2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)
Amity University, Noida, India. June 4-5, 2020

Early Prediction of Diabetes Mellitus Using Machine Learning

Gaurav Tripathi
Department of Computer Science and Engineering
Madan Mohan Malaviya University of Technology
Gorakhpur, India
01gaurav92@gmail.com

Rakesh Kumar
Department of Computer Science and Engineering
Madan Mohan Malaviya University of Technology
Gorakhpur, India
rkiitr@gmail.com

Abstract— **Diabetes mellitus is one of the noxious disease which causes abnormalities of blood glucose due to the resistance of producing insulin hormone in the body. It affects various organs in the body such as the kidney, nerves, and eyes if it is not an early diagnosis. With the advancement in technological growth, people attract to personalized healthcare. Machine learning is a very growing field in the predictive analysis and often used in healthcare applications where the prediction of diseases and their symptoms is identified in an early stage. The main objective of this work is to build a model for early prediction of diabetes by using machine learning classification algorithms under consideration of significant features related to diabetes. The proposed model gives the closest results comparing to clinical outcomes and also helps in the personalized diagnosis of patients. There are four machine learning algorithms these are Linear Discriminant Analysis (LDA), K-nearest neighbor (KNN), Support Vector Machine (SVM), and Random Forest (RF) are used in the predictive analysis of early-stage diabetes. Pima Indian Diabetes Database (PIDD) is used for experimental analysis which is taken from the UCI machine learning repository from the University of California, Irvine. The performance measures of these classification algorithms are done on various statistical measures such as sensitivity (recall), precision, specificity, F-score, and accuracy. Accuracy is the measurements of classifying correctly and incorrectly instances. The experimental results show that Random Forest (RF) gives the maximum accuracy of 87.66 % and outperformed in other classification algorithms.**

Keywords— Diabetes Mellitus; Blood Glucose; Machine Learning; SVM; KNN; LDA; RF

I. INTRODUCTION

Nowadays, people lifestyle is too much busy and most of them do not take care of their health and how to save it. It may cause us to generate many lifestyle diseases; diabetes mellitus, as usual, we can say diabetes is one of the diseases which closely related to our lifestyle. It can be the deadliest disease if it is unidentified [2] [8]. Our body needs energy for working and the main source of energy is blood glucose sourced from the food that you eat. In our body pancreas is an important organ that releases insulin. Insulin is a hormone that plays a vital role in regulating the sugar (glucose) level in the body. Glucose sourced from carbohydrates in the food that human being takes in their diet and responsible for the proper functioning of the body. Insulin maintains the sugar level in the blood such as to prevent the case of too low or too high, both are risky. Diabetes mellitus is a noxious disease that causes the organ pancreas residing in the body unable to produces enough amounts of insulin or no insulin, or due to the deficient action of the organism body does not able to respond to insulin. Due to this the abnormalities of blood glucose in the body that causes many complications in the body such as eye disease, stroke, kidney disease, hypertension, and dyslipidemia. In 2017 the data of United States, 30.3 million people have diabetes in which 23.1 million are analyzed and 7.2 million are not analyzed given by Central Disease Control and Prevention (CDC). In India also 30 million people suffering from this disease and it is expected this quantity is about to 80 million in 2030 [3] [4]. As per the International Diabetes Federation (IDF) more than 5 million people died to this disease in 2015 and the present data throughout the world is about 415 million people suffering from this disease and in India is about 50 million [8]. Diabetes has mainly three types. Type 1diabetes in which the immune system of the body destroys beta cells residing in the pancreas that produces the insulin hormone. In this, the formation of insulin is stopped in the body due to this higher abnormality of glucose level. The exact reason, why it is happening does not know. Some scientists think it is related to genes and mostly found in child age and youngers. The only solution is to give patients insulin injections with healthy food. In this, proper health check-up is very necessary [5]. In type 2 diabetes body produces less amount of insulin or it resists the production of insulin. Most of the people suffer from type 2 diabetes throughout the world. In this type of diabetes, the pancreas needs hard work in the production of insulin for the same amount of glucose in the body. The third one is gestational diabetes, it occurs during the pregnancy in the women. In the pregnancy period, the placenta resists the absorption of insulin in the body's cell so blood sugar level is high. Generally, this type of diabetes does not see after pregnancy. It can easily disappear by normal treatment and just changing the lifestyle. Nonetheless baby has a risk of type 2 diabetes [3]. Nowadays, blood samples taken from the body and send to a laboratory for analysis. There are three tests for checking the glucose level in the blood. A1C test detects the glucose level in blood which is done at a minimum of three months. If it exists between 5.7 to 6.4 % is a symptom of prediabetes and you are at the risk of diabetes. If this measures above 6.4 % then diabetes is diagnosed. In this, no fasting is required because this is a more convenient test. Fasting Plasma Glucose (FPG), check the glucose at fasting (without food). Oral Glucose Tolerance Test (OGTT) measures the blood glucose in the body before and two hours after by taking a drink which contains a measurable amount of glucose [3] [6]. we should aware of the proper treatment of disease to prevent further body damage [3] [4].

978-1-7281-7016-9/20/$31.00 ©2020 IEEE

Considering hazard from various diseases healthcare industry bring about a huge amount of valuable data such as electronic medical records, information regarding diseases, data for medication and interpretation whatever that helps in predictive analysis and determination for reducing risk management. The revolution of the intelligence analysis method in the medical field gives an unprecedented platform. In particular data mining and machine, learning field has great strength to manage a large amount of data from various sources for predictive analysis and knowledge extraction. Researchers have proved that various classifiers name as J48, SVM, KNN, Decision tree, and Random Forest (RF), etc. used in machine learning is beneficial to build a model in the area where the predictive analysis is challenging task, so often used in medical fields [2] [4]. According to various medical reports, diabetes is one of the noxious diseases and the early diagnosis of this disease is a critical problem. This work focuses on building a model using classification algorithms; these are LDA, SVM, KNN, and RF that help in the early-stage prediction of diabetes. To achieve the maximum accuracy of our model the class imbalance problems and missing values data are also considered and experiments are performed using various statistical measures.

The best parts of this paper are organized as follows. Section II presents related work. Section III gives the methodology used in the building of the model. Section IV presents experimental results and analysis. Section V presents the conclusion and future work.

II. Related Works

Sisodia et al. [2] discuss predicting the diabetes disease using three classifiers name as such as Naïve Bayes (NB), Support vector machine (SVM), and Decision tree (DT). An experiment is performed on the Pima Indian Diabetes Database (PIDD). The performance metric is measured in the term of Precision, Accuracy, Recall, and F-Score. Results obtained show Naïve Bayes outperformed among the three algorithms with 76.30 % Accuracy.

Ambilwade et al. [8] discuss the role of the Fuzzy Inference System (FIS) and Multilayer perceptron (MLP) for predicting the risk of prediabetes and also Type-2 diabetes by measuring the glucose level in the blood in the several aspects. An experiment is performed on 385 patient's data using Mat lab platform considering various statistical measures.

Wang et al. [4] discuss the predictive analysis of diabetes mellitus considering the role of imbalanced data with missing values. In their experiment, they used Naïve Bayes for data normalization by compensating the missing values. For addressing class imbalance problem oversampling with the ADASYN algorithm is used. Finally, the Random Forest (RF) is used for prediction. An experiment is performed Pima Indian Diabetes Database set and performance is checked by using the combined approach of these classifiers than they individually work to improve the results.

Sarwar et al. [15] gives a comparative study for prediction of diabetes mellitus using various machine algorithms and also discuss various statistical measures during this study. According to this if the dataset is large in size and more balance then accuracy is improved. For predictive analysis, five classifiers used the name as logistic regression (LR), k-Nearest Neighbour (KNN), Support Vector Machine (SVM), random forest (RF), and Decision Tree. SVM and KNN give better results. An experiment is performed on Pima Indian Diabetes dataset. To check the effectiveness of the model divided the dataset in the ratio of 70 and 30% where 70% of data used as a training test and 30% for the test set.

Perveen et al. [11] discuss the role of metabolic syndrome causes in the development of diabetes. Metabolic syndrome (MetS) is a collection of conditions that may cause various types of diseases in which diabetes type 2 is one. Logistic Regression is used to filter the significant conditions into MetS that arise from diabetes type 2. A comprehensive study is performed in the predictive analysis of diabetes using Naïve Bayes, Decision Tree, and J48 classifiers. For addressing the data imbalance problem experiments are performed by using K-medoids downsampling with Naïve Bayes classifier and compared the results with existing up and down sampling method and no sampling method. The obtained mean AROC accuracy of Naïve Bayes is 79%.

III. Methodology Used

Classification techniques are widely used in pattern recognition or predictive analysis for classifying the data into different classes. Machine learning and artificial neural network technology are beneficial technology that can do so due to the strength of their various classification algorithms supported by these technologies. These technologies are very frequently used in the medical field where predictive analysis is a challenging task; the cause of this is more imbalances and missing values in the data set [4]. Human beings always learn from past experiences and machine always follows the instruction given by human being. So to make a model and train this model in a particular domain, develop a valuable amount of dataset, develop the set of algorithms and check the accuracy of the model using various statistical measurements over the correctly and incorrectly classified instances [2] [10]. In this study, we aim to develop a model in the healthcare application using machine learning for predictive analysis of diabetes using significant features that are closely related to this disease.

The procedures that are used in the building of a model contain several useful steps that are described one by one that explores the logic run behind this study.

A. Dataset

For the experimental analysis, a Pima Indian Diabetes Database (PIDD) is used taken from the University of California, Irvine (UCI) Repository that contains valuable features that are closely related to this disease [17]. This dataset comprises 768 records in which 268 positive predicted classes refer to diabetes patients and 500 negative predicted classes refer to non-diabetes is in the ratio of 34.9% and 65.1% of the whole dataset, respectively. It contains 8 significant features with one outcome class which is described in table I.

B. Data pre-processing

Machine learning algorithms are completely dependent on data because it is the most crucial aspects that make model training possible. Initially when the dataset is collected from different sources is in the crude format so it may chance of many divergences, which the model may be unable to handle. So pre-processing is needed to remove all the divergences and prepare a clean data set. This included addressing missing values, calculate new features, split data

in the train-test set, data encoding means converting non-numerical data into numerical data, normalizing data, etc. Another problem that occurs during the pre-processing phase is data imbalance it means there exist more examples of one class than the other [14].

1) Missing values: Missing values are those values such as the value for some attributes in the given sample is zero. To understand this let us take an attribute diastolic blood pressure containing zero value for a person is not possible [4]. There are two approaches to solve the missing values problem 1) deletion of record 2) Imputation method. The first method such as the data deletion method is applied when the dataset is large, in that case, you can delete records that contain missing value still after that sufficient amount of data is available for prediction. But here we are dealing with health data and used dataset in this study comprises 768 records which are in not considerable amounts and also all the features are closely related to each other. So, in this case, the deletion of records that contain missing values is not a good approach. The second approach is the imputation method, in which addressing the missing values most probably feature's class mean or group median is used. Also in the handling of missing values problem can use the mean of nearest neighbor and random value method [14]. In this study feature's class mean is used for handling of missing values.

TABLE I. DESCRIPTION OF DATASET AND THEIR CHARACTERISTICS

Attributes Name	Description	Mean ± S.D
Pregnancies	Number of times pregnant	3.8 ± 3.3
Glucose (mg/dl)	Glucose concentration level	120.8 ± 31.9
Blood Pressure	Diastolic Blood Pressure (mmHg)	69.1 ± 19.3
Skin Thickness	Fold Thickness of skin (mm)	20.5 ± 15.9
BMI	Body Mass Index in (Kg/m^2)	79.7 ± 115.2
Diabetes Pedigree Function	Diabetes Pedigree Function	31.9 ± 7.8
Age	Age (Years)	0.4 ± 0.3
Outcome	Class Value Positive ('1') and Negative Class Value ('0')	33.2 ± 11.7

2) Balance and unbalance dataset: In the classification area often data unbalanced problem occurs and it is the problem that inequality in positive and negative predicted class. If the number of positive samples is the same as the number of negative samples, then the dataset is said to be balanced otherwise it is unbalanced. The advantage of a balanced dataset is that the evaluation is easier to do since there is no bias. Suppose in the given dataset positive sample is in 5% so the accuracy of the classifier which predicts the entire negative would be 95%. No doubt accuracy is very high but misleading. So to solve the problem of unbalanced dataset two techniques are available random-sampling and over-sampling. Other performance metrics such as sensitivity (recall), specificity, precision, and F-score are also evaluated to deal with imbalanced data. Replicating the minority class without any information lost falls in the original training dataset falls in the category of over-sampling. But it is prone to overfitting. In the under-sampling method, we simply remove the majority class to balance the dataset, it might discard useful information [4] [10]. In this study, the over-sampling method is used for class imbalance problem.

3) Data normalization: It is a very crucial aspect during the pre-processing phase. If you have a dataset, it may be the possibility of features of different units and scales [14]. In the given dataset some features are in low range scale and some are in high range scale, so for easy comparative analysis between them drawing them on the same scales and units is called normalization. The most probable used techniques that are used to normalize data are z-score and min-max. In this work, the min-max technique is used.

C. Algorithms used in predictive analysis

For developing the model four classifications algorithms are used in this study that is described one by one.

1) Linear Discriminant Analysis (LDA): Linear Discriminant Analysis is widely used in supervised learning classifications problem. It is based on dimensionality reduction that transforms the features from a higher dimension to a lower dimension segregated by a hyperplane [10]. The main concepts using in this classifier is to obtain the mean function of each class and estimated on the vectors with the aim to increases the distance between two class and reduces the distance between groups within the class to find the right groups.

2) K- Nearest Neighbour (KNN): KNN is widely used in classification as well as regression problem both. In this, the class of new samples is defined based on distance or similarity measure [10]. Three popular approaches Euclidean, Manhattan, and Minkowski are available for distance or similarity measure. Anyone can be used for measuring the distance. The working steps follow in KNN given below.

- The first step in the algorithm is the training phase that loads data and class levels of the training sample.
- The second step is choosing the value of K. Parameter k suggests the count of neighbors that are included in the majority vote process. The value of K is used to define the class of unlabeled into the defined class obtained by measuring the distance function.
- For selecting the value of K we used a heuristic approach.

3) Support Vector Machine (SVM): SVM is a popular supervised computational algorithm used in both area regression as well as classification. SVM draws the data item in a higher dimension space. Suppose if you have 'n' features, it draws data items in n-dimensional space. SVM draws the hyperplane between dataset that best segregate the dataset into classes. The challenging task is the selection of optimal hyperplane in the dimensional space and the right hyperplane is that plane which is on the highest margin between two classes. The points which are closer to hyperplane are called the support vectors. The mapping of the objects is according to the specified boundaries of the hyperplane. The class of the new sample is based on hyperplane that belongs to either one of the class along the hyperplane [3].

4) Random Forest (RF): RF is a very strong supervised learning algorithm, which is used in both cases classification as well as regression. It is an ensemble classifier that consists of a lot of decision tree and the prediction is based on the majority of votes collected from these trees [10]. So it gives better results in comparison to individual decision tree classifiers. It uses the concept of bagging technique to train each tree by generating the random sample features in the given sample. For generating the decision tree the commonly used algorithms are ID3 and CART. Some useful steps that are used in RF are given below [4].

- Initially load the training data that consists of 'm' features, which shows the behavior of the dataset.
- Randomly sample a subset of training (with replacement) called bagging such that select 'n' features randomly from 'm' features.
- The 'n' training features are used in the modeling of 'n' decision tree.
- Gini index is used for the selection of splitting nodes (Best node) in the case of each decision tree.
- The above steps will go on for the modeling of 'n' number of the decision trees.
- The majority voted class is calculated in the count of collected votes of all trees in predicting the target class.
- Take the mode of all prediction in the case of classification and take the mean in the case of regression.

D. Evaluation technique

K-fold cross-validation used to check the effectiveness and measures performance of the model. In this, the original dataset is prepared into a train-test set to validate the performance. Here K refers to the number of sections in which the whole data item is divided. To obtain the statistical reliable results experiments are conducted by several iterations. Suppose if the value of K is 10 the experiments will be conducted in 10 iterations. Out of K iteration for every value of K one section is selected as a test set and the remaining K-1 sections are selected as a train set. The benefit of using this strategy each section gets an equal chance to become a test set. Take the mean of obtained results after K experiments which show the performance measures of the model [4]. The estimated mean error of the k tests is given by Equation 1.

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

Where E_i is the error obtained in each pass that occurs in the test dataset.

E. Statistical Evaluation

To measures the performance of various classifiers that are used to build a model, some important statistical metrics are calculated. The metrics are accuracy, sensitivity (recall), precision, specificity, and F-score. These metrics depend on classification labels such as true positive (TP), true negative (TN), false positive (FP), and false-negative (FN) [10].

1) Accuracy: Divide the summation of TP and TN against the whole population.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \quad (2)$$

2) Sensitivity / Recall: It measures the actual true positive rate and is calculated by using the formula given below.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (3)$$

3) Specificity: It is the measure of the true negative rate and is given by the following formula.

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (4)$$

4) Precision: It is defined by dividing the true positive against whole positive class values predicted. Mathematically it can be given by the following formula, given below.

$$\text{Precision (P)} = \frac{TP}{TP + FP} \quad (5)$$

5) F-score: It is defined by dividing the true positive against whole positive class values predicted. Mathematically it can be given by the following formula, given below.

$$F - \text{score} = 2 \times \frac{p \times r}{p + r} \quad (6)$$

IV. Experimental Results And Analysis

In this study there are four classification algorithms are used to prepare a model that helps in the early-stage prediction of Diabetes Mellitus based on significant features related to this disease. These algorithms are LDA, KNN, SVM, and RF and used dataset is Pima Indian Diabetes Database (PIDD) sourced from UCI Repository [17]. Separate the data set into the train-test set and K-fold cross-validation is used by taking the value of K=10. The missing value and class imbalance

problem are solved during the experiments to achieve the maximum accuracy of the model. A missing value is replaced by features class mean and for class imbalance problem over-sampling method is used. The formula for calculating accuracy refers to Equation 2. Also have been evaluated the other important performance indicators such as sensitivity (recall), precision, specificity, and F-score. Figure 1 shows the description of the data set. It consists of 768 samples and eight attributes with one class label where '0' for negative outcome shows non-diabetic patients and '1' for positive outcome shows diabetic patients. Figure 2 shows the total outcome of diabetic and non-diabetic patients in which 268 is a count of diabetic patients and 500 is a count of non-diabetic patients. The outcome of the results is given by the tabular as well as in graphical form. Table II gives the summarized results of all classifier comprises all the performance metrics that are necessary to measures the strength of all classifier. Figure 3 shows the obtained accuracy of all classifier and figure 4 shows F-score. Figure 5 presents the precision and sensitivity (recall) outcome and figure 6 presents the specificity. The obtained results show that Random Forest (RF) classifier gives a maximum accuracy of 87.66 and is appropriated for our model in the prediction of Mellitus diabetes.

```
In [6]: data = pd.read_csv("C:/daase/diabetes.csv")
   ...: data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
Pregnancies                 768 non-null int64
Glucose                     768 non-null int64
BloodPressure               768 non-null int64
SkinThickness               768 non-null int64
Insulin                     768 non-null int64
BMI                         768 non-null float64
DiabetesPedigreeFunction    768 non-null float64
Age                         768 non-null int64
Outcome                     768 non-null int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Fig. 1. Description of the diabetes data set

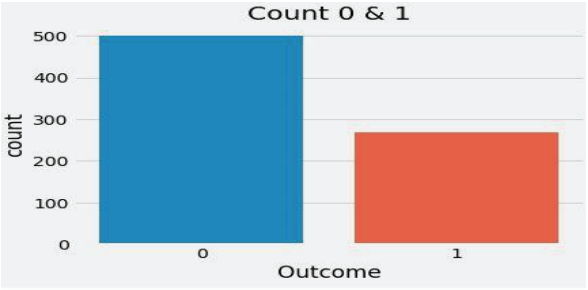

Fig. 2. Outcome of diabetic and non-diabetic patients

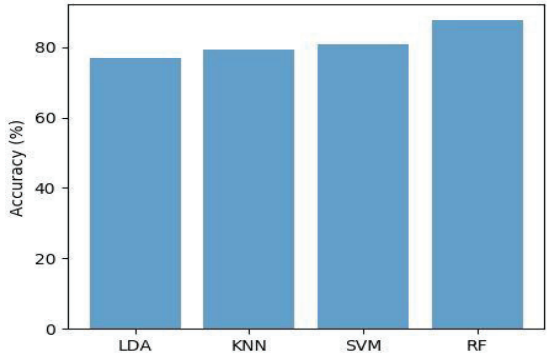

Fig. 3. Accuracy of classification algorithms

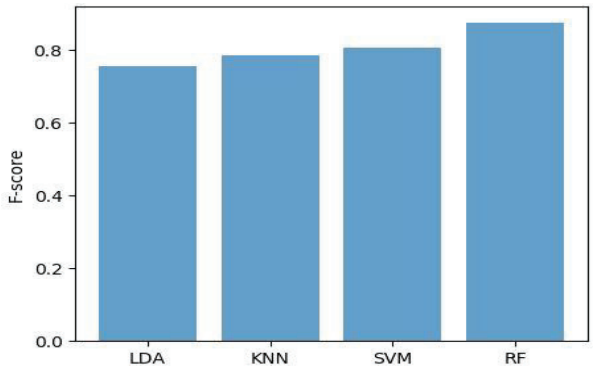

Fig. 4. F-Score measures of all classifier

TABLE II. PERFORMANCE ANALYSIS OF USED CLASSIFIERS ON VARIOUS MEASURES

Classifiers	Precision	Recall	Specificity	F-score	Accuracy
LDA	0.701	0.817	0.720	0.755	76.86
KNN	0.751	0.821	0.763	0.785	79.24
SVM	0.819	0.793	0.816	0.806	80.85
RF	0.876	0.880	0.872	0.875	87.66

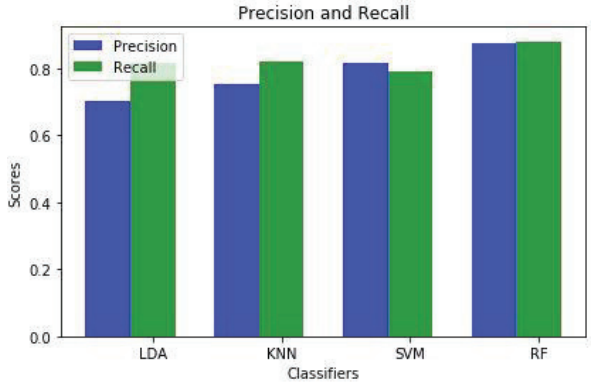

Fig. 5. Precision and Recall measures of all classifier

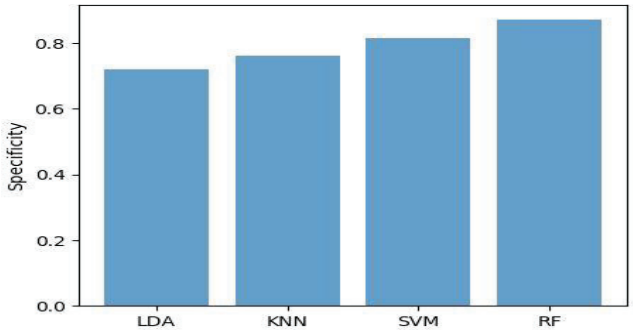

Fig. 6. Specificity measures of all classifier

V. CONCLUSION AND FUTURE WORK

Diabetes Mellitus is one of the real world noxious disease and the early diagnosis of this disease is always a challenging problem. This study uses machine learning classification algorithms in designing a model that greatly faces all the challenges and helpful in the early diagnosis of diabetes disease. PIDD dataset is used for performing the experiments by using four machine learning algorithms such as LDA, KNN, SVM, and RF. The data set comprises 768 records and 8 significant features related to diabetes with a

class label that shows the outcome of diabetic and non-diabetic patients. Our main aim is to achieve maximum accuracy of the model also the other important performance metrics have been evaluated such as precision, recall, specificity, and F-score. These performance metrics are evaluated according to confusion metrics such as true positive, true negative, false positive, false negative. The obtained results show that Random Forest (RF) gives the maximum accuracy of 87.66% and outperformed the other used classifiers. So the RF classifier is appropriated for our model.

In the future, we intend to extend our work in the prediction of other diseases like psoriasis, cancer, etc using machine learning and artificial intelligence technology.

REFERENCES

[1] R. M. Khalil and A. Al-Jumaily, "Machine learning based prediction of depression among type 2 diabetic patients," 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), Nanjing, pp. 1-5, 2017.

[2] Sisodia, D., Sisodia, D.S., "Prediction of Diabetes using Classification Algorithms," in: International Conference on Computational Intelligence and Data Science (ICCIDS 2018), ELSEVIER. Procedia Computer Science , ISSN 1877-0509,vol 132.

[3] Sneha, N., Gangil, T. ,"Analysis of diabetes mellitus for early prediction using optimal features selection," in: Journal of Big Data 6, 13 (2019).

[4] Q. Wang, W. Cao, J. Guo, J. Ren, Y. Cheng and D. N. Davis, "DMP_MI: An Effective Diabetes Mellitus Classification Algorithm on Imbalanced Data With Missing Values," in IEEE Access, vol. 7, pp. 102232-102238, 2019.

[5] J. N. Myhre, I. K. Launonen, S. Wei and F. Godtliebsen, "Controlling blood glucose levels in patients with type 1 diabetes using fitted q-iterations and functional features," 2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP), Aalborg, pp. 1-6, 2018.

[6] B. J. Lee and J. Y. Kim, "Identification of Type 2 Diabetes Risk Factors Using Phenotypes Consisting of Anthropometry and Triglycerides based on Machine Learning," in IEEE Journal of Biomedical and Health Informatics, vol. 20, no. 1, pp. 39-46, Jan. 2016.

[7] B. J. Lee, B. Ku, J. Nam, D. D. Pham and J. Y. Kim, "Prediction of Fasting Plasma Glucose Status Using Anthropometric Measures for Diagnosing Type 2 Diabetes," in IEEE Journal of Biomedical and Health Informatics, vol. 18, no. 2, pp. 555-561, March 2014.

[8] R. P. Ambilwade and R. R. Manza, "Prognosis of diabetes using fuzzy inference system and multilayer perceptron," 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I), Noida, pp. 248-252, 2016.

[9] E. M. Aiello, C. Toffanin, M. Messori, C. Cobelli and L. Magni, "Postprandial Glucose Regulation via KNN Meal Classification in Type 1 Diabetes," in IEEE Control Systems Letters, vol. 3, no. 2, pp. 230-235, April 2019.

[10] Maniruzzaman, M., Rahman, M.J., Al-MehediHasan, M., Suri, H.S., Abedin, M., El-Baz, A., Suri, J.S.," Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers," J Med Syst 42, 92(2018).

[11] S. Perveen, M. Shahbaz, K. Keshavjee and A. Guergachi, "Metabolic Syndrome and Development of Diabetes Mellitus: Predictive Modeling Based on Machine Learning Techniques," in IEEE Access, vol. 7, pp. 1365-1375, 2019.

[12] D. Sierra-Sosa, B. Garcia-Zapirain, C. Castillo, I. Oleagordia, R. Nuño-Solinis, M. Urtaran-Laresgoiti, A. Elmaghraby, "Scalable Healthcare Assessment for Diabetic Patients Using Deep Learning on Multiple GPUs," in IEEE Transactions on Industrial Informatics, vol. 15, no. 10, pp. 5682-5689, Oct. 2019.

[13] M. Goyal, N. D. Reeves, S. Rajbhandari and M. H. Yap, "Robust Methods for Real-Time Diabetic Foot Ulcer Detection and Localization on Mobile Devices," in IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 4, pp. 1730-1741, July 2019.

[14] Malley B., Ramazzotti D., Wu J.T. (2016) Data Pre-processing. In: Secondary Analysis of Electronic Health Records. Springer, Cham.

[15] M. A. Sarwar, N. Kamal, W. Hamid and M. A. Shah, 2018.," Prediction of Diabetes Using Machine Learning Algorithms in Healthcare," in: 2018 24th International Conference on Automation and Computing (ICAC), Newcastle upon Tyne, United Kingdom, pp. 1-6, 2018.

[16] Birjais, R., Mourya, A.K., Chauhan, R., Kaur, H., "Prediction and diagnosis of future diabetes risk: a machine learning approach," SN Appl. Sci. 1, 1112 (2019).

[17] K. Bache and M. Lichman, "UCI machine learning repository", 2013, University of California. URL http: //archive.ics.uci.edu/ml.