Radhakrishnan C, 2023, 11:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

Machine Learning-Based Prediction of Diabetes Using Medical Data Analysis and Classification Algorithms for Early Detection

Radhakrishnan C, Abishek B, Joandsouza A, Karnesh P
Department of Computer Science and Engineering
Kongunadu College of Engineering and Technology Tamil Nadu,India

Abstract- Diabetes mellitus is a harmful disease characterized by abnormal blood glucose levels resulting from insulin resistance. If not diagnosed early, it can lead to complications in organs such as the kidneys, nerves, and eyes. With the advent of technological advancements, people are moving toward personalized healthcare. Machine learning (ML), a rapidly evolving field in predictive analysis, is increasingly applied in healthcare to identify diseases and symptoms at early stages. This work aims to develop a machine learning model for the early prediction of diabetes using classification algorithms, considering significant features related to diabetes. The proposed model provides results comparable to clinical outcomes, assisting in personalized patient diagnoses. Four machine learning algorithms—Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest (RF)—are utilized for early diabetes prediction. The experimental analysis uses the Pima Indian Diabetes Database (PIDD) from the UCI Machine Learning Repository. The performance of these algorithms is evaluated using statistical measures such as sensitivity (recall), precision, specificity, F-score, and accuracy. Accuracy measures the correct and incorrect classification of instances. The experimental results demonstrate that the Random Forest (RF) algorithm achieves the highest accuracy of 87.66%, outperforming other algorithms.

Keywords- Diabetes, Machine Learning, Classification Algorithms, Early Detection, Random Forest, Support Vector Machine, K-Nearest Neighbors, Linear Discriminant Analysis.

I. INTRODUCTION

People's lives are too hectic these days, and the majority of them don't know how to protect their health. It could lead to the development of numerous lifestyle disorders, such as diabetes mellitus, as is customary, is one of the conditions that is strongly linked to our way of life. If the disease is unknown, it may be the most deadly [2] [8]. Our bodies require energy to function, and blood glucose, which comes from the food we eat, is the primary source of this energy. The pancreas is a vital organ in our body that secretes insulin. One hormone that is essential for controlling the body's sugar (glucose) content is insulin. derived from the carbohydrates found in meals consumed by humans, glucose is essential for the body's healthy operation. Insulin keeps blood sugar levels in check, preventing dangerously low or high blood sugar levels. Diabetes mellitus is an unpleasant illness that results in the body's pancreas producing either insufficient amounts of insulin or none at all, or the body's inability to react to insulin as a result of the

organism's poor function. This results in high blood glucose levels, which can lead to a number of health problems, including hypertension, renal disease, stroke, eye disease, and dyslipidemia. According to data from the Centers for Disease Control and Prevention (CDC), 30.3 million Americans had diabetes in 2017; 23.1 million of these cases were examined, whereas 7.2 million were not. This disease affects 30 million people in India as well, and by 2030, that number is predicted to rise to 80 million [3] [4].

According to the International Diabetes Federation (IDF), the disease claimed the lives of over 5 million people in 2015. Current statistics show that 415 million people worldwide suffer from diabetes, with over 50 million of those cases occurring in India. There are primarily three forms of diabetes. In type 1 diabetes, the body's immune system kills the beta cells that make the hormone insulin in the pancreas. In this case, the increased irregularity of the glucose level causes the body to stop producing insulin. Nobody is certain of the precise cause of it. According to some scientists, it primarily affects children and young people and is linked to DNA. Giving them insulin shots along with nutritious diet is the only way to solve the problem. A thorough medical examination is crucial in this regard [5]. The body either resists or generates less insulin when a person has type 2 diabetes. Type 2 diabetes affects the majority of individuals worldwide. 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 is gestational diabetes, which affects women throughout pregnancy. Blood sugar levels are high during pregnancy because the placenta prevents the body's cells from absorbing insulin. 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]. These days, bodily blood samples are extracted and sent to a lab for examination. To determine the blood glucose level, three tests are available. The A1C test measures blood glucose levels and is performed at least three months apart. You run the risk of developing diabetes if it is present between 5.7% and 6.4%, which is a sign of prediabetes. Diabetes is diagnosed if this value is higher than 6.4%. Since this is a more convenient test, there is no need to fast. Fasting Plasma Glucose (FPG): measure the glucose levels while fasting (without eating).

By consuming a beverage that contains a detectable amount of glucose, the Oral Glucose Tolerance Test (OGTT) determines the blood glucose levels in the body both before and two hours after [3].We ought to understand how to manage illnesses appropriately in order to stop additional bodily harm [3] [4]. Given the risk of numerous illnesses, the healthcare sector generates a vast amount of useful data, including electronic medical records, disease information, medication data, and interpretation of any information that aids in predictive analysis and decision-making for lowering risk management. The medical field now has an unparalleled platform because to the breakthrough in intelligence analytic techniques.

The fields of data mining and machine learning, in particular, are very strong at handling vast amounts of data from diverse sources for knowledge extraction and predictive analysis. Researchers have demonstrated that a variety of machine learning classifiers, such as J48, SVM, KNN, Decision tree, Random Forest (RF), etc., are useful for creating a model in domains where predictive analysis is a difficult work; for this reason, they are frequently employed in the medical domains [2] [4]. Numerous medical reports state that diabetes is one of the most harmful illnesses, and that it is crucial to diagnose it early. The goal of this work is to develop a model that aids in the early-stage prediction of diabetes using classification algorithms, specifically LDA, SVM, KNN, and RF. Class imbalance issues and missing values data are also taken into account, and tests are conducted using a variety of statistical measures in order to maximize the correctness of our model.

II. RELATED WORK

Sisodia et al. [2] talk about utilizing three classifiers— Naïve Bayes (NB), Support Vector Machine (SVM), and Decision Tree (DT)—to predict the occurrence of diabetes. Using the Pima Indian Diabetes Database (PIDD), an experiment is conducted. Precision, Accuracy, Recall, and F- Score are the metrics used to measure the performance. According to the results, Naïve Bayes fared better than the other two algorithms, with an accuracy of 76.30 percent.

Ambilwade et al. [8] talk about how the Multilayer Perceptron (MLP) and Fuzzy Inference System (FIS) measure blood glucose levels in a number of ways to predict the likelihood of Type 2 diabetes and prediabetes. Using the Mat lab platform, an experiment is conducted on the data of 385 patients while taking into account a number of statistical metrics.

The predictive analysis of diabetes mellitus taking into account the role of unbalanced data with missing values is covered by Wang et al. [4]. In their experiment, they compensated for the missing values using Naïve Bayes to normalize the data. The ADASYN method employs oversampling to address the issue of class imbalance. Lastly, prediction is done via Random Forest (RF). The Pima Indian Diabetes Database set is used in an experiment, and performance is evaluated by combining these classifiers' methods before each one works alone to enhance the outcomes.

Sarwar et al. [15] compare different machine algorithms for diabetes mellitus prediction and go over a number of statistical metrics in the process. This suggests that accuracy is increased if the dataset is larger and more balanced. Five classifiers—logistic regression (LR), k-Nearest Neighbor (KNN), Support Vector Machine (SVM), random forest (RF), and Decision Tree—were employed for predictive analysis. KNN and SVM produce superior outcomes. Using the Pima Indian Diabetes dataset, an experiment is conducted. The dataset was split 70/30 to assess the model's efficacy, with 70% of the data serving as a training test and 30% as 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 down sampling 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

In order to divide the data into distinct classes for pattern recognition or predictive analysis, classification techniques are frequently employed. The power of the several categorization methods that machine learning and artificial neural network technologies offer allows them to be useful technologies. Because of the increased imbalances and missing values in the data set, predictive analysis is a difficult task in the medical profession, where these technologies are widely used [4]. Machines constantly follow human instructions, and humans always learn from their prior experiences. Therefore, in order to create a model and train it in a certain domain, a significant amount of data must be gathered, a set of algorithms must be developed, and the quality of the model must be assessed using a variety of statistical measures over both properly and erroneously categorized examples [2] [10]. Our goal in this project is to use important variables that are strongly associated with diabetes to create a machine learning model for predictive analysis of the condition in a healthcare application.

The methods used to construct a model include a number of helpful phases that are explained one at a time, examining the reasoning behind this investigation.

Dataset

The University of California, Irvine (UCI) Repository's Pima Indian Diabetes Database (PIDD), which includes useful characteristics closely associated with this illness, is employed for the experimental study [17]. This dataset, which consists of 768 records, has 500 negative projected classes that relate to people without diabetes and 268 positive predicted classes that refer to individuals with diabetes. These two groups make up 34.9% and 65.1% of the entire dataset, respectively. It has one result class and eight important qualities, which are detailed in table I.

Data pre-processing

Since data is one of the most important components that enables model training, machine learning algorithms are entirely reliant on it. First, when the dataset is gathered from several sources in an unrefined fashion, which increases the likelihood of numerous discrepancies that the model might not be able to handle. Pre-processing is therefore required to eliminate all discrepancies and provide a clean data collection. Data encoding, which is the process of turning non-numerical data into numerical data, normalizing data, addressing missing values, calculating new features, and splitting data in the train-test set were all covered in this.

Data imbalance, which occurs when there are more samples of one class than the other, is another issue that arises during the pre-processing stage [14].

Missing Values: Missing values are those that have zero values for certain attributes in the sample.Let's look at an example to better grasp this: a person's diastolic blood pressure cannot be 0 [4]. The missing values issue can be resolved in two ways. 1) Record deletion 2) Imputation technique. When the dataset is huge, you can use the first technique, known as the data deletion approach, to remove records with missing values. This way, there will be enough data left over for predictions. However, as we are working with health data, the dataset used in this study consists of 768 records, which are not very many, and all of the attributes are intimately connected to one another. Therefore, it is not a good idea in this situation to remove records that have missing information. The second strategy is the imputation method, which uses the class mean or group median to fill in the missing numbers. The random value approach and mean of nearest neighbor can also be used to handle missing values [14]. Missing values in this study are handled using the class mean.

Table 1: 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/m2)	79.7 ± 115.2
Diabetes Pedigree Function	Diabetes Pedigree Function	31.9 ± 7.8
Outcome	Class Value Positive („1") and Negative Class Value („0")	33.2 ± 11.7

Balance and Unbalance Dataset: The issue of inequality in positive and negative predicted classes is a common occurrence in the classification field. The dataset is considered balanced if the proportion of positive samples equals that of

negative samples; if not, it is considered unbalanced. A balanced dataset has the benefit of making evaluation simpler because there is no bias. Assuming that the positive sample in the dataset is 5%, the classifier's accuracy in predicting the full negative would be 95%. Accuracy is undoubtedly quite high but deceptive. Random sampling and over-sampling are the two methods available to address the issue of an unbalanced dataset. To address unbalanced data, other performance indicators like sensitivity (recall), specificity, precision, and F-score are also assessed. Replicating the minority class in the original training dataset without losing any information is classified as excessive sampling. However, it has a tendency to over fit. The majority class is simply eliminated in the under- sampling technique to balance the dataset, it may leave out important details [4] [10]. The over-sampling technique is applied in this study to address the issue of class imbalance.

Data Normalization: It is an extremely important component of the pre-processing stage. There may be the potential for features of various scales and units provided you have a dataset [14]. Drawing the features on the same scales and units is known as normalizing since some of the characteristics in the dataset are in low range scales and some are in high range scales, making comparisons between them simple. The most often employed methods for data normalization are min-max and z-score. The min-max approach is applied in this paper.

Algorithms Used in Predictive Analysis

This study uses four classification algorithms, each of which is explained separately, to build the model.

Linear Discriminant Analysis (LDA): In supervised learning classification problems, linear discriminant analysis is frequently employed. Its foundation is dimensionality reduction, which separates features by a hyperplane and converts them from a higher dimension to a lower dimension [10]. The primary ideas behind this classifier are to locate the appropriate groups by reducing the distance between groups inside the class and increasing the distance between two classes by estimating the mean function of each class on the vectors.

K- Nearest Neighbour (KNN): KNN is frequently used for both regression and classification problems. In this case, the distance or similarity metric is used to determine the class of new samples [10]. Three well-liked methods for measuring distance or similarity are Minkowski, Manhattan, and Euclidean. The distance can be measured by anyone. The KNN working steps are listed below.

- The training phase, which loads the training sample's data and class levels, is the initial stage of the algorithm.
- Selecting the value of K is the second stage. The number of neighbors included in the majority voting process is suggested by parameter k. The class of unlabeled is determined by the distance function measurement, and the value of K is utilized to define this class.
- We applied a heuristic to choose the value of K.

Support Vector Machine (SVM): SVM is a well-liked supervised computational approach that is applied to classification and area regression. The data item is drawn in a higher dimension space via SVM. Assume that it draws data items in n-dimensional space if you have "n" characteristics. The hyperplane connecting datasets that optimally divides the dataset into classes is drawn by SVM. The hardest part is choosing the best hyperplane in the dimensional space; the right hyperplane is the one with the largest difference between two classes. The support vectors are the spots that are nearer the hyperplane. The items are mapped in accordance with the hyperplane's designated boundaries. The new sample's class is determined by a hyperplane that falls into one of the classes.

Random Forest (RF): RF is a very powerful supervised learning algorithm, which is applied to both regression and classification. This classifier is an ensemble one that consists of numerous

decision trees, with the majority of votes gathered from these trees serving as the basis for the prediction [10]. Therefore, compared to individual decision tree classifiers, it produces better results. By creating random sample features in the provided sample, it trains each tree using the bagging technique concept. The ID3 and CART algorithms are frequently used to generate the decision tree. Below are some helpful steps that are employed in RF [4].

- Load the training data first, which includes "m" features that illustrate the dataset's behavior.
- Select "n" features at random from "m" features via bagging, a technique that involves sampling a portion of training data (with replacement).
- The "n" decision tree is modeled using the "n" training characteristics.
- For every decision tree, the splitting nodes (best node) are chosen using the Gini index.
- The aforementioned procedures will be followed in order to model "n" decision trees.
- The total number of votes received by all trees in predicting the target class is used to determine the majority voted class.
- In the classification situation, take the mean of all the predictions; in the regression case, take the mean.

Evaluation Technique

K-fold cross-validation is used to assess the model's performance and efficacy. To verify the performance, the original dataset is converted into a train-test set. In this case, K is the number of parts that make up the entire data item.

Experiments are carried out over multiple iterations in order to get statistically sound results. Assume that the trials will be carried out in ten iterations if K is 10. One portion is chosen as a test set for each value of K in the iteration, while the remaining K-1 sections are chosen as a train set. The advantage of this approach is that every section has an equal chance of becoming a test set. Calculate the average of the outcomes following K trials that display the model's performance metrics [4].

Statistical Evaluation

Several significant statistical metrics are computed to assess the effectiveness of different classifiers that are utilized to construct a model. Accuracy, sensitivity (recall), precision, specificity, and F-score are the measurements.

True positive (TP), true negative (TN), false positive (FP), and false-negative (FN) are among the classification categories that these measures rely on [10].

- **Accuracy:** Accuracy measures the overall correctness of the model by calculating the proportion of correctly predicted instances (both positive and negative) out of all the instances.
- **Sensitivity:** Sensitivity, also known as recall, measures the proportion of actual positive cases correctly identified by the model. It tells us how well the model detects positive instances (e.g., diabetic patients in the case of diabetes prediction).
- **Specificity:** Specificity measures the proportion of actual negative cases correctly identified by the model. It tells us how well the model avoids false positives.
- **Precision:** Precision measures the proportion of positive predictions that are actually correct. It focuses on the accuracy of the positive class predictions and is important when false positives are critical.
- **F-score:** The F-score, or F1-score, is the harmonic mean of precision and recall. It provides a balance between the two, especially when there is a need to consider both false positives and false negatives equally.

IV. RESULT & DISCUSSION

Four classification algorithms are employed in this work to create a model that aids in the early detection of diabetes mellitus based on important characteristics associated with the condition. The Pima Indian Diabetes Database (PIDD), which was obtained from the UCI Repository, is the dataset used in these algorithms, which include LDA, KNN,

SVM, and RF [17]. Divide the data set into train and test sets, then apply K-fold cross-validation with K=10 as the value. During the experiments, the class imbalance and missing value issues are resolved to maximize the model's accuracy. The features class mean is utilized to fill in the missing value, and the over-sampling technique is applied to address the issue of class imbalance. Equation 2 is the formula used to calculate accuracy. Other significant performance metrics, including F-score, accuracy, specificity, and sensitivity (recall), have also been assessed. The data set's description is displayed in Figure 1. It has eight qualities and 768 samples with a single class label; patients with diabetes are indicated by a "1" for a good outcome and those without diabetes by a "0" for a negative outcome. The overall results of patients with and without diabetes are displayed in Figure 2, where There are 500 patients without diabetes and 268 persons with diabetes. The outcomes' outcome is provided by the both in graphical and tabular form. The results of all classifiers are given in Table II, which also includes all performance indicators required to gauge each classifier's strength. Figure 4 displays the F-score, whereas Figure 3 displays the overall classifier's acquired accuracy. The results for precision and sensitivity (recall) are shown in Figure 5, and the specificity is shown in Figure 6. According to the results, the Random Forest (RF) classifier is suitable for our model in the prediction of Mellitus diabetes and has a maximum accuracy of 87.66.

```
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

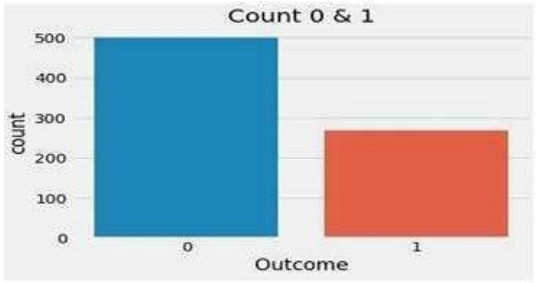

Fig. 2. Outcome of diabetic and non-diabetic patients

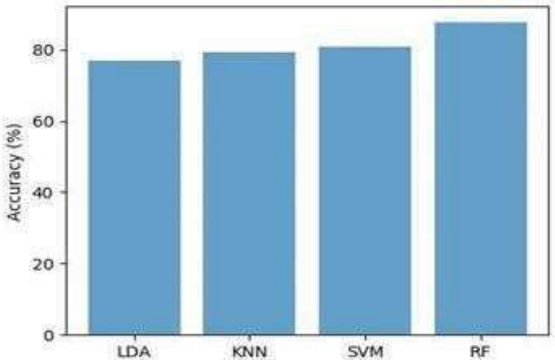

Fig. 3. Accuracy of classification algorithms

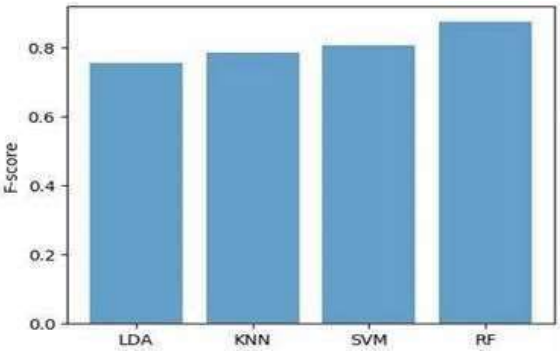

Fig. 4. F-Score measures of all classifier

Table 2. 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

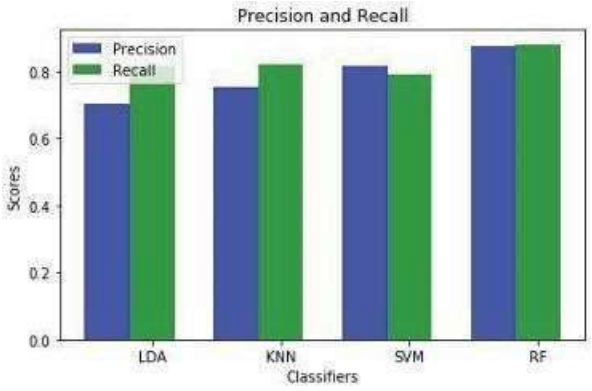

Fig. 5. Precision and Recall measures of all classifier

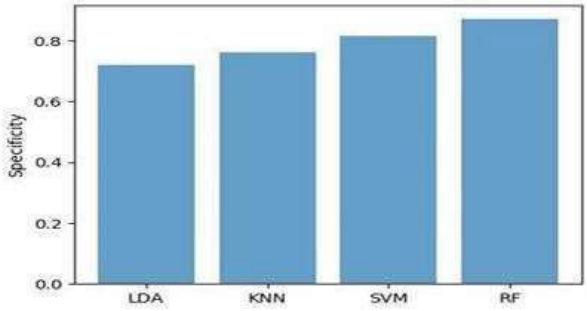

Fig. 6. Specificity measures of all classifier.

V. CONCLUSION AND FUTURE WORK

One of the most dangerous diseases in the real world is diabetes mellitus, and it can be difficult to diagnose in its early stages. This study creates a model that effectively addresses all the difficulties and aids in the early detection of diabetes using machine learning classification techniques. The studies are conducted using the PIDD dataset and four machine learning algorithms: RF, SVM, KNN, and LDA. The data set includes 768 records and 8 important diabetes- related variables with a class label that displays the outcomes of patients with and without diabetes. Our primary goal is to attain the highest precision, recall, specificity, and F-score, among other crucial performance parameters, have been assessed in addition to the model's accuracy. Confusion measures like true positive, true negative, false positive, and false negative are used to analyze these performance parameters. According to the results, Random Forest (RF) surpassed the other classifiers in use and provided the highest accuracy of 87.66%. Therefore, our model uses the RF classifier..

In the future, we hope to use machine learning and artificial intelligence technologies to expand our work in the prediction of other diseases, such as cancer and psoriasis.

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.