Effective Diabetes Prediction using an IoT-based Integrated Ensemble Machine Learning Framework

Rashi Rastogi
Department of CSE, Shobhit Institute of Engineering and Technology, Meerut, India
rastogi.rashi4@gmail.com (corresponding author)

Naveen Kumar
Salesforce Inc., Dallas, Texas, USA
nkumar5@salesforce.com

Mamta Bansal
Department of CSE, Shobhit Institute of Engineering and Technology, Meerut, India
mamta.bansal@shobhituniversity.ac.in

Ram Avtar Jaswal
Department of Electrical Engineering, UIET, Kurukshetra University, Kurukshetra, India
ravtar2015@kuk.ac.in

Priti Singla
Department of CSE, Chandigarh University, Punjab, India
pritisingla04@gmail.com

Sanjay Singla
Department of CSE, Chandigarh University, Punjab, India
dr.ssinglacs@gmail.com

Received: 31 August 2024 | Revised: 5 October 2024, 27 October 2024, and 3 November 2024 | Accepted: 25 December 2024

Licensed under a CC-BY 4.0 license | Copyright (c) by the authors | *DOI: https://doi.org/10.48084/etasr.8869*

ABSTRACT

Diabetes, a prevalent chronic disease, affects a significant global population. Identifying and being aware of key variables promptly may substantially enhance results for both patients and public health efforts. Systematic methods such as monitoring diabetic patients allow the collection of extensive data from diabetic patients. When it comes to keeping track of a patient's health, IoT sensors, such as those used in diabetic patient monitoring systems, are invaluable. Blood glucose levels, body temperature, and location of a diabetic patient can be tracked and recorded through a monitoring device. In addition to monitoring patients, these data can be classified using Machine Learning (ML) methods. This study applies three ML models to three different diabetes datasets and analyzes their performance. According to the results, the fine-tuned random forest model achieved higher accuracy, i.e., 89%, 90%, and 99%.

Keywords-diabetes prediction; machine learning; diabetes patent monitoring; IoT; chronic illness

I. INTRODUCTION

The technologies associated with IoT and its applications have experienced significant advances, resulting in increased accessibility and availability [1]. This has allowed the interconnection of numerous objects through the Internet across various domains, including, but not limited to, healthcare, home automation, and industrial manufacturing. In the domain of intelligent healthcare, particularly within the domain of patient monitoring, patient data is highly valuable. To implement an IoT application in this particular domain, it is important to ensure the comprehensive capture of a substantial volume of data obtained through the monitoring of medical indicators in patients [2-4]. Diabetes poses a significant threat

to public health and has become an important contributor to global mortality rates [5]. Effective management of this condition requires diligent surveillance and careful attention to patient well-being. Diabetes is characterized by insulin resistance, which can result in inadequate insulin secretion and subsequent fluctuations in blood glucose levels. Consequently, people with diabetes cannot effectively regulate their glucose levels within a certain range. If patients can no longer follow these conditions, urgent medical attention is necessary to prevent further deterioration [6]. The rise in the global diabetes population has led to an increased utilization of continuous Glucose Monitoring Devices (CGM) to ensure continuous monitoring [7].

A. *Effects of Diabetes*

There are broadly three effects of diabetes on the human body:

- Loss of vision: Damage to the retina and optic nerve causes retinopathy. Vision impairment at night can be caused by retinal edema, which can also reduce mental engagement. A diabetic's initial eyesight loss can be managed with a few simple tests and medications.
- Kidney neuropathy: As blood sugar levels rise, damage to the nerves and blood vessels in the kidneys (known as chronic kidney disease or diabetic neuropathy) becomes increasingly likely. The elimination of waste and the absorption of a lot of fluids are two of the kidney's many useful functions.
- Liver problems: The liver plays a crucial role in maintaining a healthy blood glucose level by breaking down starch through the gluconeogenesis and glycogenosis processes. Type 2 diabetes is associated with an increased risk of developing liver problems.

B. *Contribution of this Study*

This study responds to the recently rising issue of diabetes, presenting a new method of diabetes care in the context of IoT. This study's major contribution is the proposal of a fine-tuned Random Forest (RF) model with a hyperparameter tuning approach to increase the accuracy of diabetes prediction. Tuning can also help reduce the size of the model. This procedure offers many advantages in terms of exactness, speed, and compatibility with various datasets. To substantiate the significance of the proposed framework, the performance of the model was systematically evaluated on three actual diabetes datasets, indicating the feasibility of the proposed model for efficient and reliable diabetes tracking.

II. RELATED WORKS

In [8], an IoT application framework called the E-Healthcare Monitoring System (EHMS) was proposed, which used ML strategies to assist in accurate diagnosis. In [9], a system for real-time and remote health monitoring was developed based on the IoT infrastructure and integrated with cloud computing. Various ML algorithms were applied to public healthcare datasets stored in the cloud. The system provided suggestions based on the wealth of historical and empirical information stored in the cloud. A method for extracting information from databases was presented, illuminating hidden patterns that aid in making sound judgments. In [10], the Diabetes Prediction System with Supervised Learning (SLDPS) was presented, which is a decision tree model. The current state of diabetes research and future directions were discussed in [11]. Long-term complications are often associated with even the most common diseases, and diabetes is no exception. It is crucial to build a system that stores and analyzes diabetes data and further predicts potential risks with the help of technology. In [12], an efficient Urine-based Diabetes (UbD) monitoring system was presented, which was suitable for use in the well-being of one's own home. This approach was verified by extensive experimental simulations, outperforming state-of-the-art decision-making methods in terms of latency, classification accuracy, robustness, and stability. In [13], an intelligent patient health monitoring system was proposed to correctly and quickly identify the existence of chronic conditions, using the diabetes dataset along with six different ML methods. Diabetes mellitus was the topic in [14]. For diabetes diagnosis, it is important to survey how to manage enormous data files, define technologies, and debate using IoT sensors and management. The study in [15] focused on definitive results from diabetes research, illuminating the path taken by other studies. In [16], a mechanism for the prediction of diabetes was proposed using four ML models and testing their performance. In [17], ML algorithms were used on data from different sources related to diabetes, including IoT sensors and questionnaires. Moreover, the algorithms under discussion were used for prediction in the case of the PIMA dataset. The results showed that the proposed models had significant results in terms of accuracy and recall, and the most significant factors were age, family history, physical activity, and regular intake of medications.

III. SYSTEM MODEL

IoT refers to a network that includes interconnected physical devices that can be accessed over the Internet. The use of IoT in the healthcare sector offers the advantage of real-time data collection and analysis. Figure 1 depicts the application of this innovative approach within a hospital equipped with IoT technology. Individuals diagnosed with diabetes will receive identification cards linked to a highly secure cloud-based platform that serves as a repository for their electronic medical information. The accessibility of this information on a tablet or computer could enhance the efficacy and efficiency of patient care while facilitating ease of use for physicians and personnel. Figure 2 shows the types of sensors, such as glucose sensors, body temperature sensors, and BP sensors, used to collect data from patients. These sensors can be used for data collection and those data are preprocessed and used by ML algorithms to predict accurate diabetes results.

IV. PROPOSED FRAMEWORK

Diabetes is a chronic condition that affects millions of people worldwide and requires continuous monitoring and management of blood glucose levels. Traditional monitoring techniques involve manual blood glucose testing, which can be time-consuming and may not provide real-time data. To enhance the quality of life of diabetics, an integrated system utilizing IoT technology and ML algorithms for continuous

glucose monitoring and predictive analysis is required. The objective of this study is to design and implement a system that can effectively monitor blood glucose levels using IoT devices, gather and analyze data, and provide predictive insights to help individuals more proactively manage their diabetes. This study investigated diabetes prediction using ML-based algorithms such as Naïve Bayes (NB), KNN, and an improved fine-tuned RF (FT_RF). Initially, three different real datasets of diabetes patients were collected, as shown in Figure 3. Subsequently, the collected data were preprocessed to extract the necessary information for analysis. Furthermore, data oversampling was performed before training the models. Oversampling is a technique employed in ML to address the problem of class imbalance in datasets. Class imbalance refers to a situation in a classification issue where one class or category of data is significantly more prevalent than another. This discrepancy can lead to biased models that have insufficient performance for minority populations.

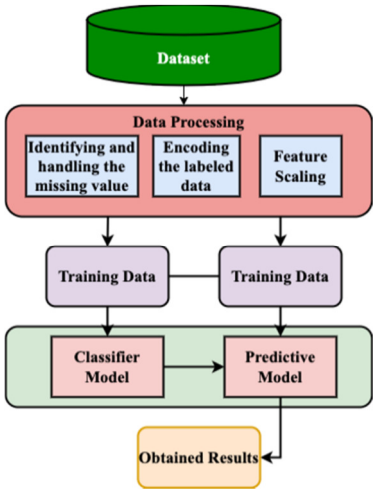

Fig. 1. Proposed system model.

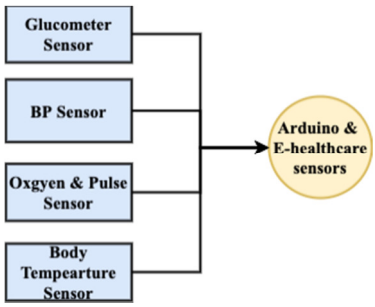

Fig. 2. IoT sensors used for diabetes monitoring.

Oversampling is augmenting the presence of the minority class by creating artificial instances or replicating existing ones. This facilitates the ML algorithm in obtaining information from a dataset that is more evenly distributed, therefore diminishing its inclination toward the dominant class. During the subsequent phase, the models underwent training and the RF model was meticulously adjusted to obtain the most favorable outcomes. Optimizing the performance of an RF model and improving its accuracy requires tuning its hyperparameters. RF is an ensemble learning technique that predicts by combining multiple decision trees. The hyperparameters of these trees can be altered to improve the predictive ability of the model.

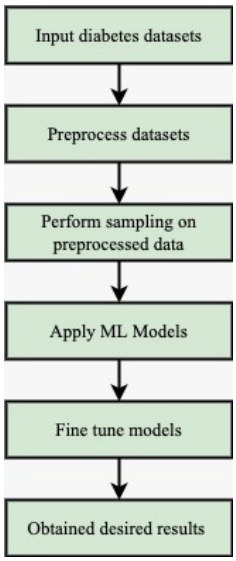

Fig. 3. The proposed framework.

This study used the mass balance equation for diabetes dataset attributes, which is depicted as:

$$V\frac{M}{dt} = Q(M_{in} - M_{out}) + R_p - R_c \quad (1)$$

where V represents volume, M denotes the levels of insulin, glucose, or glucagon, t indicates time, Q stands for the blood flow rate, and R_p and R_c correspond to the metabolic production and consumption rates of the substance in the material balance, respectively. The proposed mechanism utilized three standard differential equations:

$$\begin{cases} \frac{dG_H(t)}{dt} = \frac{1}{V_H^G}\left[Q_L^G G_L(t) + \gamma_T^G\left(G_T(t)\right)^{\alpha} - Q_H^G G_H(t) - R_H^G\right], \\ \frac{dG_L(t)}{dt} = \frac{1}{V_L^G}\left[Q_A^G G_H(t) - Q_L^G G_L(t) + R_L^G\right], \\ \frac{dG_T(t)}{dt} = \frac{1}{V_T^G}\left[Q_P^G\left(G_H(t) - \gamma_T^G G_T(t)\right) - R_T^G\right], \end{cases} \quad (2)$$

Glucagon dynamics influence the rates at which glucose is metabolized. The effects of the hormone glucagon can be modeled using a single compartment. Glucagon is important and its role is shown as:

$$\frac{d\Gamma(t)}{dt} = \frac{1}{V^{\Gamma}}\left[\eta G_H I_H - R^{\Gamma}\Gamma\right] \quad (3)$$

where R^{Γ} is the rate of glucagon secretion from the pancreas and η is a multiplicative constant describing the clearance rate of glucagon in the plasma. In addition, the rise or fall in blood sugar levels triggers the production of insulin from the pancreas. The following pancreatic mass balance equations are proposed:

$$\begin{cases} \frac{dI_s(t)}{dt} = -p_s I_s(t) + \left(I_b(t)\right)^{\sigma} G_H(t), \\ \frac{dI_s(t)}{dt} = -p_b I_b(t) + \left(I_s(t)\right)^{\delta} G_H(t) \end{cases} \quad (4)$$

V. RESULTS AND ANALYSIS

The proposed mechanism involves a sophisticated RF classifier that was developed in Python and is available on Google Colab to diagnose diabetes using three different datasets. Before applying the models to the datasets, the data was cleaned, normalized, and split into training and testing. In this regard, using GridSearchCV, the hyperparameters including the number of trees in the forest or $n\_estimators$, the maximum depth of each tree max_depth, or the minimum number of samples required to split an internal node or built-in $min\_samples\_split$, and the maximum number of features from which the split at a node is created $max\_features$ can be changed. Fine-tuning helps increase the degree of correct results and makes the model better and more efficient. The following metrics were used to evaluate the ML models:

- Accuracy is given by the total successful predictions divided by the total incidences of the event of interest.
- The confusion matrix introduces the concepts of true positive, true negative, false positive, and false negative, offering a more nuanced way to illustrate the performance of a classifier.
- Precision measures the number of positive observations of the model relative to the actual true observations.
- The F1 score is defined as twice the precision times the recall divided by the precision plus the recall.
- Recall measures the ratio of accurate positive predictions to the number of actual positive observations.

A. Datasets Used

Three different diabetes datasets, having different modalities and sizes, were used to evaluate the performance of the proposed model.

- Pima Indian Diabetes Dataset: This dataset was originally obtained from the National Institute of Diabetes and Digestive and Kidney Diseases and includes diagnostic measurements to determine if a patient has diabetes or not. All patients in this dataset are females at least 21 years old of Pima Indian heritage [18].
- Diabetes Health Indicator Dataset: This dataset is based on the BRFSS survey administered by the CDC in 2015. This dataset comprises survey answers from more than 455 people and includes 330 elements associated with risk attributes to health, ongoing health issues, and the use of precautionary measures [19].
- Diabetes Prediction Dataset 1.5: This dataset contains medical and demographical profiles of the patients along with positive or negative diagnoses of diabetes. In that way, researchers can examine the association between different medical and demographic factors with the propensity of developing diabetes [20].

TABLE I. DATASET DETAILS

Dataset	Value
Dataset 1: Pima Indians Diabetes Dataset [18]	Rows and columns: 768, 9
Dataset 2: Diabetes Health Indicators Dataset [19]	Rows and columns: 253680, 22
Dataset 3: Diabetes Prediction Dataset 1.5 [20]	Rows and columns: 100000, 9

B. Results

1) Results on Pima Indians Diabetes Dataset

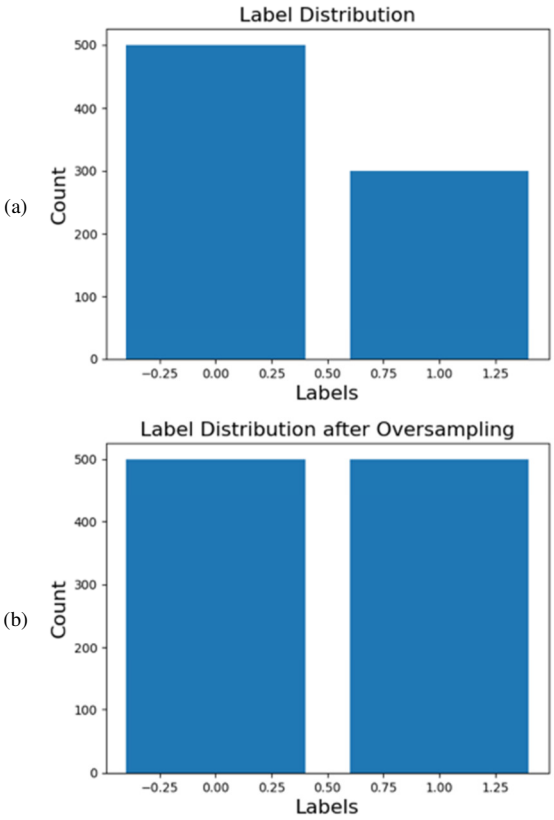

Fig. 4. Dataset before (a) and after oversampling (b).

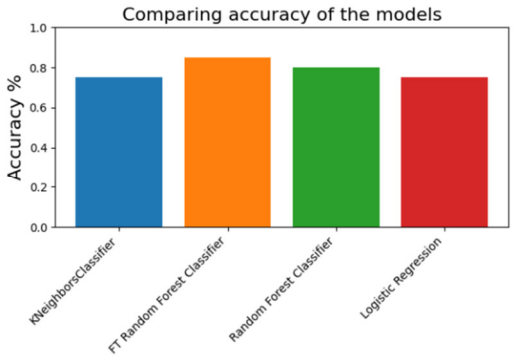

Fig. 5. Accuracy comparison of all models.

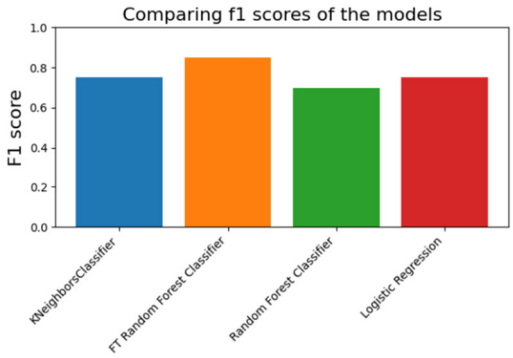

Fig. 6. F1 score comparison of all models.

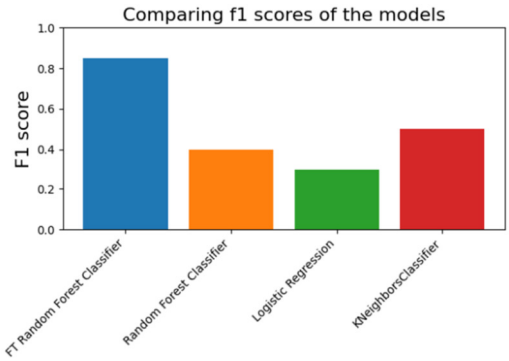

Fig. 9. F1 score comparison of all models

2) Results on Diabetes Health Indicators Dataset

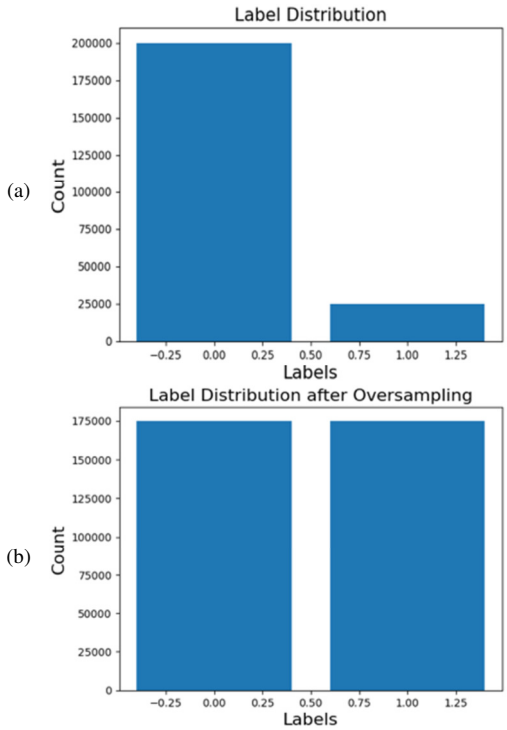

Fig. 7. Dataset before (a) and after oversampling (b).

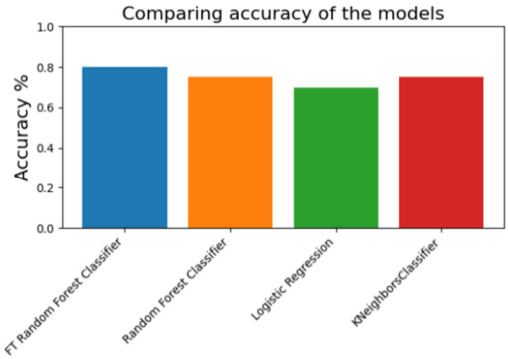

Fig. 8. Accuracy comparison of all models.

3) Results on Diabetes Prediction Dataset 1.5

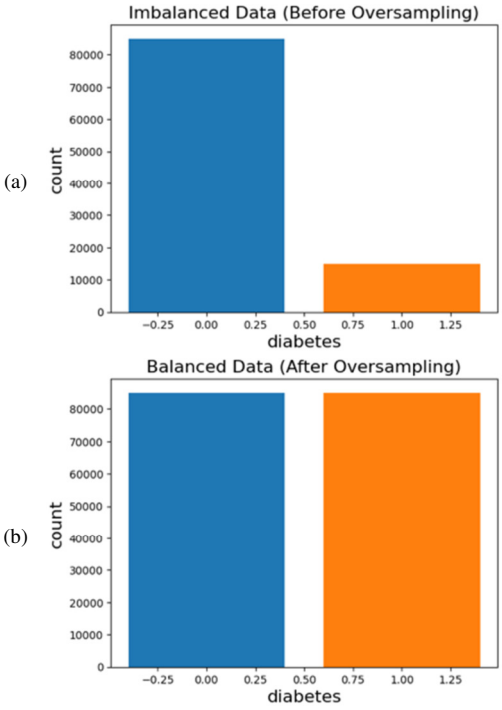

Fig. 10. Dataset before (a) and after oversampling (b).

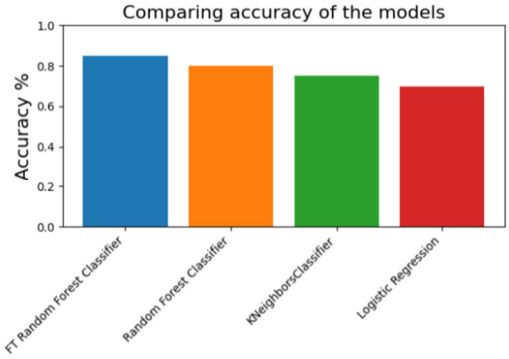

Fig. 11. Accuracy comparison of all models.

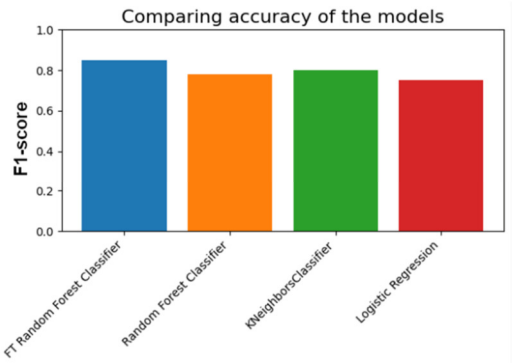

Fig. 12. F-1 score comparison of all models.

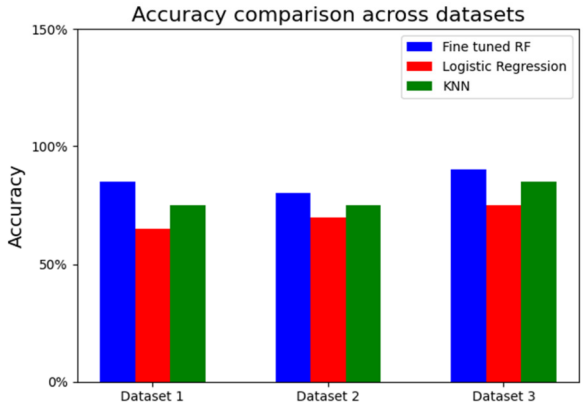

Fig. 13. Accuracy comparison of fine-tuned RF, LR, and KNN.

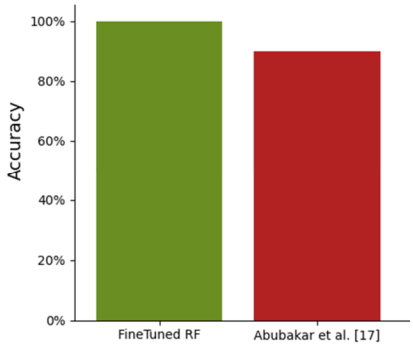

Fig. 14. Accuracy comparison of the proposed FT_RF and [17].

C. Discussion

The performance of the proposed FT_RF was compared with that of LR and KNN across three different datasets. Figures 4, 7, and 10 show the datasets before and after oversampling. Oversampling can reduce the distortion of the data so that actual data can be recovered for further processing. Figures 5, 8, and 11 compare the accuracy of all models on separate datasets. FT_RF achieved higher accuracy than the other models. Similarly, Figures 6, 9, and 12 compare the F1 scores of all models considered. Furthermore, Figure 13 compares the accuracy of all models on the three datasets. The accuracy of FT_RF was 90%, 93%, and 99%, which was higher than that of the other two models. Furthermore, Figure 14 illustrates the comparative performance of two techniques in terms of accuracy: FT_RF and [17]. The FT_RF model yielded a higher accuracy (99%) than [17]. The accuracy achieved by the technique in [17] was almost 90%. These improvements tend to support the hypothesis that the FT_RF model is a better solution compared to previous methods.

VI. CONCLUSION

Significant progress in the field of genetics and the development of robust public healthcare procedures have resulted in a substantial generation of crucial healthcare data. Through the application of smart data analysis tools, numerous intriguing patterns emerge related to the timely detection and prevention of various serious diseases. This study provides a framework for predicting diabetes patient outcomes on three different real datasets using three standard ML models. In addition, the RF model was fine-tuned to achieve higher accuracy. The results show that the FT_RF model achieved accuracy of 90%, 93%, and 99%. IoT devices can continuously monitor glucose levels and other vital parameters. ML algorithms can provide real-time feedback and decision support to patients and healthcare providers, helping them make informed decisions about insulin dosages, diet, exercise, and medication adjustments.

REFERENCES

[1] P. Sonar and K. JayaMalini, "Diabetes Prediction Using Different Machine Learning Approaches," in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, Mar. 2019, pp. 367–371, https://doi.org/10.1109/ICCMC.2019.8819841.

[2] M. A. R. Refat, Md. A. Amin, C. Kaushal, M. N. Yeasmin, and M. K. Islam, "A Comparative Analysis of Early Stage Diabetes Prediction using Machine Learning and Deep Learning Approach," in *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*, Solan, India, Oct. 2021, pp. 654–659, https://doi.org/10.1109/ISPCC53510.2021.9609364.

[3] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," *Procedia Computer Science*, vol. 165, pp. 292–299, Jan. 2019, https://doi.org/10.1016/j.procs.2020.01.047.

[4] R. Rastogi and M. Bansal, "Assessment on Different IoT-Based Healthcare Services and Applications," in *Emergent Converging Technologies and Biomedical Systems*, 2023, pp. 445–461, https://doi.org/10.1007/978-981-99-2271-0_35.

[5] S. Larabi-Marie-Sainte, L. Aburahmah, R. Almohaini, and T. Saba, "Current Techniques for Diabetes Prediction: Review and Case Study," *Applied Sciences*, vol. 9, no. 21, Jan. 2019, Art. no. 4604, https://doi.org/10.3390/app9214604.

[6] A. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe, "A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques," in *2019 1st International Informatics and Software Engineering Conference (UBMYK)*, Ankara, Turkey, Nov. 2019, pp. 1–4, https://doi.org/10.1109/UBMYK48245.2019.8965556.

[7] Y. Liu *et al.*, "Anthocyanins' effects on diabetes mellitus and islet transplantation," *Critical Reviews in Food Science and Nutrition*, vol. 63, no. 33, pp. 12102–12125, Dec. 2023, https://doi.org/10.1080/10408398.2022.2098464.

[8] B. Godi, S. Viswanadham, A. S. Muttipati, O. Prakash Samantray, and S. R. Gadiraju Student, "E-Healthcare Monitoring System using IoT with Machine Learning Approaches," in *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)*, Gunupur, India, Mar. 2020, pp. 1–5, https://doi.org/10.1109/ICCSEA49143.2020.9132937.

[9] P. Kaur, R. Kumar, and M. Kumar, "A healthcare monitoring system using random forest and internet of things (IoT)," *Multimedia Tools and*

Applications, vol. 78, no. 14, pp. 19905–19916, Jul. 2019, https://doi.org/10.1007/s11042-019-7327-8.

[10] V. R. Allugunti, C. Kishor Kumar Reddy, N. M. Elango, and P. R. Anisha, "Prediction of Diabetes Using Internet of Things (IoT) and Decision Trees: SLDPS," in *Intelligent Data Engineering and Analytics*, 2021, pp. 453–461, https://doi.org/10.1007/978-981-15-5679-1_43.

[11] N. Sharma and A. Singh, "Diabetes Detection and Prediction Using Machine Learning/IoT: A Survey," in *Advanced Informatics for Computing Research*, 2019, pp. 471–479, https://doi.org/10.1007/978-981-13-3140-4_42.

[12] M. Bhatia, S. Kaur, S. K. Sood, and V. Behal, "Internet of things-inspired healthcare system for urine-based diabetes prediction," *Artificial Intelligence in Medicine*, vol. 107, Jul. 2020, Art. no. 101913, https://doi.org/10.1016/j.artmed.2020.101913.

[13] A. Naseem, R. Habib, T. Naz, M. Atif, M. Arif, and S. Allaoua Chelloug, "Novel Internet of Things based approach toward diabetes prediction using deep learning models," *Frontiers in Public Health*, vol. 10, Aug. 2022, https://doi.org/10.3389/fpubh.2022.914106.

[14] R. Biswas, S. Pal, N. H. H. Cuong, and A. Chakrabarty, "A Novel IoT-Based Approach Towards Diabetes Prediction Using Big Data," in *Intelligent Computing in Engineering*, 2020, pp. 163–170, https://doi.org/10.1007/978-981-15-2780-7_20.

[15] P. Bide and A. Padalkar, "Survey on Diabetes Mellitus and incorporation of Big data, Machine Learning and IoT to mitigate it," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India, Mar. 2020, pp. 1–10, https://doi.org/10.1109/ICACCS48705.2020.9074202.

[16] R. Rastogi and M. Bansal, "Diabetes prediction model using data mining techniques," *Measurement: Sensors*, vol. 25, Feb. 2023, Art. no. 100605, https://doi.org/10.1016/j.measen.2022.100605.

[17] J. A. Abubakar, A. E. Odianose, and O. F. Ademola, "IoT-Enabled Machine Learning for Enhanced Diagnosis of Diabetes and Heart Disease in Resource-Limited Settings," in *Artificial Intelligence of Things for Achieving Sustainable Development Goals*, S. Misra, K. Siakas, and G. Lampropoulos, Eds. Springer Nature Switzerland, 2024, pp. 181–205.

[18] "Pima Indians Diabetes Database." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

[19] A. Teboul, "Diabetes Health Indicators Dataset." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset.

[20] M. Mustafa, "Diabetes prediction dataset." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset.