Automated Diabetes Diagnosis and Risk Assessment using Machine Learning

Deep Mathukiya
Department of Computer Science and Engineering, School of Technology, Pandit Deendayal Energy University, Gandhinagar, India

Yogesh Kumar
Department of Computer Science and Engineering, School of Technology, Pandit Deendayal Energy University, Gandhinagar, India

Apeksha Koul
Punjabi University, Patiala, Punjab, India

ABSTRACT: Currently, Diabetes is a prevalent issue worldwide, affecting millions of people. The detection machinery for this disease is only available at healthcare centers, so individuals with diabetes are only aware of their condition when they visit and undergo a checkup. Prolonged diabetes can lead to various complications such as heart disorders, nerve damage, kidney disease, and diabetic retinopathy. However, the risk can be reduced if the disease at its early stage is predicted. By utilizing a deep learning approach, this model can easily determine if an individual is diabetic or not. The aim of the research is to develop machine learning model that can identify diabetic people on the basis of their clinical data. The machine learning algorithms utilized in this study include Naive Bayes (NB), Decision tree (DT), Random Forest (RF), k-nearest neighbor (KNN), Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), and Logistic Regression (LR). We have implemented efficient pre-processing techniques to detect and handle missing/ NAN and duplicate values and later trained our models with the cleaned dataset. During execution, we found that Random Forest as well as Decision Tree performed well for both diabetic and non-diabetic class by computing 100% and 99% accuracy respectively.

1 INTRODUCTION

The transcendence of diabetes, a steady metabolic issue portrayed by raised blood glucose levels, has shown up at upsetting degrees generally, making it one of the most crushing clinical consideration troubles of the 21st 100 years. According to the World Prosperity Affiliation (WHO), a normal 463 million persons had diabetes in 2019, and if current trends continue, this figure is expected to increase to 700 million by 2045. Reasonable organization and ideal intervention are essential to thwart troubles and work on the individual fulfilment for individuals living with diabetes [1]. Simulated intelligence (ML) strategies have emerged as essential resources for the assumption, gathering, and the chiefs of diabetes in light of their ability to research complex data models and concentrate huge encounters. The prevalence of diabetes on a global scale has reached epidemic levels, necessitating creative techniques for timely discovery and effective management. In this particular circumstance, AI (ML) algorithms have emerged as essential tools for predicting and classifying diabetes in light of a vast array of characteristics [2]. This study focuses on a comprehensive set of markers, including hypertension, orientation, age, a history of cardiovascular illness, a tendency to smoke, BMI, hbA1c levels, and glucose levels. Together, these characteristics present a large dataset with the potential to significantly increase the accuracy of diabetes prediction and categorization. By outfitting the force of ML calculations, medical services experts and specialists can use the complicated connections among these factors to fabricate prescient models that can recognize people in danger of diabetes or characterize them as diabetic or non-diabetic. The incorporation of different boundaries like orientation, age, and way of life propensities recognizes the intricacy of diabetes as a condition impacted by hereditary, ecological, and conduct factors. This approach adds to exact individualized expectations as well as establishes the ground work for a more nuanced comprehension of diabetes, the study of disease transmission on a worldwide scale.

2 LITERATURE WORK

Zhou et al. [3**]** presented in their work that in contemporary times, diabetes stands apart as one of the most pervasive constant sicknesses all around the world, conveying huge dangers because of expected confusions. Identifying diabetes at a beginning phase accepts fundamental significance as it empowers brief mediation, accordingly stopping the infection's movement. The technique introduced not just holds the possibility to estimate future diabetes events yet additionally has the capacity to recognize the particular diabetes type an individual might experience. Given the significant contrasts in treatment approaches between type 1 and type 2 diabetes, the technique serves to precisely tailor treatment plans for individual patients. Approach changes the test into an order task, dominatingly developing a model through the idle layers of a profound brain organization. Dropout regularization is consolidated to counter overfitting. Through careful boundary tuning and utilizing the twofold cross-entropy misfortune capability, a profound brain network forecast model of exceptional exactness is accomplished. Observational results highlight the viability and fittingness of the proposed DLPD (Profound Learning for Anticipating Diabetes) model. Quiet, the best preparation exactness for the diabetes type dataset comes to 94.02174%, while for the Pima Indians diabetes dataset, it achieves a noteworthy 99.4112% preparation precision. **Luis Fregoso et al. [4]** mentioned that when blood glucose levels exceed certain thresholds, the condition known as diabetes mellitus develops. In recent times, learning models have been harnessed to forecast diabetes and its associated problems. However, within the realm of constructing predictive models for type 2 diabetes, researchers and developers continue to confront two primary hurdles. The systematic review of Luis et al. endeavours to address the aforementioned challenges comprehensively. The review adheres primarily to the PRISMA methodology, augmented by the approach advocated by Keele and Durham Universities. **Muneer Butt et al. [5]** presented a sizable quantity of crucial and sensitive healthcare data has been produced as a result of significant advancements in biotechnology and the growth of the public healthcare infrastructure. Intelligent data analysis methods have been used to reveal fascinating patterns that have made it easier to identify and stop the onset of a number of serious diseases. Due to its role in potentially fatal illnesses like heart, kidney, and nerve damage, diabetes mellitus has an especially dire position. **Nadeem et al. [6]** mentioned about making informative driven applications and administrations for diagnosing and sorting critical ailments faces difficulties due to deficient and below average logical information for preparing and approving calculations. Internationally, the mounting costs related with overseeing diabetes, a far and wide persistent infirmity portrayed by delayed times of raised glucose levels, are applying huge tensions on medical services suppliers. The proposed arrangement holds the possibility to raise endurance rates among PwD by working with individualized treatment suggestions. Likewise, **Tasin Ullah & Sajida [7]** discussed that globally, diabetes affects approximately 537 million people and make it the most prevalent as well as deadly non-communicable illness. Several factors, such as being overweight, having abnormal cholesterol levels, lacking any physical activity, history of any family member having diabetes, and unhealthy eating habits are responsible to increase the risk of giving rise to diabetes. A paper of them presents an automated system for predicting diabetes, where the machine learning models are trained with the private dataset collected from female patients of Bangladesh. **Ahmed et al. [8]** mentioned in their paper that diabetes is a widespread disease and it can be associated with serious complications such as kidney failure and heart disease. Early detection of diabetes can lead to a longer and healthier life. Using various supervised machine learning models trained on suitable data, it is possible to diagnose diabetes in its early stages. Preprocessing techniques such as label encoding and normalization are used to improve model accuracy. In addition, different feature selection methods are used to identify and priorities key risk factors. Rigorous experiments evaluate the performance of the model on different datasets. **Birjais et al. [9]** stated that when paired with data mining techniques, machine learning, a subset of artificial intelligence, offers enormous potential, particularly in the area of forecasting. Today, data generation is growing, but if it isn't turned into useful information, it’s worth remains untapped. This is also true in the field of health care, where there is a wealth of information that requires its collection to improve prognosis, diagnosis, treatment, drug development and the general advancement of health care. In study, they mainly focus on the diagnosis of diabetes, a chronic disease that is rapidly spreading worldwide, as defined by the World Health Organization in 2014. **Similalry, Shafi & Ansari [10]** mentioned that Diabetes is a global health problem that often lasts throughout the patient's life. Its influence extends to all age groups. Embracing technological advances provides an unusual opportunity to predict diabetes, giving accurate results and improving efficiency. A considerable number of studies that have sought to predict diabetes have used the Pima Indian dataset. Study of Shafi gives a paradigm for calculating a patient's likelihood of having diabetes as accurately as possible. In order to identify diabetes at an early stage, the authors apply machine learning techniques, including decision trees, SVM, and Naive Bayes. The study makes use of the Pima Indian Diabetes dataset, which is available at the UCI Library, to expedite the procedure and produce accurate results. The work attempts to maximize the precision and effectiveness of diabetes detection using these techniques. **Fatih Aslan & Kadir Sabanci [11]** proposed an innovative approach of deep learning for the detection of diabetes at an early stage. As PIMA datasets typically only contain numeric values, it is not feasible to use well-known CNN models. To address issue, the study generated visualisations from numerical data based on the significance of features that leverage the powerful performance of CNN models in the early diagnosis process. The resultant diabetes image data is classified using three distinct classification strategies. Initially, the original strategy was to feed diabetes images to a pair of CNN models, such as ResNet18 or ResNet50. In a second approach, the deep features from the ResNet model are combined with support vector machines (SVM) and classified using SVM. Finally, a third approach involves the classification of selected fusion features using support vector machine. **Awasthi et al. [12]** conducted assessment of multiple diabetes prediction models stemmed from the need to locate, critically evaluate, and amalgamate pertinent and high-quality individual research findings. The primary goal of study was to identify optimal approaches for the selection and synthesis of high-quality studies. Medical data, which is predominantly nonlinear, characterized by intricate correlation structures, poses a significant analytical challenge. Notably, the utilization of machine learning in healthcare and medical imaging has been excluded from consideration.

3 FRAMEWORK FOR DIABETIC DETECTION

The system design as per the Figure 1 provides a comprehensive visualization of the end-to-end process for predicting diabetes using machine learning. It starts by introducing the dataset and proceeds through crucial stages like data analysis and preprocessing, ensuring data readiness followed by data splitting. Various machine learning models have been applied and later their performances have been examined for both diabetic and non-diabetic classes.

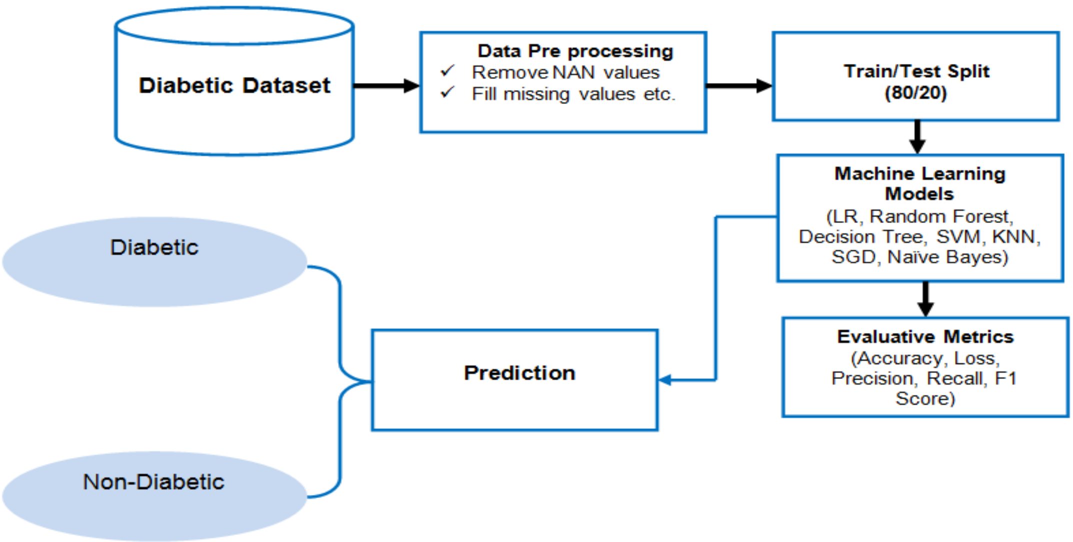

Fig.1 System design for the prediction of Diabetes

3.1 *Description of the dataset*

The Diabetes prediction dataset encompasses medical and demographic information sourced from patients, accompanied by their diabetes status (positive or negative). This dataset comprises attributes like age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. It serves as a foundational resource for constructing machine learning models intended to forecast diabetes in patients based on their medical records and demographic particulars. This dataset holds potential utility for healthcare professionals in identifying individuals who might be prone to diabetes development, facilitating the formulation of customised treatment strategies. Table 1 presents few attributes of the dataset.

Table 1: Diabetes dataset attributes with their description

Attributes	Description
Gender	It defines the sexuality of an individual and also impacts the suspension of having diabetes
Age	It also plays an important role because diabetes is mostly seen in adult generation.
Hypertension	It is blood pressure that consistently elevates in the arteries of the heart.
Heart Disease	This medical illness is also linked with the increasing risk of having diabetes.
Body Mass Index	It is for calculating the fat in the human body by considering their height as well as weight. In this dataset, the value of body mass index ranges from 10.16 to 71.55 where <18.5 indicates underweight and 18.5 to 24.9 classifies overweight and above 30 is obese.

3.2 *Data preprocessing*

Pre-processing the dataset is an important step to extract meaningful insights and construct accurate predictive models. Initially, Pandas DataFrame is used to load the data and enables efficient data manipulation and exploration. It provides robust functionality to identify missing values in any column as well as identify duplicate records and process them to ensure dataset accuracy. This thorough approach to data pre-processing lays a solid foundation for robust analysis and modeling.

3.3 *Train-Test splitting*

Performing a train-test split on dataset is crucial to assess the performance of machine learning model for the detection of diabetic and non-diabetic patients. In this research paper, we have Scikit library to split the dataset in training and testing in a ration 80:20. After the split, the training set trains the model's parameters, while the testing set assesses its predictive accuracy. This approach helps prevent overfitting, where the model performs well on the training data but fails to generalise to new data.

3.4 *Applied machine learning classifiers*

Logistic regression, a fundamental binary classification algorithm, is commonly employed in diabetic and non-diabetic detection due to its interpretability and efficiency. By modeling the probability of an input that belongs to a specific class, it efficiently categorizes data into two distinct groups [12]. For binary classification, it can be computed by using equation (i):

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \ldots + \beta_3 x_3 \qquad \text{(i)}$$

Here, p is the probability, ln is the natural logarithm, β_0 is the intercept of the equation and rest of them are the coefficients associated with the predicted variables x_1, $x_2 \ldots x_n$. ***Naive Bayes***, despite its simplistic assumption of independence, is surprisingly effective in tasks like diabetes prediction, especially in text-based contexts like medical records [13]. It is computed by equation (ii)

$$Posterior\ Prob\ of\ class\ A = \frac{Prior\ Prob\ of\ class\ A\ \text{x}\ Likelihood\ prob\ of\ class\ A}{Prior\ prob\ without\ any\ specific\ class} \qquad \text{(ii)}$$

Stochastic Gradient Descent, an optimization technique, is invaluable for training large-scale models crucial in diabetes detection. Its primary objective is to iteratively minimize the loss function by updating the model parameters. It is particularly valuable for training large models and is often used with lot of machine learning approach [14]. ***K-Nearest Neighbors'*** versatility makes it applicable for diabetic and non-diabetic classification, as it predicts based on the majority class of its nearest neighbors, a useful characteristic in handling various diabetes-related features [15]. It is computed by equation (iii).

$$Distance(X_{new}, X_i) = \sqrt{\sum_{j-1}^{n}(X_{new.j} - X_{i,j})^2} \quad \text{(iii)}$$

Decision trees, with their intuitive tree-like structures, make feature-based decisions aiding in diabetes prediction scenarios. They are popular models for classification and regression tasks. They create a tree-like structure that partitions the feature space on the basis of attribute values. Every internal node and leaf node represents a feature-based decision, and a predicted class or value respectively while as ***Random Forest,*** an ensemble method, merges predictions from multiple decision trees, enhancing accuracy and mitigating overfitting, making it pivotal in diabetes prognosis [16]. ***Support Vector Machine,*** highly effective in high-dimensional data, is adept at discerning clear boundaries between diabetic and non-diabetic classes, a vital aspect in precise diabetes detection and management [17].

3.5 *Evaluative Parameters*

Various parameters, as shown in Table 2, have been used to examine the performance of the applied learning models while being trained with the diabetic dataset [18].

Table 2: Evaluative parameters of the model

Parameter	Definition
Accuracy	The proportionate count of accurate predictions
Loss	The degree to which the model's predictions are inaccurate.
Precision	The proportion of positive predicts that prove to be accurate.
Recall	The proportion of accurate predictions made for true positives.
F1 Score	A single value that integrates precision and recall as a metric.

4 RESULTS

The assessment of different machine learning models in a binary classification task involving non-diabetic and diabetic cases, as shown in Table 3.

Table 3: Accuracy and loss values for each model on diabetic and non-diabetic

Models	Diabetic		Non-Diabetic	
	Accuracy	Loss	Accuracy	Loss
Logistic Regression	0.99	0.011	0.63	0.37
Naïve Bayes	0.93	0.073	0.65	0.35
SGD	0.97	0.025	0.71	0.29
KNN	1	0.0039	0.6	0.4
Decision Tree	1	1.4e-05	0.99	0.0084
Random Forest	1	2.7e-05	0.99	0.0084
SVM	1	1.4e-05	0.39	0.61

The accuracy values, along with their corresponding "loss" values, were evaluated for each model. For instance, the Logistic Regression achieved a high accuracy of 0.99 for non-diabetic cases and 0.63

for diabetic cases, corresponding to "loss" values of 0.011 and 0.37, respectively. Similarly, the Naïve Bayes model demonstrated an accuracy of 0.93 for non-diabetic cases and 0.65 for diabetic cases, with corresponding "loss" values of 0.073 and 0.35. Stochastic Gradient Descent achieved an accuracy of 0.97 for non-diabetic cases and 0.71 for diabetic cases, resulting in "loss" values of 0.025 and 0.29 for the respective classes. The K-Nearest Neighbours model attained perfect accuracy (1) for non-diabetic cases and an accuracy of 0.6 for diabetic cases, leading to "loss" values of 0.0039 and 0.4. Remarkably, both the Random Forest and Decision Tree models achieved perfect accuracy (1) for both non-diabetic and diabetic cases, accompanied by extremely low "loss" values of 1.4e-05 and 2.7e-05 for both classes, respectively. While the Support Vector Machine (SVM) reached perfect accuracy (1) for non-diabetic cases, its accuracy for diabetic cases was 0.39. This discrepancy resulted in "loss" values of 1.4e-05 for non-diabetic and 0.61 for diabetic cases.

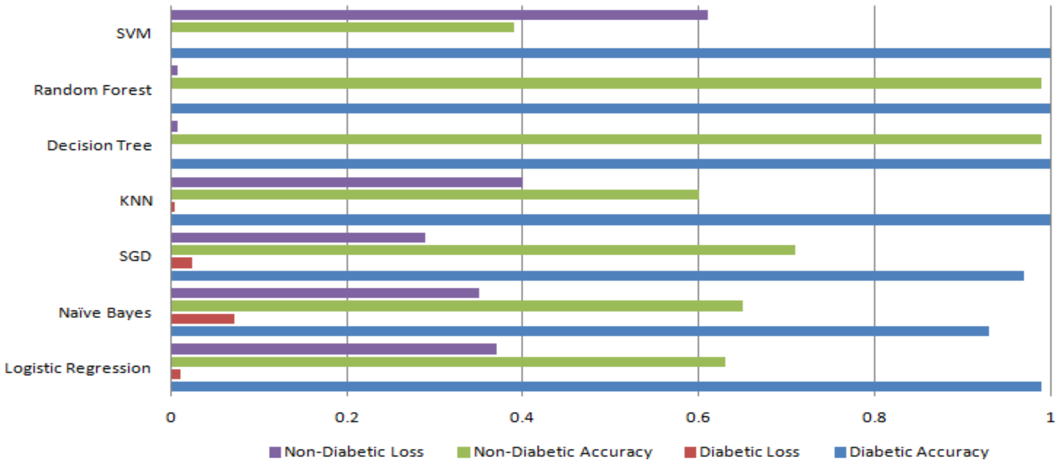

Fig.2 Graphical analysis of models

Remark of Figure 2 depicts that the Decision Tree as well as Random Forest models demonstrated exceptional accuracy by achieving perfect scores across both classes. Other models also showed favourable performances with varying accuracy levels. Selecting the most suitable model should be based on specific task requirements and trade-offs.

The performance of various machine learning models in predicting diabetic and non-diabetic cases was assessed using precision, recall, and F1 score metrics, as shown in Table 4.

Table 4: Evaluation parameters for diabetic and non-diabetic

Models	Diabetic			Non-diabetic		
	Precision	Recall	F1 score	Precision	Recall	F1 score
Logistic Regression	0.84	0.63	0.72	0.97	0.99	0.98
Random Forest	1	1	1	1	0.99	1
Naïve Bayes	0.45	0.65	0.53	0.97	0.93	0.95
SVM	1	0.39	0.56	0.95	1	0.97
KNN	0.94	0.60	0.73	0.96	1	0.98
Decision Tree	1	0.99	1	1	1	1
SGD	0.62	0.76	0.68	0.98	0.96	0.97

Among the models evaluated, Logistic Regression demonstrated a respectable balance between precision, recall, and F1 score by computing 0.84, 0.63, and 0.72 respectively for diabetic cases. On the contrary, it has achieved the highest precision of 0.97, recall of 0.99, and F1 score of 0.98 for non-diabetic cases. Naïve Bayes, although having a lower precision (0.45) for diabetic cases, showed a relatively higher recall (0.65). However, its overall F1 score (0.53) indicated room for improvement. SGD exhibited decent performance with a precision of 0.62 and recall of 0.76 for diabetic cases, along with a respectable F1 score of 0.68. KNN showcased high precision (0.94) but a relatively lower recall (0.60) for diabetic cases, leading to an F1 score of 0.73. Decision Tree and Random Forest models performed exceptionally well, achieving perfect precision, recall, and F1 score for both diabetic and non-diabetic cases. Lastly, SVM demonstrated perfect precision (1) for diabetic cases but had a lower recall (0.39) and F1 score (0.56), indicating challenges in identifying all diabetic cases. In summary, Decision Tree and Random Forest models outperformed others, exhibiting flawless accuracy, while Logistic Regression also showed commendable overall performance in classifying diabetic and non-diabetic cases. The same parameters are also being graphically presented in Figure 3.

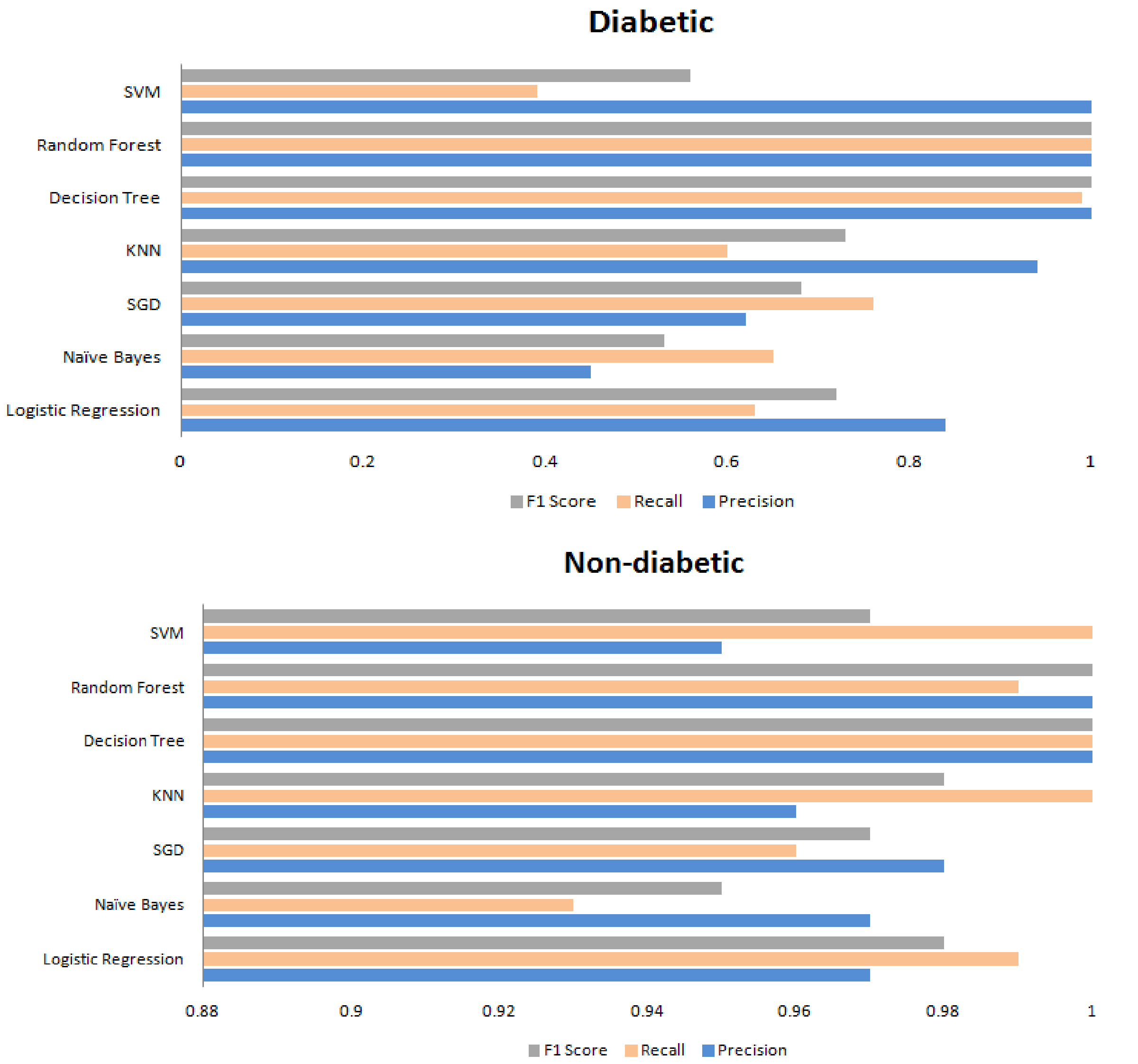

Fig.3 Evaluation of models for diabetic and non-diabetic class

5 CONCLUSION

In this study, a comprehensive evaluation of several machine learning models was performed, and the results are summarized in the accompanying graphs. These models which include naive Bayes, random forest, stochastic gradient, support vector machine, logistic regression, decision tree, and K-nearest neighbours, are evaluated using three basic performance metrics: precision, recall and F1 score. These metrics are key indicators of the models' performance in various classification tasks. In particular, Decide Tree and Random Forest achieved perfect scores for precision and recall, highlighting their potential suitability for applications where it is important to achieve the highest levels of precision and recall. In contrast, Naive Bayes showed a trade-off between precision and recall, making it a worthy choice when the main consideration is between false positives and false negatives. These findings provide valuable information for selecting the most appropriate machine learning model based on specific application requirements. Nevertheless, it is crucial to recognize the constraints of our study. The success of machine learning models relies heavily on the caliber and volume of data. With the ongoing progress in machine learning algorithms and healthcare data collection, there is potential for future research to delve into more advanced algorithms and diverse datasets. This could lead to even more accurate and widely applicable results.

6 REFERENCES

[1] Parimala, G., Kayalvizhi, R. and Nithiya, S., 2023, January. Diabetes Prediction using Machine Learning. In *2023 International Conference on Computer Communication and Informatics (ICCCI)* (pp. 1-10). IEEE.

[2] Kaur, S., Bansal, K. and Kumar, Y., 2022, December. Artificial Intelligence approaches for Predicting Hypertension Diseases: Open Challenges and Research Issues. In *2022 5th International Conference on Contemporary Computing and Informatics (IC3I)* (pp. 338-343). IEEE

[3] Zhou, H., Myrzashova, R. and Zheng, R., 2020. Diabetes prediction model based on an enhanced deep neural network. *EURASIP Journal on Wireless Communications and Networking*, *2020*, pp.1-13.

[4] Fregoso-Aparicio, L., Noguez, J., Montesinos, L. and García-García, J.A., 2021. Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. *Diabetology & metabolic syndrome*, *13*(1), pp.1-22.

[5] Butt, U.M., Letchmunan, S., Ali, M., Hassan, F.H., Baqir, A. and Sherazi, H.H.R., 2021. Machine learning based diabetes classification and prediction for healthcare applications. *Journal of healthcare engineering*, *2021*.

[6] Nadeem, M.W., Goh, H.G., Ponnusamy, V., Andonovic, I., Khan, M.A. and Hussain, M., 2021, October. A fusion-based machine learning approach for the prediction of the onset of diabetes. In *Healthcare* (Vol. 9, No. 10, p. 1393). MDPI.

[7] Tasin, I., Nabil, T.U., Islam, S. and Khan, R., 2023. Diabetes prediction using machine learning and explainable AI techniques. *Healthcare Technology Letters*, *10*(1-2), pp.1-10.

[8] Ahmed, N., Ahammed, R., Islam, M.M., Uddin, M.A., Akhter, A., Talukder, M.A. and Paul, B.K., 2021. Machine learning based diabetes prediction and development of smart web application. *International Journal of Cognitive Computing in Engineering*, *2*, pp.229-241.

[9] Birjais, R., Mourya, A.K., Chauhan, R. and Kaur, H., 2019. Prediction and diagnosis of future diabetes risk: a machine learning approach. *SN Applied Sciences*, *1*, pp.1-8.

[10] Shafi, S. and Ansari, G.A., 2021, May. Early prediction of diabetes disease & classification of algorithms using machine learning approach. In *Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021)*.

[11] Aslan, M.F. and Sabanci, K., 2023. A novel proposal for deep learning-based diabetes prediction: Converting clinical data to image data. *Diagnostics*, *13*(4), p.796.

[12] A. Awasthi, I. Gangwal, M. Jain, 2022 Diabetes Prediction using Machine Learning: A Review International Journal for Research., ISSN: 2321-9653; IC Value: 45.98.

[13] Kaur, K., Singh, C. and Kumar, Y., 2022, December. Artificial Intelligence Techniques for the Detections of Congenital Diseases: Challenges and Research Perspectives. In *2022 5th International Conference on Contemporary Computing and Informatics (IC3I)* (pp. 888-893). IEEE.

[14] Thakur, K., Kaur, M. and Kumar, Y., 2022, October. Artificial Intelligence Techniques to Predict the Infectious Diseases: Open Challenges and Research Issues. In *2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS)* (pp. 109-114). IEEE.

[15] Koul, A., Kumar, Y. and Gupta, A., 2022, October. A study on bladder cancer detection using AI-based learning techniques. In *2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS)* (pp. 600-604). IEEE.

[16] Bhardwaj, P., Kumar, S. and Kumar, Y., 2022, May. Deep Learning Techniques in Gastric Cancer Prediction and Diagnosis. In *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)* (Vol. 1, pp. 843-850). IEEE.

[17] Kumar, Y., Patel, N.P., Koul, A. and Gupta, A., 2022, February. Early prediction of neonatal jaundice using artificial intelligence techniques. In *2022 2nd international conference on innovative practices in technology and management (ICIPTM)* (Vol. 2, pp. 222-226). IEEE.