 2013.
20. Y. Hayashi and S. Yukita, "Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset," Informatics in Medicine Unlocked, vol. 2, pp. 92-104, 2016.
21. F. Pourpanah, et al., "A hybrid model of fuzzy ARTMAP and genetic algorithm for data classification and rule extraction," Expert Systems with applications, vol. 49, pp. 74-85, 2016.
22. B. Dennis and S. Muthukrishnan, "AGFS: Adaptive Genetic Fuzzy System for medical data classification," Applied Soft Computing, vol. 25, pp. 242-252, 2014.
23. R. Cheruku, et al., "SM-RuleMiner: Spider monkey based rule miner using novel fitness function for diabetes classification," Computers in biology and medicine, vol. 81, pp. 79-92, 2017.
24. G. hospital Frankfurt, kaggle, "Diabetes," https://www.kaggle.com/johndasilva/diabetes, Access 5, May, 2019.
25. Chen Jin, Luo De-lin and Mu Fen-xiang, "An improved ID3 decision tree algorithm," 2009 4th International Conference on Computer Science & Education, Nanning, 2009, pp. 127-130.doi: 10.1109/ICCSE.2009.5228509.
26. Valencia, Rafael, and Juan Andrade-Cetto. "Active pose SLAM." Mapping, Planning and Exploration with Pose SLAM. Springer, Cham, 2018. 89-108.
27. X. Wu, et al., "Top 10 algorithms in data mining," Springer, pp. 1-37, 4,December 2007
28. P. W. Wagacha, "Induction of Decision Trees," Foundations of Learning and Adaptive Systems, vol. 12, pp. 1-14, 9,May 2003.
29. Amin Ul Haq et.al "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms". ", Hindawi Mobile Information Systems Volume 2018, Article ID 3860146, 21 pages https://doi.org/10.1155/2018/3860146.
30. S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," in IEEE Transactions on Systems, Man, and Cybernetics, vol. 21, no. 3, pp. 660-674, May-June 1991. doi: 10.1109/21.97458.
31. Mingers, J. (1989). An empirical comparison of pruning methods for decision tree induction. Machine learning, 4(2), 227-243.
32. Pal, Mahesh, and Paul M. Mather. "An assessment of the effectiveness of decision tree methods for land cover classification." Remote sensing of environment 86.4 (2003): 554-565.
33. Shouman, M., Turner, T., & Stocker, R. (2011, December). Using decision tree for diagnosing heart disease patients. In Proceedings of the Ninth Australasian Data Mining Conference-Volume 121 (pp. 23-30).
34. Chasmer, L., et al. "A decision-tree classification for low-lying complex land cover types within the zone of discontinuous permafrost." Remote Sensing of Environment 143 (2014): 73-84.

35. Amin Ul Haq et.al, Comparative Analysis of the Classification Performance of Machine Learning Classifiers and Deep Neural Network Classifier for Prediction of Parkinson Disease, 2018 15th International computer Conference on Wavelet Active Media Technology and Information Processing(ICCWAMTIP), IEEE,14-16 Dec 2018.
36. Amin Ul Haq et.al, "Feature Selection Based on L1-Norm Support Vector Machine and Effective Recognition System for Parkinson's Disease Using Voice Recordings" IEEE Access, issue date December 2019, volume: 7, Issue: 1, on page 37718-37734, Print ISSN:2169-3536.
37. A. Tsanas, et al., "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," IEEE Transactions on biomedical engineering, vol. 59, pp. 1264-1271, January 2012.
38. Barakat, Nahla, Andrew P. Bradley, and Mohamed Nabil H. Barakat. "Intelligible support vector machines for diagnosis of diabetes mellitus." IEEE transactions on information technology in biomedicine 14.4 (2010): 1114-1120.
39. Ramezani, Rohollah, Mansoureh Maadi, and Seyedeh Malihe Khatami. "A novel hybrid intelligent system with missing value imputation for diabetes diagnosis." Alexandria engineering journal 57.3 (2018): 1883-1891.
40. Yilmaz, Nihat, Onur Inan, and Mustafa Serter Uzer. "A New Data Preparation Method Based on Clustering Algorithms for Diagnosis Systems of Heart and Diabetes Diseases." Journal of Medical Systems 38.5 (2014): 1.
41. Alić, Berina, Lejla Gurbeta, and Almir Badnjević. "Machine learning techniques for classification of diabetes and cardiovascular diseases." 2017 6th Mediterranean Conference on Embedded Computing (MECO). IEEE, 2017.
42. Zangooei, Mohammad Hossein, Jafar Habibi, and Roohallah Alizadehsani. "Disease Diagnosis with a hybrid method SVR using NSGA-II." Neurocomputing 136 (2014): 14-29.
43. Li, Wei, et al. "Point process analysis in brain networks of patients with diabetes." Neurocomputing 145 (2014): 182-189.
44. L. Naranjo, et al., "A two-stage variable selection and classification approach for Parkinson's disease detection by using voice recording replications," Computer methods and programs in biomedicine, vol. 142, pp. 147-156, 22 Feb 2017.
45. Z. CAI, et al., "A new hybrid intelligent framework for predicting Parkinson's disease," IEEE Access, vol. 5, pp. 17188-17200, 19, September 2017.
46. Kotsiantis, S. B., Dimitris Kanellopoulos, and P. E. Pintelas. "Data preprocessing for supervised leaning." International Journal of Computer Science 1.2 (2006): 111-117.
47. Alasadi, Suad A., and Wesam S. Bhaya. "Review of data preprocessing techniques in data mining." Journal of Engineering and Applied Sciences 12.16 (2017): 4102-4107.
48. Wang, Zhiqiong, et al. "Breast Cancer Detection Using Extreme Learning Machine Based on Feature Fusion with CNN Deep Features." IEEE Access (2019).
49. Everitt, B. S. (1977). The analysis of contingency tables. London: Chapman and Hall.
50. Y.Freund, R.Shapire. A Short Introduction to Boosting. Journal of Japanese Society for Artificial Intelligence, 14(5):771-780, 1999.
51. Breiman L., (2001), Random forests, machine learning, 2001Kluwer Academic Publishers, 45(1), 5-32.
52. P. S. Kohli and S. Arora, "Application of Machine Learning in Disease Prediction," 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2018, pp. 1-4. doi: 10.1109/CCAA.2018.8777449.
53. S. K. Dey, A. Hossain and M. M. Rahman, "Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm," 2018 21st International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2018, pp. 1-5.doi:10.1109/ICCITECHN.2018.8631968.
54. F. Aofa, P. S. Sasongko, Sutikno, Suhartono and W. A. Adzani, "Early Detection System Of Diabetes Mellitus Disease Using Artificial Neural Network Backpropagation With Adaptive Learning Rate And Particle Swarm Optimization," 2018 2nd International Conference on Informatics and Computational Sciences (ICICoS), Semarang, Indonesia, 2018, pp. 1-5. doi:10.1109/ICICOS.2018.8621683.

55. 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.
56. Y. Liu et al., "Detecting Diseases by Human-Physiological-Parameter-Based Deep Learning," in IEEE Access, vol. 7, pp. 22002-22010, 2019.
doi: 10.1109/ACCESS.2019.2893877.
57. F. Ferri, et al., "Comparative study of techniques for large-scale feature selection," in Machine Intelligence and Pattern Recognition. vol. 16, ed: Elsevier, 1994, pp. 403-413.
58. P. Pudil, et al., "Floating search methods in feature selection," Pattern recognition letters, vol. 15, pp. 1119-1125, 19 ,Jun 1994.
59. C. Rother, et al., "Grabcut: Interactive foreground extraction using iterated graph cuts," in ACM transactions on graphics (TOG), 2004, pp. 309-314.
60. Amin ul haq, et al. "A novel integrated diagnosis method for breast cancer detection." Journal of Intelligent & Fuzzy Systems Preprint: 1-16.
61. Ani, R., et al. "Iot based patient monitoring and diagnostic prediction tool using ensemble classifier." 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, 2017.
62. Yang, Zhe, et al. "An IoT-cloud based wearable ECG monitoring system for smart healthcare." Journal of medical systems 40.12 (2016): 286.
63. Khan, Jalaluddin, et al. "SMSH: Secure Surveillance Mechanism on Smart Healthcare IoT System With Probabilistic Image Encryption." *IEEE Access* 8 (2020): 15747-15767.