Type 2 Diabetes Mellitus: Early Detection using Machine Learning Classification

Gowthami S[1], Venkata Siva Reddy[2], Mohammed Riyaz Ahmed[3]
Research Scholar, School of Electronics and Communication Engineering, REVA University, Bangalore-560064[1]
School of Electronics and Communication Engineering, REVA University, Bangalore-560064[2,3]

***Abstract*—Type 2 Diabetes Mellitus (T2DM) is a growing global health problem that significantly impacts patient's quality of life and longevity. Early detection of T2DM is crucial in preventing or delaying the onset of its associated complications. This study aims to evaluate the use of machine learning algorithms for the early detection of T2DM. A classification model is developed using a dataset of patients diagnosed with T2DM and healthy controls, incorporating feature selection techniques. The model will be trained and tested on machine learning algorithms such as Logistic Regression, K-Nearest Neighbors, Decision Trees, Random Forest, and Support Vector Machines. The results showed that the Random Forest algorithm achieved the highest accuracy in detecting T2DM, with an accuracy of 98%. This high accuracy rate highlights the potential of machine learning algorithms in early T2DM detection and the importance of incorporating such methods in the clinical decision-making process. The findings of this study will contribute to the development of a more efficient precision medicine screening process for T2DM that can help healthcare providers detect the disease at its earliest stages, leading to improved patient outcomes.**

***Keywords*—Diabetes Mellitus Type II; feature selection; machine learning methods; precision medicine**

I. Introduction

The phenomenon of urbanization in recent years has brought about significant lifestyle changes, contributing to the rising incidence of diabetes. Diabetes, also known as Diabetes Mellitus (DM), occurs when the blood sugar levels become elevated. This condition can occur due to either inadequate production of insulin by the body or the cells' inability to respond to insulin. Insulin is a hormone that plays a crucial role in regulating glucose levels in the blood [1]. When the body fails to utilize glucose for energy production, it accumulates in the bloodstream, leading to a condition known as hyperglycemia.

The World Health Organization (WHO) has projected that by the year 2040, approximately 600 million people worldwide will be affected by diabetes. This staggering statistic underscores the urgent need to address the growing prevalence of this disease. The number of diagnosed cases continues to escalate [2], highlighting the pressing need for effective strategies to combat diabetes and mitigate its impact on global health.

Fig. 1 illustrates the significant impact of diabetes, where 90% to 95% of diabetes cases are attributed to T2D. Inadequate diabetes management affects glucose control and increases the risk of various comorbidities such as stroke, cancer,

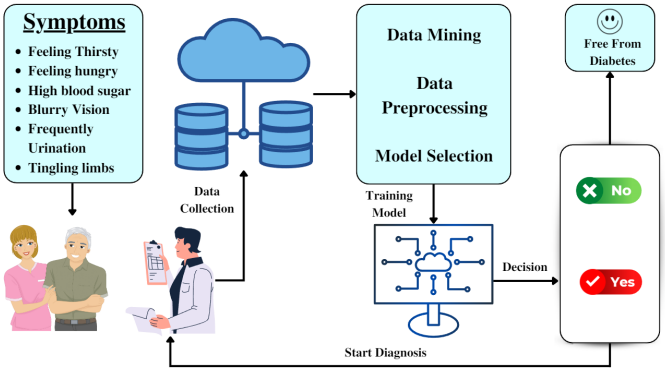


Fig. 1. Block diagram representation of precision medicine employing various data analytics and machine learning algorithms.

Alzheimer's disease, hypertension, and cardiovascular disease. Thus, early detection plays a crucial role in predicting and managing Diabetes Mellitus or T2D effectively.

Type 2 diabetes mellitus occurs when insulin is not produced sufficiently to meet the body's needs. On average, people are diagnosed to type 2 diabetes around the age of 40 due to Obesity, seniority, and parents' inheritance. Apart from Obesity, other factors such as age, gender, socioeconomic class, place of residency (rural or urban), smoking addiction, liquor infusion, nutrition frequency, and so on are strongly linked with Type 2 Diabetes (T2D) [3]. Few of these characteristics are flexible and thus play a vital role in T2D management.

Since the dawn of machine learning, many researchers have suggested classification models for the prediction of diabetes mellitus (Table I). However, even though machine learning algorithms can predict diabetes early, it is not yet established for clinical operations. The proposed method constitutes of various classification of ML algorithms and attributes are choosen from the dataset using feature selection algoritm. The selected attributes are fed into various ML algorithms and results are compared with existing models.

Diabetes data analysis is a difficult task since most of the medical information is non-linear, irregular, correlation structured, and very tangled in nature. Therefore, using machine learning techniques in T2D analysis is a critical process for extracting knowledge from large amounts of available diabetes-related data and also aids in accurately diagnosing diabetes [4]. Traditional blood analysis methods may be more painful

TABLE I. TIME LINE OF VARIOUS MACHINE LEARNING METHODS FOR T2DM

Ref	Year	Methodology	Independent Variables
[6]	1993	Static and Dynamic Regression Model	Gender, Age, Frequency of diabetes
[7]	2004	Multiple Regression Analysis	Age, Usage of cigarette, Chronic pancreas history, Physical activity
[8]	2010	Pearson Partial Correlation	Age, Hyper tension status, Race, Waist Circumference
[9]	2015	Wilcoxon Signed-Rank Regression Model	Age, Diabetic state, Over weight, BMI, Obesity
[10]	2016	General Regression Networks	Diastolic Blood pressure, Two hours serum insulin, Age, Pedigree Function, triceps subcutaneous thickness
[11]	2016	Artificial Neural Network	Age,Gender, Height, Weight, BMI,High Blood Pressure History, Pregnancy history, Gestational diabetes history
[12]	2017	Dynamic Markov Model	Type-1 and type-2 Diabetes Status, Death, Undiagnosis diabetic state
[13]	2022	Ensembler Random Forest	Gender, Age, Polyuria, Irritability, Genital Thrush

and time-consuming, and because of this, physicians frequently make decisions based on a patient's current blood analysis report. Therefore, it has become a challenging task for both physicians and patients the diagnosis of diabetes mellitus. Hence, monitoring blood glucose (BG) levels plays a vital role in avoiding and relieving diabetic complications.

A portative Self-Monitoring of Blood Glucose (SMBG) device, a combination of advanced Information and Communication Technology (ICT) and biosensors, nourishes an efficient real-time monitoring administration technique for the healthiness state of diabetic patients. As a result, a patient can independently monitor the differences in his blood glucose levels [5]. In addition, using CGM (continuous glucose monitoring) sensors, users can better understand changes in blood glucose levels. Thus, many of the intelligent diabetes diagnosis systems have been designed by inspiring human biological constructors. These advanced methods predict whether or not the patient has diabetes mellitus without taking into account of any surgeon's progress [14].

Further, the paper consists of a brief description of the different sections discussed in the paper. The subsequent sections include a comprehensive Literature Survey in the second section, which examines relevant research and existing knowledge in the field. Following the Literature Survey, the experimental setup outlines the methodology and techniques to develop the diabetes diagnosis model. Finally, the paper presents the Results and Discussion section, where the performance and effectiveness of the machine learning models are analyzed and interpreted in detail.

II. LITERATURE SURVEY

Early detection and diagnosis are paramount to effectively prevent the progression of diabetes. Several studies have found that lifestyle changes and pharmacological interventions can decrease the chance of developing diabetes. Furthermore, for recently interpreted intensive lifestyle, diabetic patients intervention, earlier short-term intensive insulin treatment, and metabolic therapy can result in prolonged glycaemic remission without additional antidiabetic treatment. As a result, identifying people at increased risk of acquiring diabetes is critical for diabetes prevention programs. These studies can be summarized as follows (Table II).

Type 2 diabetes mellitus (T2DM) is a chronic metabolic disorder, and recent studies have shown that machine learning techniques can accurately predict the early stages of the disease. For instance, Oladimeji et al. (2021) [15] work highlights the potential of combining feature selection and classification algorithms for predicting diabetes at an early stage, Bhavya et al. (2020) [16] achieved their goal of predicting the presence of diabetes using machine learning techniques and achieved high accuracy rates, while Tigga et al. (2021) [17] developed a logistic regression model to predict the occurrence of type 2 diabetes with an accuracy of 83.3%. Moreover, Liu et al. (2022) [18] used machine learning techniques to accurately predict the risk of incident type 2 diabetes mellitus in the Chinese elderly. Boshra et al. (2021) [19] diagnosed diabetes using machine learning techniques and achieved high accuracy. Butt et al. (2021) [20] highlight the benefits of using machine learning algorithms in the healthcare industry, such as improved accuracy, faster diagnosis, and lower costs. In another study, Barik et al. (2021) [21] analyzed the prediction and accuracy of diabetes using classifiers and hybrid machine learning techniques and found that hybrid techniques provided higher prediction accuracy. Mounika et al. (2021) [22] compared the performance of different machine learning algorithms. They achieved a high accuracy rate of 95.3% in predicting type-2 diabetes. In contrast, Sneha et al. (2019) [23] used techniques like entropy and correlation-based feature selection to predict diabetes mellitus early by selecting optimal features. However, limitations such as dataset size and diversity need to be addressed in these studies to improve the generalizability of their results.

Khallel et al. (2021) [24] applied machine learning algorithms such as a support vector machine, decision tree, and k-nearest neighbor to predict the incidence of type 2 diabetes in a Korean population with metabolic syndrome. Their results showed that machine learning models could detect individuals at high risk of developing type 2 diabetes. In another study, Hernández-Lemus et al. (2022) [25] proposed a new machine-learning approach using a joint embedding of DNA methylation data and clinical variables to predict T2DM status. The authors showed a higher predictive performance of their model compared to current state-of-the-art methods. Furthermore, the authors highlighted their approach's potential to identify key methylation signatures linked to T2DM. These latest studies have shown promising results and highlight the potential of machine learning techniques to enhance early detection and improve the management of T2DM. Recent studies that demonstrate the potential of machine learning in predicting diabetes include the work of Lu et al. (2021) [26], who developed a machine-learning model to predict prediabetes and T2DM using medical claim data from a large health maintenance organization. Their model achieved high sensitivity and specificity, indicating its potential to iden-

TABLE II. LITERATURE SURVEY

Authors	Algorithms Used	Summary	Limitations
Oladosu et.al 2021 [15]	Random Forest Naive Bayes J48 KNN	The author uses the feature selection method to prevent overfitting and remove redundant data, which helps in providing an optimal model for predicting diabetes in the earlier stage by emphasizing more on feature selection	The accuracy of the output can by increased by considering body size, height and BMI attributes, which lacks in this paper.
Bhavya et.al 2020 [16]	KNN	The authors extensively explore the popular techniques in Machine Learning and uses KNN algorithm to identify the diabetes	The paper lacks in verifying the different algorithms available in Machine Learning such as Navie Bayes, SVM, Decision Tree, ID3 etc.
Neha and Shruthi 2020 [17]	LR KNN, SVM Naive Bayes DT Random Forest	The author aims to evaluate the risk of diabetes among people based on their family background and lifestyle. In this paper the model is developed by choosing 952 instances collected through online and offline questioner. The authors conclude that RF is most accurate.	The authors failed to consider biological attributes which plays an important role in predicting diabetes.
Qing Liu et.al 2022 [18]	LR DT XGBoost RF	The authors aim to make effective prediction models using machine learning algorithms for risk type of T2DM in Chinese elderly. The proposed model uses Lasso Regression for feature selection and features are applied to ML algorithms, they got best accuracy in XGBoost algorithm.	The paper fails to implement on different algorithms which might have given more accuracy because they have choosen more feature selected attributes.
Boshra et.al 2021 [19]	Logistic Regression SVM Decision Tree XGBoost Random Forest ADABoost	The author aims to implement the diagnosis of diabetes employing Machine Learning algorithms and evaluate the performance comparison of diverse models for classifying diagnosis. According to this paper the ADABoost has more accuracy.	The accuracy can be increased if authors had considered more attributes.

tify high-risk individuals and improve preventive strategies. Similarly, Alqudah et al. (2022) [27] developed a machine-learning model utilizing retinal images to predict T2DM. The authors utilized deep learning techniques to predict the presence of T2DM. They found that their model outperformed existing methods, offering the potential for an inexpensive, non-invasive approach to diabetes screening.

Yilmaz et al. (2019) [28] investigated the effectiveness of machine learning models in predicting the risk of developing T2DM using non-invasive retinal imaging. The authors utilized deep learning techniques to analyze retinal images and predict the presence of T2DM. Their findings suggest that the analysis of retinal images may provide an effective and convenient screening tool for detecting early signs of T2DM with the potential for widespread implementation. Rasmy et al. (2021) [29] developed a machine-learning model for the early prediction of T2DM using electronic health record data from a large healthcare system. The authors utilized a combination of feature selection algorithms and machine learning models to predict the risk of T2DM and achieved high accuracy rates. Their findings suggest that machine learning-based approaches improve early detection and intervention for T2DM, ultimately preventing severe complications and reducing healthcare costs. In a recent study, Wang et al. (2022) [30] developed a machine-learning model for predicting incident T2DM using electronic health record data. The authors used a large dataset of over 2 million patients, and the model achieved high accuracy rates in predicting incident T2DM up to 36 months in advance. Their model may be incorporated into clinical decision support systems to aid in the early detection and prevention of T2DM.

A study by Huang et al. (2022) [31] developed and validated a machine learning-based nomogram for predicting the risk of T2DM in Chinese adults. The authors used data from the China Health and Retirement Longitudinal Study and developed a nomogram that integrated various risk factors such as age, sex, BMI, and lifestyle factors. The results showed that the nomogram achieved high accuracy in predicting T2DM risk and could be clinically applicable for the early detection and prevention of T2DM. Another study by Wu et al. (2023) [32] proposed a novel machine learning algorithm that combines deep neural networks (DNNs) and attention mechanisms to predict T2DM. The authors used electronic health record data from a large hospital in China and compared their model's performance to other widely used models. Their results showed that their DNN-attention model outperformed other models in predicting, indicating its potential for clinical application in early detection and personalized intervention. In a study by Lui et al. (2023) [33], the authors used a deep learning-based approach to predict T2DM using electronic health record data. The authors developed a convolutional neural network (CNN) model incorporating structured and unstructured data, such as lab test results and clinical notes. The model achieved high accuracy rates in T2DM prediction and showed improved performance compared to traditional machine learning models, suggesting its potential for clinical application in diabetes prediction and management. Finally, a recent study by Sim et al. (2023) [34] aimed to develop a machine-learning model that predicts T2DM based on lifestyle factors. The authors used data from a large cohort study and developed a model incorporating physical activity levels, diet, and sleep patterns. The results showed that the model achieved high accuracy in predicting T2DM risk based on lifestyle factors alone, demonstrating the potential of machine learning approaches to personalize diabetes prevention and management strategies based on lifestyle habits.

III. EXPERIMENTAL SET UP

Fig. 2 shows the proposed methodology of this work. The method demonstrated incorporates feature selection as an important step in the machine learning classification process. It helps to reduce the dataset's dimensionality and eliminate irrelevant or redundant features that can lead to overfitting or reduced accuracy. Further, the proposed model helps combine multiple feature selection techniques, which also helps to identify the most important features that contribute to the T2DM detection model and improve its overall accuracy. This is a crucial advantage over other methods that do not incorporate feature selection, leading to a more robust and accurate T2DM detection model.

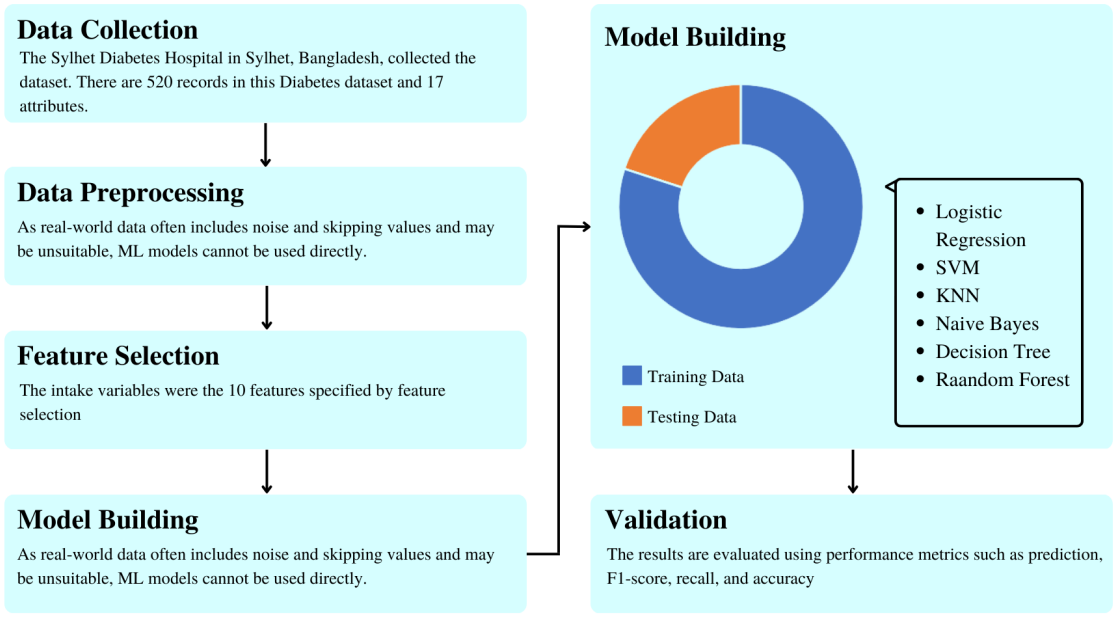


Fig. 2. Flowchart depicting the five stages of the machine learning process with data collection, data preprocessing, feature selection, and model building for validation.

The gathering and comprehension of data enable the examination of patterns and trends, which aids in outcome prediction and evaluation. The Sylhet Diabetes Hospital in Sylhet, Bangladesh, collects the dataset directly from the patients by having them complete a questionnaire, and a doctor then verifies it. Below is a description of the dataset. There are 520 records in this Diabetes dataset and 17 attributes.

As real-world data often includes noise and skipping values and may be unsuitable, ML models cannot be used directly. Instead, data cleaning and preparation for a machine learning sample are required for data pre-processing, which increases the precision and effectiveness of the machine learning model. Finding lost data, encoding categorical data, exploratory data analysis (Fig. 3), dividing the dataset into a training and trial set, feature selection, and feature scaling are some of the tasks it entails.

Every machine learning pipeline comprises Key Performance Indicators (KPIs). Since the result from classification models are discrete, we require a metric that contrasts different classes somehow. Classification Metrics evaluate a model's execution and show if the classification is valid, but they do so in various ways. It evaluates models based on their performance, time complexity, accuracy, recall, sensitivity, and the most promising metric, which is the confusion matrix. The truth labels versus model predictions are displayed in a confusion matrix using tables. An individual row of the confusion matrix defines an instance in a class label, but an individual column represents an instance in a real class. The Confusion Matrix isn't a performance indicator but provides the framework for other metrics to assess the outcomes. Therefore, a metric is required for all machine learning models to evaluate their effectiveness.

The dataset was randomly split into train and trial data with a proportion of 80% to 20%. Since the data is extremely unbalanced, it is a good idea to use the Sklearn module in Python to standardize the dataset's input features and normalize. Using the Python Standard Scaler function from the Sklearn pre-processing library, the training cluster was standardized to have a mean of Zero and a variance of One, and the test set was normalized using the mean and standard deviation of the training dataset.

We trained the LR, DT, RF, KNN, and SVM models and implemented them using the Python Sklearn package. The intake variables were the 10 features specified by feature selection (Table III). The Primary purpose of using hyperparameters is to specify the attributes before training begins. It expresses the major characteristics of the model, such as its efficiency, robustness, and intricacy. Therefore, it has evolved to become immensely popular for adjusting hyperparameters in machine-learning algorithms. The most suitable hyperparameters for RF were as observed: max_depth = 100, max_features = 10, min_samples_leaf = 5, n_estimators = 80, min_samples_split = 69, criterion = entropy and random state = 0.

TABLE III. ATTRIBUTES

SL No.	Attributes	Score
1	Age	18.845767
2	Gender	38.747637
3	Polyuria	116.184593
4	Polydipsia	120.785515
5	sudden weight loss	57.749309
6	Polyphagia	33.198418
7	visual blurring	18.124571
8	Irritability	35.334127
9	partial paresis	55.314286
10	AgeAlopecia	24.402793

Though AI and its subsets like machine learning are quite

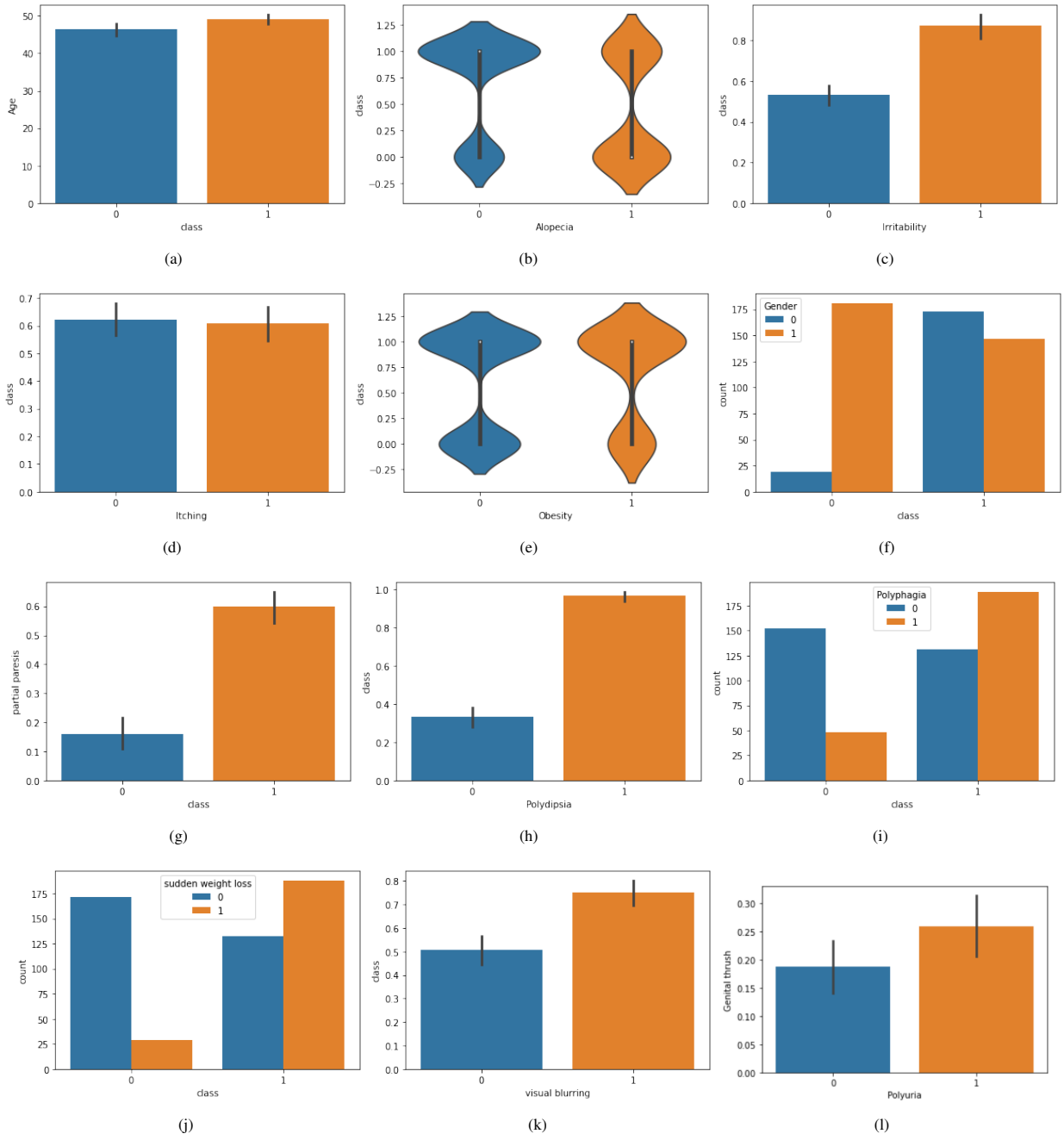
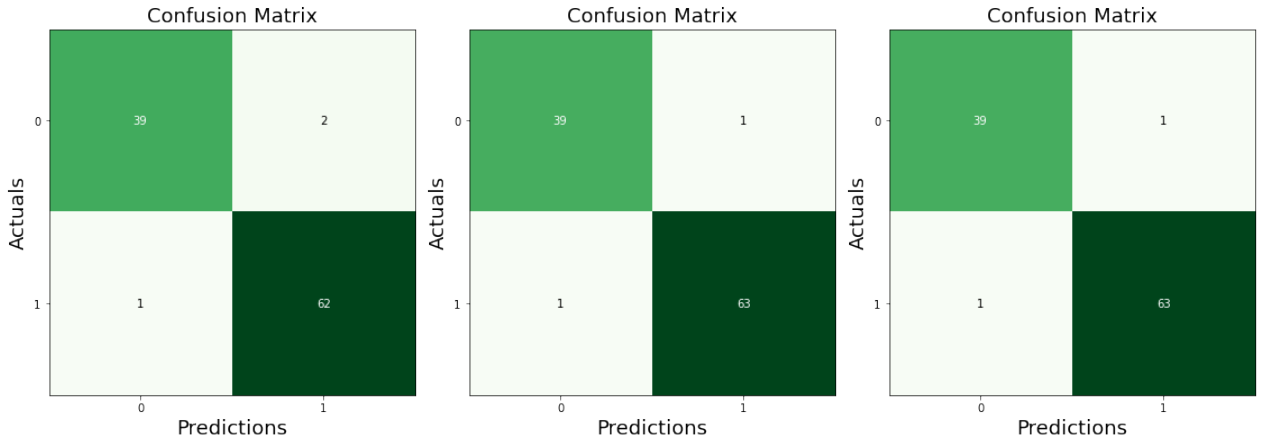


Fig. 3. Exploratory data analysis for the extracted features from the dataset - a)age-class, b)alopecia, c)irritability, d)itching, e)gender, f)obesity, g)partial paresis, h)polydipsia, i)polyphagia, j)sudden weight loss, k)visual blurring, l)ploryuria Vs gential thursh.

promising with their results, it is still a challenge to practice these methods in real-time for clinical diagnosis. The main reason is that AI needs a large amount of data to work with, and this requires time. In the case of AI, this is called data transfer learning. And it means that you need a large amount of labeled data to train your algorithm with. Moreover, the results may again vary based on the data collected from a geographic location such as Asia Pacific, USA, and European regions. The datasets available on open-source platforms are outdated and don't meet the expected results as the prediction needs to be dynamic, which means based on the patient's current reports. Also, the results will vary based on the top features or attributes selected during the process of feature selection. The challenge of building a machine learning model is that it requires extensive data, which is often not available in the medical field. Therefore, there are limitations to building a model based on the available data.

IV. Results and Discussion

Our research presents compelling results on the detection of Type 2 Diabetes (T2D) using feature selection algorithms.



(a) (b) (c)

Fig. 4. Confusion matrix depicting true positive, true negative, false positive and false negative to calculate precision, recall, F1-score and accuracy for - a)decision tree, b)support vector machine, c)random forest.

Through the utilization of carefully curated datasets, we employed various algorithms to assess the probability of individuals having T2D. The evaluation of our results was based on performance metrics such as prediction, F1-score, recall, and accuracy, which are presented in Table IV.

Fig. 4 highlights the notable findings of our study, emphasizing the significant improvement in T2D prediction accuracy achieved by addressing imbalanced datasets using feature selection techniques. The Confusion Matrix visually represents the superior performance of our approach, clearly demonstrating the enhanced accuracy obtained through our methodology.

Of particular interest is the remarkable performance of the Random Forest (RF) algorithm, outperforming other algorithms examined in our study. This discovery holds great promise for the healthcare industry, as our proposed model proves to be cost-effective and efficient in terms of implementation time. Unlike traditional diagnostic tests such as Oral Glucose Tolerance Test (OGTT) and Hemoglobin A1c (HbA1c), our model eliminates the need for expensive lab reagents and specialized skills. This breakthrough paves the way for early detection of diabetes, especially in remote areas with limited access to healthcare facilities.

Looking ahead, the integration of 5G and Artificial Intelligence (AI) technologies offers even more potential for improving healthcare outcomes. By leveraging deep learning algorithms, the latency at the edge can be significantly reduced, enabling real-time data analysis and decision-making. This cognitive system architecture represents the inception of a smart healthcare system, empowering remote areas and underserved communities with access to timely and accurate diabetes detection.

Moreover, the potential impact of our research extends beyond the realm of T2D detection. The application of feature selection algorithms and the success of the RF algorithm in this study open avenues for exploring similar approaches in the diagnosis and prediction of other medical conditions. By leveraging the power of data analysis and machine learning, we can unlock new insights into various diseases and develop efficient and accessible diagnostic models. This not only has the potential to transform healthcare delivery but also contributes to the advancement of personalized medicine, where early detection and targeted interventions can significantly improve patient outcomes and overall population health. Our research lays a solid foundation for future investigations into leveraging feature selection and algorithmic techniques to enhance medical diagnostics across diverse healthcare domains.

TABLE IV. THE ML ALGORITHMS WITH THE KPI AFTER BEING TRAINED WITH THE PROPOSED MODEL

SL No.	Algorithms	Precision	Recall	F1-score	Accuracy
1	Logistic Regression	92%	91%	91%	89%
2	SVM Linear	94%	91%	92%	90%
3	SVM Radial	98%	98%	98%	98%
4	KNN	98%	98%	98%	98%
5	Naive bayes	89%	88%	88%	86%
6	Decision Tree	97%	97%	97%	97%
7	Random Forest	98%	98%	98%	98%

The dataset used in this research was obtained from the Sylhet Diabetes Hospital in Bangladesh. It consists of 520 records with 17 attributes. The data was collected through patient questionnaires, verified by medical professionals. However, the dataset's regional specificity may limit generalizability. The sample size, though substantial, should be considered for its representativeness. Potential limitations include self-reporting bias and missing variables. Despite these considerations, the dataset provides a valuable foundation for studying the relationship between urbanization and diabetes.

In summary, our results demonstrate the effectiveness of feature selection algorithms, with the RF algorithm standing out as a powerful tool for T2D prediction. The cost-effectiveness, efficiency, and accessibility of our proposed model present exciting opportunities for revolutionizing healthcare and fostering the development of smart healthcare sys-

tems.

V. Conclusion

Our research on early detection of Type 2 Diabetes Mellitus (T2DM) using machine learning classification has yielded significant findings. By employing feature selection techniques and effective data management, we have developed an innovative model for diagnosing diabetes at its early stages. The combination of various feature selection methods with machine learning algorithms has proven to be highly effective in predicting the presence of T2DM, with the Random Forest algorithm achieving an impressive accuracy of 98%. This research highlights the importance of early identification of diabetes due to its widespread prevalence and impact on individuals' health. Furthermore, the application of machine learning classification in the healthcare sector not only facilitates more efficient decision-making but also lowers the cost of diagnosis, making it more accessible to a broader population. Our contribution lies in the significance of feature selection and appropriate data management techniques, which greatly improve the predictive power of the model. However, further validation is necessary through larger sample sizes and diverse populations to ensure the model's robustness and generalizability. Future research should explore the impact of variables such as body size, height, and body mass index (BMI) in the early identification of diabetes, potentially enhancing the accuracy of the diagnostic model. Continued advancements in machine learning algorithms and refinement of the feature selection process can further enhance the model's performance and applicability. Overall, our study demonstrates the promising potential of machine learning classification with feature selection in early T2DM detection, aiding in better management and prevention of this prevalent disease.

Acknowledgment

The authors acknowledge the support from REVA University for the facilities provided to carry out the research.

References

[1] Pati, A., Parhi, M., & Pattanayak, B. K. (2023). A review on prediction of diabetes using machine learning and data mining classification techniques. International Journal of Biomedical Engineering and Technology, 41(1), 83-109.

[2] Qi, H., Song, X., Liu, S., Zhang, Y., & Wong, K. K. (2023). KFPredict: An ensemble learning prediction framework for diabetes based on fusion of key features. Computer Methods and Programs in Biomedicine, 107378.

[3] Deepthi, Y., Kalyan, K. P., Vyas, M., Radhika, K., Babu, D. K., & Krishna Rao, N. V. (2020). Disease prediction based on symptoms using machine learning. In Energy Systems, Drives and Automations: Proceedings of ESDA 2019 (pp. 561-569). Singapore: Springer Singapore.

[4] Ma, J. (2020, October). Machine learning in predicting diabetes in the early stage. In 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI) (pp. 167-172). IEEE.

[5] Hebbale, A., Vinay, G. H. R., Krishna, B. V., & Shah, J. (2021, October). IoT and Machine Learning based Self Care System for Diabetes Monitoring and Prediction. In 2021 2nd Global Conference for Advancement in Technology (GCAT) (pp. 1-7). IEEE.

[6] Ruwaard, D., Hoogenveen, R. T., Verkleij, H., Kromhout, D., Casparie, A. F., & Van der Veen, E. A. (1993). Forecasting the number of diabetic patients in The Netherlands in 2005. American journal of public health, 83(7), 989-995.

[7] Rosenthal, A. D., Jin, F., Shu, X. O., Yang, G., Elasy, T. A., Chow, W. H., ... & Zheng, W. (2004). Body fat distribution and risk of diabetes among Chinese women. International journal of obesity, 28(4), 594-599.

[8] Holman, N., Forouhi, N. G., Goyder, E., & Wild, S. H. (2011). The Association of Public Health Observatories (APHO) diabetes prevalence model: estimates of total diabetes prevalence for England, 2010–2030. Diabetic Medicine, 28(5), 575-582.

[9] Nanri, A., Nakagawa, T., Kuwahara, K., Yamamoto, S., Honda, T., Okazaki, H., ... & Japan Epidemiology Collaboration on Occupational Health Study Group. (2015). Development of risk score for predicting 3-year incidence of type 2 diabetes: Japan epidemiology collaboration on occupational health study. PLoS One, 10(11), e0142779.

[10] Alby, S., & Shivakumar, B. L. (2016). A prediction model for type 2 diabetes risk among Indian women. ARPN Journal of Engineering and Applied Sciences, 11(3), 2037-2043.

[11] Chen, L. S., & Cai, S. J. (2015). Neural-network-based resampling method for detecting diabetes mellitus. Journal of Medical and Biological Engineering, 35, 824-832.

[12] Saidi, O., O'Flaherty, M., Mansour, N. B., Aissi, W., Lassoued, O., Capewell, S., ... & EC FP7 funded MEDCHAMPS project. (2015). Forecasting Tunisian type 2 diabetes prevalence to 2027: validation of a simple model. BMC public health, 15, 1-8.

[13] Nagaraj, P., Muneeswaran, V., & Deshik, G. (2022, August). Ensemble Machine Learning (Grid Search & Random Forest) based Enhanced Medical Expert Recommendation System for Diabetes Mellitus Prediction. In 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 757-765). IEEE.

[14] van Doorn, W. P., Foreman, Y. D., Schaper, N. C., Savelberg, H. H., Koster, A., van der Kallen, C. J., ... & Brouwers, M. C. (2021). Machine learning-based glucose prediction with use of continuous glucose and physical activity monitoring data: The Maastricht Study. PloS one, 16(6), e0253125.

[15] Oladimeji, O. O., Oladimeji, A., & Oladimeji, O. Classification models for likelihood prediction of diabetes at early stage using feature selection. Applied Computing and Informatics, 2021.

[16] Bhavya, M. R., & Rao, S. Diabetes Prediction using Machine Learning. International Journal of Advanced Research in Computer and Communication Engineering, 9(7), 2020.

[17] Tigga, N. P., & Garg, S. Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science, 167, 706-716, 2020.

[18] Liu, Q., Zhang, M., He, Y., Zhang, L., Zou, J., Yan, Y., & Guo, Y. Predicting the Risk of Incident Type 2 Diabetes Mellitus in Chinese Elderly Using Machine Learning Techniques. Journal of Personalized Medicine, 12(6), 905,2022.

[19] Farajollahi, B., Mehmannavaz, M., Mehrjoo, H., Moghbeli, F., & Sayadi, M. J. Diabetes diagnosis using machine learning. Frontiers in Health Informatics, 10(1), 65, 2021.

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

[21] Barik, S., Mohanty, S., Mohanty, S., & Singh, D. (2021). Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques. In Intelligent and Cloud Computing: Proceedings of ICICC 2019, Volume 2 (pp. 399-409). Springer Singapore.

[22] Mounika, V., Neeli, D. S., Sree, G. S., Mourya, P., & Babu, M. A. (2021, March). Prediction of type-2 diabetes using machine learning algorithms. In 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) (pp. 127-131). IEEE.

[23] Sneha, N., & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big data, 6(1), 1-19.

[24] Khaleel, F. A., & Al-Bakry, A. M. (2023). Diagnosis of diabetes using machine learning algorithms. Materials Today: Proceedings, 80, 3200-3203.

[25] Martínez-García, M., & Hernández-Lemus, E. (2022). Data integration challenges for machine learning in precision medicine. Frontiers in medicine, 8, 3082.

[26] Lu, H., Uddin, S., Hajati, F., Moni, M. A., & Khushi, M. (2022). A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus. Applied Intelligence, 52(3), 2411-2422.

[27] Alqudah, A. M., & Alqudah, A. (2022). Improving machine learning recognition of colorectal cancer using 3D GLCM applied to different color spaces. Multimedia Tools and Applications, 81(8), 10839-10860.

[28] Yilmaz, T., Foster, R., & Hao, Y. (2019). Radio-frequency and microwave techniques for non-invasive measurement of blood glucose levels. Diagnostics, 9(1), 6.

[29] Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ digital medicine, 4(1), 86.

[30] Dong, Z., Wang, Q., Ke, Y., Zhang, W., Hong, Q., Liu, C., ... & Chen, X. (2022). Prediction of 3-year risk of diabetic kidney disease using machine learning based on electronic medical records. Journal of Translational Medicine, 20(1), 1-10.

[31] Guo, C., Ye, Y., Yuan, Y., Wong, Y. L., Li, X., Huang, Y., ... & Chen, H. (2022). Development and validation of a novel nomogram for predicting the occurrence of myopia in schoolchildren: A prospective cohort study. American Journal of Ophthalmology, 242, 96-106.

[32] Wu, J., Fang, Y., Tan, X., Kang, S., Yue, X., Rao, Y., ... & Yap, P. T. (2023). Detecting type 2 diabetes mellitus cognitive impairment using whole-brain functional connectivity. Scientific Reports, 13(1), 3940.

[33] Lui, G., Leung, H. S., Lee, J., Wong, C. K., Li, X., Ho, M., ... & Zee, B. (2023). An efficient approach to estimate the risk of coronary artery disease for people living with HIV using machine-learning-based retinal image analysis. Plos one, 18(2), e0281701.

[34] Sim, R., Chong, C. W., Loganadan, N. K., Adam, N. L., Hussein, Z., & Lee, S. W. H. (2023). Comparison of a chronic kidney disease predictive model for type 2 diabetes mellitus in Malaysia using Cox regression versus machine learning approach. Clinical kidney journal, 16(3), 549-559.