A Comparative Analysis and Risk Prediction of Diabetes at Early Stage using Machine Learning Approach

[1]Ahmed Kareem, [2]Lei Shi, [3]Lin Wei, [4]Yongcai Tao

[1]PhD Scholar, School of Information Engineering, Zhengzhou University, Zhengzhou, China.
Email Id: oleiwi@gs.zzu.edu.cn

[2]Professor, School of Information Engineering, Zhengzhou University, Zhengzhou, China.
Email Id: shilei@zzu.edu.cn

[3]Associate Professor, School of Software, Zhengzhou University, Zhengzhou, China.
Email Id: weilin@zzu.edu.cn

[4]PhD Scholar, School of Information Engineering, Zhengzhou University, Zhengzhou, China.
Email Id: ieyctao@zzu.edu.cn

Abstract

Nowadays, diabetes is one of the fastest growing chronic life threatening diseases has become a common disease to the mankind from young to the old persons.The growth of the diabetic patients that has already affected 422 million people worldwide according to the report of World Health Organization (WHO), in 2018, now also it is increasing day by-day due to various causes such as bacterial or viral infection, toxic or chemical contents mix with the food, auto immune reaction, obesity, bad diet, change in lifestyles, eating habit, environment pollution, etc. Hence, diagnosing the diabetes is very essential to save the human life from diabetes. Around 50% of all people suffering from diabetes are undiagnosed because of its long-term asymptomatic phase is a chronic disease or group of metabolic disease where a person suffers from an extended level of blood glucose in the body, which is either the insulin production is inadequate, or because the body's cells do not respond properly to insulin. The objective of this research is to make use of significant features, design a prediction algorithm using Machine learning and find the optimal classifier to give the closest result comparing to clinical outcomes. Moreover, this paper presents a diabetes prediction system to diagnosis diabetes and to improve the accuracy in diabetes prediction using medical data with various machine learning algorithms.Finally, the result shows the Multilayer Perceptron (MLP) algorithm and the Radial Basis Function Network (RBF/RBFN) has the highest specificity of 95% and 98.72%, respectively holds best for the analysis of diabetic data. Using tenfold Cross- Validation evaluation techniquesRadial Basis Function Network outcome states the best accuracy of 98.80%.

Keywords *Diabetes risk, Symptom, Early stage, RF, MLP, RBFN, Evaluation model and Performance Metric.*

1. INTRODUCTION

Diabetes is one of deadliest diseases in theworld. It is not only a disease but also a creator of different kinds of diseases like heart attack, blindness, kidney diseases, etc. The normal identifying process is that patients need to visit a diagnostic center, consult their doctor, and sit tight for a day or more to get their reports. Moreover, every time they want to get their diagnosis report, they have to waste their money in vain. Diabetes Mellitus (DM) is defined as a group of metabolic disorders mainly caused by abnormal insulin secretion and/or action. Insulin deficiency results in elevated blood glucose levels (hyperglycaemia) and impaired metabolism of carbohydrates, fat and proteins. DM is one of the most common endocrine disorders, affecting more than 200 million people worldwide.The common symptoms of diabetes are polyuria, polydipsia, polyphagia, sudden weight loss (usually Type 1), weakness, obesity (usually Type 2), delayed healing, visual blurring, itching, irritability, genital thrush, partial paresis, muscle stiffness, alopecia, etc.

Insulin is a hormone that regulates blood sugar. Hyperglycaemia, or raised blood sugar, is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body's systems, especially the nerves and blood vessels. In 2014, 8.5% of adults aged 18 years and older had diabetes. In 2016, diabetes was the direct cause of 1.6 million deaths and in 2012 high blood glucose was the cause of another 2.2 million deaths. Between 2000 and 2016, there was a 5% increase in premature mortality from diabetes. In high-income countries the premature mortality rate due to diabetes decreased from 2000 to 2010 but then increased in 2010-2016. In lower-middle-income countries, the premature mortality rate due to diabetes increased across both periods. By contrast, the probability of dying from any one of the four main non-communicable diseases (cardiovascular diseases, cancer, chronic respiratory diseases or diabetes) between the ages of 30 and 70 decreased by 18% globally between 2000 and 2016.

1.1 Types of Diabetes:

It’s not often that people will know about the 6 different types of diabetes, let alone the most common: type 1 and type 2 diabetes. [1][2] Due to the complexity of the condition, it’s hard to properly diagnosis and distinguish between the different types of diabetes. But with more precise groupings, it will aid diagnosis and help towards responsive treatment.

1.1.1 Type 1 Diabetes

Type 1 Diabetes is an autoimmune disease that was once known as juvenile diabetes. Type 1 Diabetes occurs when the immune system attacks and destroys the insulin producing (beta) cells in the pancreas. Approximately 5% of people with diabetes have this form. Symptoms can come on suddenly and progressively worse. Symptoms of Type 1 Diabetes include: Increased Thirst, Frequent Urination, Bed-wetting (in children), Extreme Hunger, Weight Loss, Irritability, and Fatigue, Weakness, and Blurred vision. (If you notice these symptoms seek medical attention right away). People with Type 1 Diabetes need to inject insulin every day in order for the glucose they eat to be used for energy. Diet and/or exercise is NOT a cure for Type 1 Diabetes. There is no known cure, but researchers believe genetics and environmental factors play a factor [1][2].

1.1.2 Type 2 Diabetes

Type 2 Diabetes is a metabolic condition where the body doesn’t produce enough insulin or becomes resistant to it. Type 2 is the most common form and occurs in approximately 90% of people with diabetes. [1][2] It can sometimes be controlled with proper diet and exercise, or a drug to enhance sensitivity to the body’s insulin production. But sometimes natural insulin production is insufficient and insulin injections are then needed to sustain normal blood glucose levels. You if you are age 45 or older, have a family history of diabetes, or are overweight or obese.

1.1.3 Gestational Diabetes

Gestational Diabetes is a form of diabetes that is diagnosed during pregnancy. Approximately 2-5% of women pregnant women will develop this condition. Gestational Diabetes is normally detected in the middle of the pregnancy around 24 to 28 weeks. A glucose test will be conducted by giving the patient a sweet liquid to drink. If higher than normal glucose levels are detected in the urine, further testing will be done to verify if the patient is producing enough insulin. Once there is a proper diagnosis, the patient can manage their diabetes with proper diet, exercise, and monitoring blood glucose levels. If treated effectively, there is little risk of complications. Women with gestational diabetes can have healthy babies and the condition (normally) goes away after delivery.

1.1.4 LADA

LADA stands for Latent Autoimmune Diabetes of Adulthood. Like Type 1 Diabetes, LADA or (Type 1.5) occurs when the body stops producing adequate insulin. The difference is LADA progresses slowly and insulin may still be produced even after diagnosis.LADA is usually diagnosed in adulthood. LADA often gets confused and misdiagnosed with Type 2 Diabetes because of the same symptoms. Proper diagnosis of LADA is difficult and requires proficient testing of antibodies. The treatment of LADA patients will be similar to Type 1 Diabetes once insulin production is gone completely.

1.5 MODY

MODY or (Maturity Onset Diabetes of the Young) is a rare form of diabetes. MODY is caused by a mutation or change in a single gene disrupting insulin production. MODY affects 1-2% of people with diabetes. It is normally diagnosed in ages 20 and younger but can affect any age.MODY is a dominant genetic condition meaning a gene can be inherited and passed down by either mother or father. There are 11 different types of diabetes (MODY) and diagnosis will determine different treatments.MODY 1, 3, and 4 can be managed with a type of medicine called sulfonylurea therapy.MODY 2 can be treated with a proper diet and exercise.MODY 5 may need multiple treatments because it can affect other health problems.MODY 7-11 was recently discovered and patients will likely respond to treatments given to other types of MODY.[3]

1.6 NDM (Neonatal Diabetes Mellitus)

NDM a monogenic form of diabetes – like MODY.NDM is diagnosed anywhere from birth to 6 months old. It accounts for 1 in 400,000 infants in the United States. NDM is often mistaken for Type 1 Diabetes, but Type 1 is rarely seen in patients before 6 months of age. There are 20 different genes that can cause NDM. It can be temporary and disappear later in life called transient neonatal diabetes mellitus (TNDM) or permanent (PNDM). NDM can be present at conception and affect fetuses' growth and development. [3]

2. LITERATURE REVIEW

Heart disease is one of the most common and lethal complication associated with type 2 diabetes which is why research into preventing its impact is so important. The Diabetes Australia Research Trust is supporting research at the Baker Heart and Diabetes Institute where researchers believe they may have identified a potential new therapy. Dr.Arpeeta Sharma said that while many people with type 2 diabetes were taking medication to control blood pressure or cholesterol levels, they did not always respond adequately to reduce their risk of heart disease. People with diabetes who don't respond adequately to standard blood pressure lowering medication can be at an increased risk of a range of diabetes-related cardiovascular complications, Dr Sharma said. “Recent research points to the role inflammatory processes play in damaging blood vessels and impairing the heart's function. “If we can reduce the impact of inflammation on the heart then we may be able to develop more effective ways of treating diabetes-related heart conditions. “With the help of a grant from the Diabetes Australia Research Trust we will be studying a small molecule inflammasome inhibitor which we think we can show helps decrease the impact of inflammation and, in turn, helps reduces the likelihood of diabetes-related heart damage.”[4]

Insulin resistance is the process where the body stops responding to insulin like it should. It can lead to type 2 diabetes. “Most studies of insulin resistance look at how glucose is transported into tissues within the body after it is consumed in a meal, however we have exciting data that suggests the way glucose is processed within tissues may be just as important,” Dr Krycer said. “Using a new technique known as metabolomics we are able to see how the body processes glucose whether that is burning it for energy or storing it away in the form of fat or glycogen. “Understanding how this change during insulin resistance can help us develop targeted strategies that improve the way the body processes glucose."[4]

One of the big challenges surrounding the type 2 diabetes epidemic is tackling the inter-generational cycle of the disease. What do we mean by the inter-generational cycle? When someone develops type 2 diabetes it can cause changes in the body that may increase their children's risk of developing type 2 diabetes which, in turn, increases their children's risk of developing type 2 diabetes. However, a new study lead by Professor Mary Wlodek at The University of Melbourne is looking at ways of breaking this cycle. [4]

We sit at work. We sit at home. We sit in the car. We sit down to eat. We sit at in front of the TV. We sit at the movies. It is clear we spend a lot of our lives sitting. But, as plenty of new research seems to confirm, all this sitting is not be good for us. One thing researchers have learnt is that prolonged sitting is associated with sustained higher blood glucose levels which, over time, can contribute to the development of a range of diabetes-related complications. To date most of the research into sitting and diabetes has focused on type 2 diabetes, but a new study from Professor David Dunstan, Head of the Physical Activity Laboratory at the Baker Heart and Diabetes Institute, is looking into whether frequent breaks from sitting can help improve diabetes management in people with type 1 diabetes. A lot of the research into how prolonged sitting affects people with diabetes has focused on type 2 diabetes, but we want to see how it impacts people with type 1 diabetes," Professor Dunstan said. "One thing we know is that in people with type 2 diabetes prolonged sitting can exaggerate hyperglycaemia after meals."This could be the same for people with type 1 diabetes but nobody has done the research."We want see if people can improve their blood glucose management by breaking up periods of sitting."We are trying to build the evidence-base relating to workplace interventions that support people with type 1 diabetes."This could be a series of light exercises that can be performed in the workplace like an office walk at a light pace or simple resistance activities such as squatting and calf raises at your desk."[4]

Two population-based groups of white patients with non-insulin-dependent diabetes (NIDDM) in the United States and Australia were studied. Prevalence of retinopathy and duration of diabetes subsequent to clinical diagnosis were determined for all subjects. Weighted linear regression was used to examine the relationship between diabetes duration and prevalence of retinopathy [5].Sociodemographic and anthropometric data and data on blood pressure and blood glucose levels were obtained for 7541 adults aged 35 years or more from the biomarker sample of the 2011 Bangladesh Demographic and Health Survey (DHS), which was a nationally representative survey with a stratified, multistage, cluster sampling design. Risk factors for diabetes and pre-diabetes were identified using multilevel logistic regression models, with adjustment for clustering within households and communities. Almost one in ten adults in Bangladesh was found to have diabetes, which has recently become a major public health issue. Urgent action is needed to counter the rise in diabetes through better detection, awareness, prevention and treatment [6].

Diabetes mellitus is a serious metabolic disease, affecting people of all geographic, ethnic or racial origin and its prevalence is increasing globally [7]. Burden from this costly disease is high on the low and middle income countries (LMIC) where the impacts of modernization and urbanization have caused marked adverse changes in lifestyle parameters.In 2013, of the estimated 382 million people with diabetes globally, more than 80 per cent lived in LMIC. It was estimated that India had 65.1 million adults with diabetes in 2013, and had the 2nd position among the top 10 countries with the largest number of diabetes. This number is predicted to increase to 109 million by 2035 unless steps are taken to prevent new cases of diabetes [7]. Primary prevention of diabetes is feasible and strategies such as lifestyle modification are shown to be effective in populations of varied ethnicity [8, 9]. However, for implementation of the strategies at the population level, national programmes which are culturally and socially acceptable and practical have to be formulated which are currently lacking in most of the developed and developing countries. Early diagnosis and institution of appropriate therapeutic measures yield the desired glycaemic outcomes and prevent the vascular complications [10].

Type 2 diabetes which accounts for 85-95 per cent of all diabetes has a latent, asymptomatic period of sub-clinical stages which often remains undiagnosed for several years [7]. As a result, in many patients the vascular complications are already present at the time of diagnosis of diabetes, which is often detected by an opportunistic testing. Asian populations in general, particularly Asian Indians have a high risk of developing diabetes at a younger age when compared with the western populations [11]. Therefore, it is essential that efforts are made to diagnose diabetes early so that the long term sufferings by the patients and the societal burden can be considerably mitigated.

Diabetes Mellitus (DM) is defined as a group of metabolic disorders mainly caused by abnormal insulin secretion and/or action [12]. Insulin deficiency results in elevated blood glucose levels (hyperglycemia) and impaired metabolism of carbohydrates, fat and proteins. DM is one of the most common endocrine disorders, affecting more than 200 million people worldwide. The onset of diabetes is estimated to rise dramatically in the upcoming years. DM can be divided into several distinct types. However, there are two major clinical types, type 1 diabetes (T1D) and type 2 diabetes (T2D), according to the etiopathology of the disorder. T2D appears to be the most common form of diabetes (90% of all diabetic patients), mainly characterized by insulin resistance. The main causes of T2D include lifestyle, physical activity, dietary habits and heredity, whereas T1D is thought to be due to autoimmunological destruction of the Langerhans islets hosting pancreatic-β cells. T1D affects almost 10% of all diabetic patients worldwide, with 10% of them ultimately developing idiopathic diabetes. Other forms of DM, classified on the basis of insulin secretion profile and/or onset, include Gestational Diabetes, endocrinopathies, MODY (Maturity Onset Diabetes of the Young), neonatal, mitochondrial, and pregnancy diabetes. The symptoms of DM include polyurea, polydipsia, and significant weight loss among others. Diagnosis depends on blood glucose levels (fasting plasma glucose = 7.0 mmol/L) [13].

DM progression is strongly linked to several complications, mainly due to chronic hyperglycemia. It is well-known that DM covers a wide range of heterogeneous pathophysiological conditions. The most common complications are divided into micro- and macro-vascular disorders, including diabetic nephropathy, retinopathy, neuropathy, diabetic coma and cardiovascular disease. Due to high DM mortality and morbidity as well as related disorders, prevention and treatment attracts broad and significant interest. Insulin administration is the main treatment for T1D, although insulin is also provided in certain cases of T2D patients, when hyperglycemia cannot be controlled through diet, weight loss, exercise and oral medication. Current medication targets primarily a) saving one's life and alleviating the disease symptoms, and b) prevention of long term diabetic complications and/or elimination of several risk factors, thereby increasing longevity. The most common anti-diabetic agents include sulfonylurea, metformin, alpha-glucosidase inhibitor, peptide analogs, non-sulfonylurea secretagogues, etc. [14].

The majority of the present anti-diabetic agents, however, exhibit numerous side-effects. In addition, insulin therapy is related to weight gain and hypoglycemic events. Hence, anti-diabetic drug design and discovery is of great concern and concurrently a research challenge [15][16][17][18].Although extensive research in DM has provided significant knowledge, over the past decades, on the a) etiopathology (genetic or environmental factors and cellular mechanisms), b) treatment, and c) screening and management of the disease, there is still much to be discovered, unraveled, clarified and delineated. Through such processes, diagnosis, prognostic evaluation of appropriate treatment and clinical administration could gain significant ground toward medical handling of the disease. In such an effort, reliance on a large and fast increasing body of research and clinical data serves to establish a significant basis for safe diagnosis and follow-up treatment. Thus, data mining and machine learning emerge as key processes, contributing decisively to the decision-making clinician. The aspiration, therefore, is to link data assessment to diagnosis and appropriate decision-making in drug administration.

A person with diabetes has a condition in which the quantity of glucose in the blood [19] is too elevated (hyperglycemia). Because of any of these two cases that either enough insulin is not being produced in our body, or no insulin, or has cells that do not respond properly to the insulin the pancreas produces. This excess blood glucose eventually passes out of the body in urine [19]. So, even though the blood has plenty of glucose, the cells are not getting it for their essential energy and growth requirements.

Karahoca, Adem, and M. Alper Tunga used multivariate data partitioning method which is called Indexing HDMR to manage drug dosage. As dataset of this study, 142 diabetic of type 2 medical records were used to implement the model. As the result of this study a polynomial structures was obtained for the dosage planning by using Indexing HDMR method with high performance [20]. Liu, Haifeng, et al. proposed a model by using data mining techniques. The model was used to help physicians to control the glucose level in diabetes type 2 only. In the first, all factors affect the treatment plan were identified. The model performance was validated by using HPA1c value [21]. Habibi et al. proposed a model to diagnosis diabetes type 2 by using decision tree (J48). To generate the model, 22,398 medical records were used. The precision of the model was 0.717. The age factor was found as very important factor in the classification tree. The ROC curve was indicated that the model has high quality [22].Toussi, Massoud, et al. analyzed the type 2 management for French national guidelines. By using C5.0 algorithm, the physicians' prescriptions were obtained. About 463 medical records were used to develop the model. The model was consisted of 72 rules and 12 of them are related to the treatment types. As result, the model was useful a tool for the physicians in taking their decision in treatment plans [23]. Ramezankhani, Azra, et al. developed a prediction model to identify low factors for the type 2 diabetes incidence. About 6647 records and decision tree method were used to generate the model. The result mentioned that the accuracy was 90.5% and 97.9% specificity [24]. ALjumah et al proposed a classification model by using support vector machine algorithm in the Oracle Data Miner (ODM). The model was aimed to treating diabetes. From World Health Organization (WHO), the datasets were collected to generate the model. All datasets were related to Saudi diabetic patients. The model generated several treatment types for two age groups (Oldand young) [25]. Patil et al. developed a hybrid classification model for identifying type 2 diabeticpatients.

Diabetes is becoming a pandemic in world and with 62 million diabetic patients; India is one ofthe significant contributors [26]. Over the past 30 years, the prevalence of diabetes has increasedto12-18% in urban India and 3-6% in rural India. This rate of increase is 50-80% higher thanChina (10%) [27].According to International Diabetes Federation (IDF), India is the home of most number of diabetic patients and hence it is rightly termed as the “diabetes capital of the world” [29]. The estimated burden for properly treating diabetes is USD 2.2 billion in India,while government was spending only USD 61 per capita on healthcare in year 2012[30, 31].The following table shows all major studies done on prevalence of diabetes in India. Most of the studies are regional and none of them are done in Eastern U.P and Bihar. The literature on the studies underlines the fact, that there is a rising trend in the prevalence of Type 2 diabetes inUrban India [28].

3. DATASET DESCRIPTION

This dataset contains reports of diabetes-related symptoms of 520 persons [37]. It includes data about peoples including symptoms that may cause diabetes. This dataset has been created from a direct questionnaire to people who have recently become diabetic, or who are still no diabetic but having few or more symptoms. The data has been collected from the patients using direct questionnaire from Sylhet Diabetes Hospitalof Sylhet,Bangladesh.The data pre-processing has been conducted by handling the missing values following the technique of ignoring the tuples with incomplete values.

Table. 1 Dataset Description

	Number of attributes	Number of instances
Diabetes symptom dataset	16	520

Table. 2 Attribute Descriptions

Attributes	Description
Age	1.20–35, 2.36–45, 3.46–55,4.56–65, 6.above 65
Sex	1.Male, 2.Female
Polyuria	1.Yes, 2.No.
Polydipsia	1.Yes, 2.No.
Sudden Weight loss	1.Yes, 2.No.
Weakness	1.Yes, 2.No.
Polyphagia	1.Yes, 2.No.
Genital thrush	1.Yes, 2.No.
Visual blurring	1.Yes, 2.No.
Itching	1.Yes, 2.No.
Irritability	1.Yes, 2.No.
Delayed Healing	1.Yes, 2.No.
Partial Paresis	1.Yes, 2.No.
Muscle Stiffness	1.Yes, 2.No.
Alopecia	1.Yes, 2.No.
Obesity	1.Yes, 2.No.
Class	1.Positive, 2.Negative

After pre-processing, 520 instances have been remained in total. Among them, 320 are positive values and 200 are negative values. The detail description of the dataset and the attributes are shown in Tables 1 and 2. Two class variables are used to find whether the patient is having a risk of diabetes (positive) or not (negative).

4. METHODOLOGY

The aim of the research work is to analyze the diabetes dataset over the classification techniques. Our research concentrates, to reduce the complications of diabetes through early predictions and to improve the prognosis (lives) of the people. A person with diabetes has considerable features for the cause of disease depending on the age, glucose level, heredity, and other factors, as well these features vary from one type to another type.

The dataset containing the information about the symptoms of the patients will be fed to the prediction algorithms Random Forest (RF), Multilayer Perceptron (MLP) and Radial Basis Function

Network (RBF/RBFN) algorithm (Fig. 1). Then the performance of the algorithms will be tested with appropriate evaluation model, in particular, tenfold cross-validation techniques. Then the best algorithm chooses to build the system for the end users using the dataset as database. Taking the symptom from the user as input, the system will support the user for risk prediction. The dataset was analyzed using the following classification algorithms.

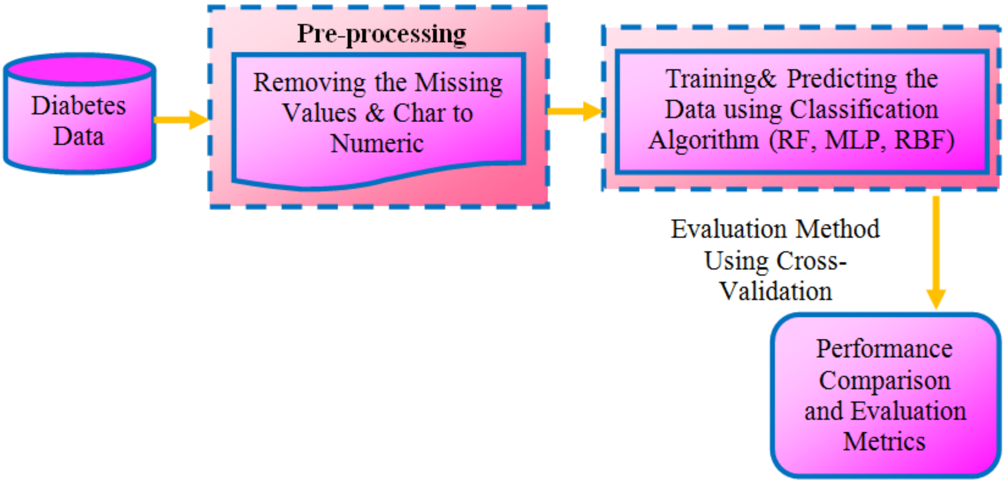
Diabetes Data
Pre-processing
Removing the Missing Values & Char to Numeric
Training& Predicting the Data using Classification Algorithm (RF, MLP, RBF)
Evaluation Method Using Cross-Validation
Performance Comparison and Evaluation Metrics

Fig. 1 Block Diagram for the overall research work

4.1Random Forest (RF)

Random Forest is a supervised learning, used for both classification and Regression. The logic behind the random forest [31, 32] is bagging technique to create random sample features. The difference between the decision tree and the random forest is the process of finding the root node and splitting the feature node will run randomly. The Steps are given below

- Load the data where it consists number of features representing the behaviour of the dataset.
- The training algorithm of random forest is called bootstrap algorithm or bagging technique to select n feature randomly from m features, i.e. to create random samples, this model trains the new sample to out of bag sample(1/3rd of the data) used to determine the unbiased OOB error.
- Calculate the node d using the best split. Split the node into sub-nodes.
- Repeat the steps, to find n number of trees.
- Calculate the total number of votes of each tree for the predicting target. The highest voted class is the final prediction of the random forest.

4.2 Multilayer Perceptron (MLP)

An artificial neural network (ANN) has three layers: input layer, hidden layer and output layer. The hidden layer vastly increases the learning power of the MLP. The transfer or activation function of the network modifies the input to give a desired output. Multilayer Perceptron (MLP) [33] network models are the popular network architectures used in most of the research applications in medicine, engineering, mathematical modelling, etc.In MLP, the weighted sum of the inputs and bias term are passed to activation level through a transfer function to produce the output, and the units are arranged in a layered feed-forward topology called Feed Forward Neural Network (FFNN). The schematic representation of FFNN with n inputs, m hidden units and one output unit along with the bias term of the input unit and hidden unit. The transfer or activation function of the network modifies the input to give a desired output. This step is continued iteratively until there is a convergence of weights.

- The input layer receives the signal and sends it to the hidden layers which subsequently propagate it to the nodes in output layer [34].
- The multilayer perceptron operates in two phases.
  - The first phase is the forward pass where the signal is propagated from input to output layer and the predicted error is computed.
  - The second phase is the backward pass where updating of weights occur corresponding to various nodes in the network so as to reduce the error.
- Adjust the error rate still the error rate is significantly reduced.

4.3 Radial Basis Function Network (RBF/RBFN)

The radial basis function network offers a viable alternative to the two-layer neural network in many applications of signal processing. A common learning algorithm for radial basis function networks is based on first choosing randomly some data points as radial basis function and then using singular value decomposition to solve for the weights of the network Radial basis networks can require more neurons than standard feedforward backpropagation networks, but often they can be designed in a fraction of the time it takes to train standard feedforward networks. The RBF network has a feed forward structure consisting of a single hidden layer of J locally tuned units, which are fully interconnected to an output layer of L linear units [33]. All hidden units simultaneously receive the n-dimensional real valued input vector X. Both layers have biases.RBF networks are also good at modelling nonlinear data and can be trained in one stage rather than using an iterative process as in MLP and also learn the given application quickly. They are useful in solving problems where the input data are corrupted with additive noise. The transformation functions used are based on a Gaussian distribution. The main difference from that of MLP is the absence of hidden-layer weights. The following steps are repeated until the network's mean squared error falls below goal.

- The network is simulated.
- The input vector with the greatest error is found. This layer contains D number of neurons, where D is input pattern dimension [35]. This layer is fully connected with pattern layer.
- A radial basis neuron is added with weights equal to that vector. This pattern layer is fully connected with summation layer. Every neuron of pattern layer is mathematically described by a gaussian radial basis function weights are redesigned to minimize error.
- The summation layer has small number of neurons, with linear activation functions. Size of this layer is limited to number of distinct classes in the training data. Output of the number of neuron in summation layer.
- Decision layer contains only one neuron. This outputs the class label of testing (unseen) pattern

5. EXPERIMENTAL AND RESULT ANALYSIS

The diabetes medical dataset (Early Stage diabetes dataset) has been collected from University of California, Irvine (UCI) machine learning repository [36]. This dataset contains medical report of 520 persons. This medical report (dataset) includes 16 features of the persons and the results such as whether the person has diabetes (positive) or not (negative).This dataset has been created from a direct questionnaire to people who have recently become diabetic, or who are still non-diabetic but having few or more symptoms. The data has been collected from the patients using direct questionnaire from Sylhet Diabetes Hospital of Sylhet, Bangladesh. The data pre-processing has been conducted by handling the missing values following the technique of replacing the tuples with incomplete values.

Table 3 Comparison of evaluation metrics using tenfold cross-validation

	RF	MLP	RBF
Accuracy	92.31	95.19	98.8
Sensitivity	100	95.31	100
Specificity	82.61	95	98.72
F_Measure	93.55	96.06	90.57
ROC	91.3	95.16	99.36

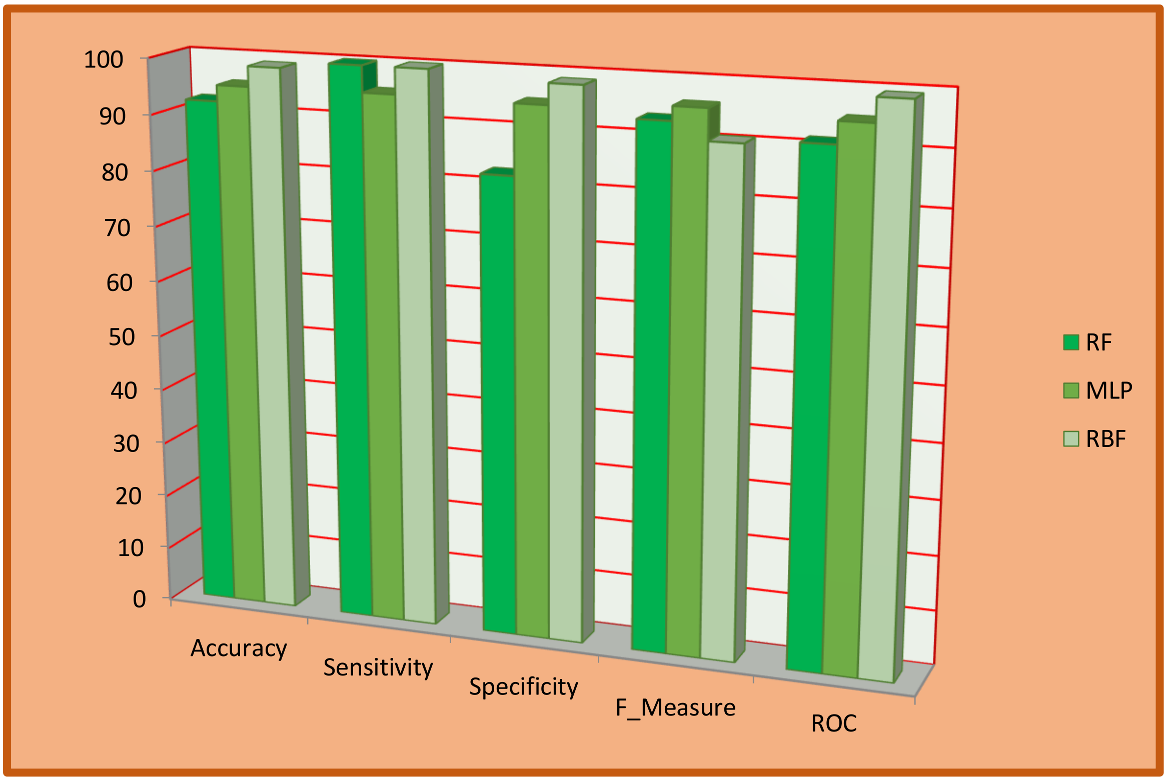

Fig. 2 Performance of classification algorithms using cross-validation technique

Performance of different Machine Learning techniques on our dataset with exhaustive accuracy information is represented in the following tables. While Radial Basis Function Network classifier is one of the most popular algorithms for data prediction, in case of our dataset; the accuracy of it was the lowest for the cross-validation method. However, the best result was achieved using Radial Basis Function Network Algorithm where using tenfold cross-validation 97.4% instances were classified correctly as shown in Table 3. For the more semantic view of the performance of used algorithms using evaluation techniques are illustrate in graphs. In Fig. 2, the performance of the algorithms using cross-validation evaluation is illustrated shown to represent the comparative accuracy of the used algorithms.

6. CONCLUSION

Diabetes is a heterogeneous group of diseases and it characterized by chronic elevation of glucose in the blood. This paper presented a diabetes prediction system for diabetes diagnosis. In order to develop this system, the dataset is collected from the University of California, Irvine (UCI) repository. Different machine learning algorithm namely decision tree-based random forests (RF) , function-based multilayer perceptron (MLP) and radial basis function network (RBF), are used to build the machine learning model to carry out the diagnosis of diabetes. Furthermore, the machine learning model is tested with evaluation methods using 10-fold cross validation (FCV) to evaluate the performance of the machine learning model in terms of accuracy. Our comparative analysis work also performs the analysis of the features in the dataset and found that the Radial basis function network (RBF) algorithm had performed with the best accuracy in cross validation evaluation test and the RBF is the best algorithm for the prediction of newly created datasets made for diabetic risk prediction. It gives the best fit to the data with respect to the diabetic and non-diabetic patients. Finally, in the future work the feature selection algorithm and other classification algorithms can be used to improve the accuracy and prediction based system or can create a better tool.

7. REFERENCE

1. The 6 Different Types of Diabetes: (5 Mar 2018). The diabetic journey.https://thediabeticjourney.com/the-6-different-types-of-diabetes
2. Statistics about Diabetes: American Diabetes Association, 22 Mar 2018. https://www.diabetes.org
3. Diabetes, World Health Organization (WHO): 30 Oct 2018. https://www.who.int/news-room/fact-sheets/detail/diabetes
4. Failure to detect type 2 diabetes early costing $700 million per year, Diabetes Australia, 8 July2018. https://www.diabetesaustralia.com.au
5. Harris, M.I., et al.: Onset of NIDDM occurs at least 4–7 yr before clinical diagnosis. Diabetes Care **15**(7), 815–819 (1992)
6. Akter, S., et al.: Prevalence of diabetes and prediabetes and their risk factors among Bangladesh iadults: a nationwide survey. Bull. World Health Organ. **92**, 204–213A (2014)
7. IDF Diabetes Atlas. International Diabetes Federation. 6th ed. 2013. [Accessed on January 6, 2014]. Available from: www. idf.org/diabetesatlas .
8. Alberti KGMM, Zimmet P, Shaw J. International Diabetes Federation: a consensus on type 2 diabetes prevention. Diabet Med. 2007; 24:451–63.
9. Ramachandran A, Snehalatha A, Samith Shetty A, Nanditha A. Primary prevention of type 2 diabetes in South Asians-challenges and the way forward. Diabet Med. 2013; 30:26–34.
10. Abdul-Ghani MA, DeFronzo RA. Pathophysiology of prediabetes. CurrDiab Rep. 2009; 9:193–9.
11. Ramachandran A, Ma RC, Snehalatha C. Diabetes in Asia. Lancet. 2010; 375:408–18.
12. American Diabetes Association Diagnosis and classification of diabetes mellitus. Diabetes Care. 2009; 32(Suppl. 1):S62–S67.
13. Cox E.M., Elelman D. Test for screening and diagnosis of type 2 diabetes. Clin Diabetes. 2009; 4(27):132–138.
14. Krentz A.J., Bailey C.J. Oral antidiabetic agents: current role in type 2 diabetes mellitus. Drugs. 2005; 65(3):385–411.
15. Tsave O., Halevas E., Yavropoulou M.P., Kosmidis Papadimitriou A., Yovos J.G., Hatzidimitriou A. Structure-specific adipogenic capacity of novel, well-defined ternary Zn (II)-Schiff base materials. Biomolecular correlations in zinc-induced differentiation of 3T3-L1 pre-adipocytes to adipocytes. J InorgBiochem. Nov 2015; 152:123–137. [Epub 2015 Aug 11]
16. Halevas E., Tsave O., Yavropoulou M.P., Hatzidimitriou A., Yovos J.G., Psycharis V. Design, synthesis and characterization of novel binary V(V)-Schiff base materials linked with insulin-

mimetic vanadium-induced differentiation of 3T3-L1 fibroblasts to adipocytes. Structure–function correlations at the molecular level. J InorgBiochem. Jun 2015; 147:99–115. [Epub 2015 Mar 26]

17. Tsave O., Yavropoulou M.P., Kafantari M., Gabriel C., Yovos J.G., Salifoglou A. The adipogenic potential of Cr (III). A molecular approach exemplifying metal-induced enhancement of insulin mimesis in diabetes mellitus II. J Inorg Biochem. Oct 2016; 163:323–331.
18. Sakurai H., Kojima Y., Yoshikawa Y., Kawabe K., Yasui H. Antidiabetic vanadium(IV) and zinc(II) complexes review article coordination. Chem Rev. March 2002; 226(1–2):187–198.
19. DeeptiSisodia, Dilip Singh Sisodia National Institute of Technology Prediction of Diabetes using Classification Algorithms, 2018.
20. Karahoca, Adem, and M. Alper Tunga."Dosage planning for type 2 diabetes mellitus patients using Indexing HDMR." Expert Systems with Applications39.8 (2012): 7207-7215.
21. Liu, Haifeng, et al. "An efficacy driven approach for medication recommendation intype 2 diabetes treatment using data mining techniques. "Studies in health technology andinformatics 192 (2012): 1071-1071.
22. Habibi, Shafi, Maryam Ahmadi, and SomayehAlizadeh. "Type 2 Diabetes Mellitus Screening and Risk Factors Using Decision Tree: Results of Data Mining." Global journal of health science 7.5 2015): 304.
23. Toussi, Massoud, et al. "Using data mining techniques to explore physicians' therapeutic decisions when clinical guidelines do not provide recommendations: methods and example for type 2 diabetes." BMC medical informatics and decision making 9.1 (2009):1.
24. Ramezankhani, Azra, et al. "Applying decision tree for identification of a low risk population for type 2 diabetes. Tehran Lipid and Glucose Study." Diabetes research and clinical practice 105.3 (2014): 391-398.
25. Abdullah A. Aljumah, Mohammed GulamAhamad, Mohammad KhubebSiddiqui, Application of data mining: Diabetes healthcare in young and old patients, Journal of King Saud University - Computer and Information Sciences, Volume 25, Issue 2,July 2013, Pages 127-136
26. Seema Abhijeet Kaveeshwa, Jon Cornwall, The current state of diabetes mellitus in India, Australia Med J. 2014; 7(1): 45–48.
27. Mohan V, Sandeep S, Deepa R, Shah B, Varghese C. Epidemiology of type 2 diabetes: Indian scenario. Indian J Med Res. Mar 2007;125(3):217-230
28. A. Ramchandran, Socio-Economic Burden of Diabetes in India, JULY 2007 VOL. 55
29. Guariguata L, Whiting DR, Hambleton I, Beagley J, Linnenkamp U, Shaw JE: Global estimates of diabetes prevalence for 2013 and projections for 2035 for the IDF Diabetes Atlas. Diabetes Res Clin Pract 2013, 49.
30. Ramchandran A: Socio-economic burden of diabetes in India Assoc Physicians India 2007,55(L):9
31. Lee JW, Lee JB, Park M, Song SH. An extensive evaluation of recent classification tools applied to microarray data. Comput Stat Data Anal. 2005; 48:869–85.
32. Yeung KY, Bumgarner RE, Raftery AE. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. Bioinformatics. 2005; 21:2394–402.
33. Venkatesan, Perumal & Anitha, S. (2006). Application of a radial basis function neural network for diagnosis of diabetes mellitus. Current Science. 91.
34. Mishra, Sushruta & Tripathy, H. & Mishra, Brojo. (2018). Implementation of biologically motivated optimisation approach for tumour categorisation. International Journal of Computer Aided Engineering and Technology. 10. 244.

35. Cheruku, Ramalingaswamy & Edla, Damodar & Kuppili, Venkatanareshbabu. (2017). Diabetes Classification using Radial Basis Function Network by Combining Cluster Validity Index and BAT Optimization with Novel Fitness Function. International Journal of Computational Intelligence Systems.
36. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
37. Islam, MM Faniqul, et al. 'Likelihood prediction of diabetes at early stage using data mining techniques.' Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, Singapore, 2020. 113-125.