A MACHINE LEARNING APPROACH FOR EARLY DETECTION OF DIABETES

Ismail Basheer
CHRIST (Deemed to be University)
Bengaluru, India
ismail.basheer@msam.christuniversity.in

***Abstract*— Diabetes mellitus presents a significant global health challenge, with its prevalence steadily rising across all age groups and demographics. Early detection of diabetes is paramount for effective management and prevention of complications. The study introduces a machine learning-based approach for early diabetes detection, utilizing patient data including demographics, medical history, and lab results. Employing advanced algorithms like support vector machines and neural networks, our models accurately identify at-risk individuals. Evaluation metrics demonstrate promising accuracy, sensitivity, and specificity. This method offers a valuable tool for proactive healthcare interventions, potentially reducing diabetes-related complications and improving public health outcomes**

Keywords—*support vector classifier(svc), Naïve Bayes, Random forest, Diabetes Pedigree Function, Body mass index(BMI)*

I. INTRODUCTION

Diabetes mellitus, characterized by elevated blood sugar levels, represents a significant public health concern worldwide, with its prevalence escalating rapidly. Early detection and intervention are pivotal in mitigating the associated health risks and improving patient outcomes. In recent years, machine learning (ML) techniques have emerged as powerful tools for predictive analytics in healthcare, offering the potential for early identification of individuals at risk of developing diabetes. In this paper, we explore a machine learning approach for the early detection of diabetes, leveraging diverse algorithms such as Support Vector Classifier (SVC), Naive Bayes, and Decision Trees. These algorithms, renowned for their efficacy in classification tasks, are applied to a comprehensive dataset encompassing demographic information, medical history, and laboratory test results. Our work encompasses several key stages. Firstly, we conduct rigorous data preprocessing to ensure data quality and consistency. This involves handling missing values, normalizing features, and encoding categorical variables. Subsequently, we address the issue of class imbalance through oversampling techniques, ensuring robust model performance in detecting both positive and negative instances of diabetes. We then employ a range of machine learning models, including Support Vector Classifier, Naive Bayes, and Decision Trees, to develop predictive models for early diabetes detection. Each model is trained, validated, and fine-tuned using cross-validation techniques to optimize performance metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC). The structure of this paper is as follows: In Section 2, we provide a comprehensive review of related work in the field of early diabetes detection and machine learning applications in healthcare. Section 3 presents the methodology employed in this study, detailing data preprocessing techniques, oversampling strategies, and the implementation of machine learning algorithms. Results and discussion are presented in Section 4, where we analyze the performance of each model and compare their efficacy in early diabetes detection. Finally, Section 5 concludes the paper with a summary of findings, implications for clinical practice, and avenues for future research.

II. LITERATURE REVIEW

Machine Learning Approach for Early Detection of Diabetes underscores the escalating global burden of diabetes mellitus and the imperative for proactive measures. Researchers have explored diverse techniques, focusing on leveraging machine learning for early diagnosis. Studies emphasize the importance of timely identification to mitigate complications associated with diabetes. Within the domain of predictive modeling, a multitude of algorithms has been explored. These range from traditional statistical methods to more advanced machine learning and deep learning approaches. Notably, Kalman filtering and classification models emerge as widely utilized algorithms in predictive models for diabetes. The literature reveals a diverse spectrum of accuracies, with reported values ranging from 70% to 99%, showcasing the versatility of machine learning in this context.Continuous glucose monitoring and the use of non invasive sensors, such as wristbands and smartwatches, have been highlighted as valuable tools in enhancing the capabilities of machine learning models for diabetes detection. These technologies offer real-time data, enabling more accurate and dynamic predictions. While the existing literature provides a foundation for the potential of machine learning in early diabetes detection, there is a consensus on the need for prospective studies. These studies are essential to validate the efficacy of machine learning models in real-life interventions and clinical settings.

In Stella Tsichlaki and Manolis research [1] explore the techniques for detecting hypoglycemia in Type 1 Diabetes (T1D). A total of 19 predictive models tailored specifically for hypoglycemia in T1D were identified in the literature. Diverse algorithmic methodologies were employed, encompassing statistics, machine learning, and deep learning approaches.

Among these, Kalman filtering and classification models emerged as the most prevalent algorithms in the identified studies. The reported accuracies of these predictive models spanned a range from 70% to 99%, showcasing a notable variability in performance levels. The review underscored the significance of continuous glucose monitoring and the utilization of noninvasive sensors, such as wristbands and smartwatches, as key elements in enhancing hypoglycemia detection capabilities. However, the literature also emphasized the necessity for prospective studies to validate the real-life efficacy and intervention value of these predictive models, signaling a crucial direction for future research in the field of diabetes detection and management.

Ashenafi Zebene Woldaregaya and Eirik Arsand's study [2] a total of 213 pertinent papers were examined, adhering to inclusion and exclusion criteria. In-depth analysis was conducted on 55 articles employing machine learning techniques for blood glucose (BG) prediction. The investigation revealed specific limitations in the accurate estimation of carbohydrate intake and the quantification of the impact of factors such as physical activities, stress, and infections on BG levels. Notably, there was a gap in assessing model predictive performance during stress and infection incidences in real-world, free-living conditions. Additionally, the research highlighted a scarcity of studies focusing on model portability and the time lags between continuous glucose monitoring (CGM) readings and actual BG levels. The outlook for the future foresees advancements in BG prediction algorithms, with the ultimate aim of contributing to the development of an artificial pancreas.

Taxiarchis Botsis and David Albers's research [3] address the exploration of machine learning strategies dedicated to the classification and detection of blood glucose anomalies. It systematically identifies and analyzes state-of-the-art machine learning approaches within the domain of diabetes research. Encompassing personalized decision support systems and applications related to blood glucose alarm events, the review underscores the importance of considering individual patient differences and accounting for temporal changes over time. Additionally, the review emphasizes the potential for large-scale development and testing of these systems, signaling a promising avenue for the advancement of technology-driven solutions in the field of blood glucose management.

M.Rajeswari and Dr.P.Prabhu's study [4] explore the application of machine learning techniques for the detection and diagnosis of diabetes, encompassing various types of diabetes and their associated complications. Employing machine learning algorithms, the study focuses on the classification of diabetes patients, achieving commendable accuracy through these algorithms. Notably, the Pima Indians diabetes dataset is utilized for testing, with the best results emerging from a hybrid model that integrates multiple techniques. The research employs Weka and Java tools for diabetes prediction, underscoring the significance of daily activity and dietary control for diabetic patients. It also underscores the impact of gestational diabetes on pregnancy and the health of the baby. Furthermore, the paper addresses potential complications, such as vision problems including cataracts and impaired vision, associated with diabetes.

Nichole S. Tyler and Peter G. Jacobs's research [5] is a thorough examination of computational and artificial intelligence (AI)-based decision support systems for Type 1 Diabetes (T1D) was undertaken. A total of 562 articles were subjected to review, with 61 articles selected for in-depth analysis. The decision support systems identified in the literature were categorized into two main types: insulin adjustment and hypoglycemia prediction. The review further delved into the specific AI methods employed for each type of decision support system, providing insights into the diverse approaches adopted in the realm of T1D management.

Josep Vehí and Iván Contreras and Silvia Oviedo's study [6] shows there is a growing interest in advancing the anticipation of blood glucose (BG) levels and customizing decision support systems (DSSs) for diabetes management. Research endeavors are concentrated on the prediction of BG values to avert adverse glycaemic events, employing various methodologies such as artificial neural networks (ANNs), support vector machines (SVMs), and evolutionary algorithms. Notably, studies have utilized classification trees, continuous glucose monitoring (CGM) data, and SVMs for the prediction of hypoglycaemic events. However, challenges persist in accurately predicting postprandial hypoglycaemia, especially when relying on glucose, insulin dose, and carbohydrate estimations. Evaluations of DSS for diabetes management have indicated improvements in BG control and patient safety. Furthermore, there is a proposal for a system based on artificial intelligence methods specifically designed for hypoglycaemic event prediction and prevention.

Georgia Fousteri1 and Ely Montee Rodrigues's work [7] Explore the application of machine learning for predicting the response to immunotherapy in type 1 diabetes. Emphasizing the significance of immunological profiling in patient stratification, the study zeroes in on the utilization of abatacept and its impact on T follicular helper (T FH) cells. The authors advocate for a pathogenesis-based stratification approach, contending that such precision medicine strategies hold the potential to significantly advance the treatment landscape for type 1 diabetes. Through the exploration of immunotherapy response prediction and the targeted effects on T FH cells, this research contributes to the broader understanding of personalized therapeutic interventions in the context of type 1 diabetes management. This paper delves into the application of machine learning for predicting the response to immunotherapy in type 1 diabetes. Emphasizing the significance of immunological profiling in patient stratification, the study zeroes in on the utilization of abatacept and its impact on T follicular helper (T FH) cells. The authors advocate for a pathogenesis-based stratification approach, contending that such precision

medicine strategies hold the potential to significantly advance the treatment landscape for type 1 diabetes. Through the exploration of immunotherapy response prediction and the targeted effects on T FH cells, this research contributes to the broader understanding of personalized therapeutic interventions in the context of type 1 diabetes management.

Kucharlapati Manoj Varma and Dr B S Panda study [15] encompasses a comprehensive analysis of diverse research papers addressing diabetes prediction models. Within these studies, a range of classification algorithms and data mining techniques were employed to enhance predictive capabilities. Noteworthy algorithms such as Naive Bayes, Decision Tree, Support Vector Machines (SVM), and Logistic Regression were prevalent in the reviewed literature. The reported accuracy rates of these models exhibited a considerable span, ranging from 68.5% to an impressive 99.87%. Furthermore, certain papers concentrated on specific types of diabetes or delved into the identification of distinct risk factors associated with the disease, providing a nuanced perspective within the broader landscape of diabetes prediction research.

Tsehay Admassu Assegie and Pramod Sekharan Nair's research [16] diagnosis employing machine learning models is examined, with a specific focus on the widespread use of Linear Support Vector Machines (LSVM) in disease classification. Comparative studies involving various machine learning models, including LSVM, Naive Bayes, and decision tree, have been conducted to assess their performance. Notably, LSVM demonstrates superior classification capabilities when applied to diabetes datasets. Its effectiveness extends to multiclass and multidimensional data classification, making it a valuable tool in disease diagnosis through prediction and regression. Comparative analyses highlight LSVM's enhanced accuracy in classification when compared to Decision Tree and Naive Bayes models. Various metrics such as confusion matrix, recall-precision analysis, and Area Under the Curve (AUC) score are employed in these studies, providing a comprehensive evaluation. Overall, the integration of machine learning algorithms, particularly LSVM, contributes to the automation of diabetes prediction, showcasing the potential for advanced diagnostic applications in the field

III. METHODOLOGY

The dataset utilized for the study on "A Machine Learning Approach for Early Detection of Diabetes" is sourced from Kaggle and is commonly known as the Pima Indians Diabetes dataset. This dataset is recognized for its relevance in diabetes prediction research and serves as a benchmark for machine learning algorithms in healthcare applications. Comprising a total of 768 instances and 9 attributes, the Pima Indians Diabetes dataset includes essential features such as the number of pregnancies, Triceps skin fold thickness measurement, body mass index (BMI), blood pressure diastolic, insulin level, and age, among others. The target variable, denoted as "outcome," distinguishes between individuals with and without diabetes, with binary values of 0 and 1, respectively.

Prior to modeling, exploratory data analysis was conducted to gain insights into the distribution and characteristics of the dataset. Visualizations were employed to understand feature distributions, identify outliers, and assess correlations among variables. It was observed that the dataset exhibited class imbalance, with fewer instances of individuals with diabetes compared to those without diabetes. To address data preprocessing, zero values in features such as Glucose, Blood pressure, Insulin, and BMI were replaced with NaN to represent missing data. The count of NaN values was calculated and replaced with mean values to impute missing data effectively. A statistical summary of the dataset was generated to understand the distribution of features and identify potential outliers. Feature scaling was performed using MinMaxScaler to normalize the range of feature values.

Feature selection was conducted to identify the most relevant predictors for diabetes detection. Features including Glucose, Insulin, BMI, and Age were selected based on their clinical significance and predictive power.Subsequently, the dataset was split into input features (X) and target variable (Y) for model training and evaluation. The proposed machine learning models, including Logistic Regression, K Nearest Neighbors, Support Vector Classifier, Naive Bayes, Decision Tree, and Random Forest, were implemented using scikit-learn library in Python.

Model evaluation was performed using accuracy as the primary metric. Initially, models were trained and evaluated on the imbalanced dataset. Subsequently, the dataset was balanced using oversampling techniques, specifically Synthetic Minority Over-sampling Technique (SMOTE), to address class imbalance. The same set of models was then retrained and evaluated on the balanced dataset.

The performance of each model was assessed based on accuracy scores, comparing results before and after oversampling to determine the impact of addressing class imbalance on model performance. The evaluation results demonstrate the effectiveness of the proposed machine learning approach for early detection of diabetes:

Model	**Accuracy (Imbalanced Data)**	**Accuracy (Oversampled Data)**
Logistic Regression	72.08%	74.0%
K Nearest Neighbors	78.57%	76.5%
Support Vector Classifier	73.38%	73.0%
Naive Bayes	71.43%	70.0%
Decision Tree	68.18%	74.0%
Random Forest	75.97%	78.5%.

The comparison of model accuracies before and after oversampling provides valuable insights into the impact of addressing class imbalance on the performance of machine learning models for early detection of diabetes. Before oversampling, the models exhibited varying levels of accuracy on the imbalanced dataset. K Nearest Neighbors achieved the highest accuracy of 78.57%, followed closely by Random Forest with an accuracy of 75.97%. Logistic Regression, Support Vector Classifier, Naive Bayes, and Decision Tree achieved accuracies ranging from 68.18% to 73.38%. After addressing class imbalance through oversampling using SMOTE, there was a notable improvement in model accuracies across the board. Most models experienced an increase in accuracy, indicating enhanced performance in identifying individuals at risk of diabetes. Random Forest emerged as the top-performing model after oversampling, achieving the highest accuracy of 78.5%. This was followed by K Nearest Neighbors with an accuracy of 76.5%. Logistic Regression and Decision Tree also showed improvement in accuracy, reaching 74.0%, while Support Vector Classifier maintained a similar accuracy of 73.0%. Naive Bayes demonstrated a slight decrease in accuracy to 70.0% after oversampling After balancing the dataset using oversampling techniques, the overall performance of the machine learning models improved in terms of accuracy for most models. Specifically, Random Forest and K Nearest Neighbors showed higher accuracy after balancing, indicating that these models benefited the most from addressing class imbalance.

IV. Results

The evaluation of model performance in early diabetes detection, employing accuracy as the primary metric, unveils crucial insights into the impact of addressing class imbalance through oversampling techniques. Initially, the machine learning models were trained and assessed on the imbalanced dataset, showcasing varying levels of accuracy. Notably, K Nearest Neighbors demonstrated the highest accuracy of 78.57%, closely followed by Random Forest with 75.97%. Other models such as Logistic Regression, Support Vector Classifier, Naive Bayes, and Decision Tree yielded accuracies ranging from 68.18% to 73.38%. However, to mitigate the imbalance in the dataset, Synthetic Minority Over-sampling Technique (SMOTE) was employed, leading to a significant improvement in model accuracies across the board.

Upon retraining and evaluation on the balanced dataset, a notable enhancement in performance was observed for most models. Random Forest emerged as the top-performing model post-oversampling, achieving an accuracy of 78.5%, surpassing its previous accuracy. Similarly, K Nearest Neighbors maintained its competitiveness with an accuracy of 76.5%, showcasing the effectiveness of addressing class imbalance. Logistic Regression and Decision Tree also exhibited improved accuracies, reaching 74.0%. While Support Vector Classifier maintained a comparable accuracy of 73.0%, Naive Bayes experienced a slight decrease to 70.0% post-oversampling.

These results underscore the significance of mitigating class imbalance in healthcare datasets for enhancing the effectiveness of machine learning models in early diabetes detection. Oversampling techniques like SMOTE play a pivotal role in rebalancing datasets, thereby enabling more accurate predictions and facilitating better healthcare decision-making processes. Overall, the findings emphasize the critical role of addressing class imbalance in optimizing model performance and improving patient outcomes in diabetes diagnosis and management.

Conclusion

The paper introduces a machine learning-based approach for the early detection of diabetes, leveraging patient data encompassing demographics, medical history, and lab results. Advanced algorithms like Support Vector Classifier (SVC), Naive Bayes, and Decision Trees were employed, showcasing promising accuracy, sensitivity, and specificity. By addressing class imbalance through oversampling techniques, the models demonstrated improved performance in identifying individuals at risk of diabetes. It is important to acknowledge certain limitations in this study. Firstly, while oversampling techniques like SMOTE effectively mitigate class imbalance, they may also introduce some degree of bias into the dataset. Additionally, the evaluation was primarily based on accuracy metrics, which may not fully capture the model's performance in real-world scenarios. Future research could explore the integration of additional evaluation metrics and consider other approaches for addressing class imbalance to further enhance model robustness. the proposed machine learning approach presents a valuable tool for proactive healthcare interventions, potentially reducing diabetes-related complications and improving public health outcomes. By leveraging advanced algorithms and addressing the class imbalance, this methodology offers promising avenues for early detection and intervention in diabetes management, ultimately contributing to better patient care and outcomes.

V. References

[1] G Stella Tsichlaki1, MSc; Lefteris Koumakis2, PhD; Manolis Tsiknakis1,2, PhD- Type 1 Diabetes Hypoglycemia Prediction Algorithms:Systematic Review. 2022;7(3):e34699) doi: 10.2196/34699 [1]

[2] Ashenafi Zebene Woldaregaya,, Eirik Årsandb, Ståle Walderhaugb,c, David Albersd, Lena Mamykinad, Taxiarchis Botsise, Gunnar Hartvigsena- Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes https://doi.org/10.1016/j.artmed.2019.07.007.[2]

[3] Ashenafi Zebene Woldaregay1, MSc; Eirik Årsand2, PhD; Taxiarchis Botsis3, PhD; David Albers4, PhD; Lena Mamykina4, PhD; Gunnar Hartvigsen1, PhD-Data-Driven Blood Glucose Pattern Classification and Anomalies Detection: Machine-Learning Applications in Type 1 J Med Internet Res 2019;21(5):e11030) doi: 10.2196/11030 [3]

[4] M.Rajeswari1, Dr.P.Prabhu2 - A Review of Diabetic Prediction Using Machine Learning Techniques, ISSN: 2395-1303http://www.ijetjournal.org [4]

[5] Nichole S. Tyler and Peter G. Jacobs - Artificial Intelligence in Decision Support Systems for Type 1 Diabetes, 2020, 20, 3214; doi:10.3390/s20113214 [5]

[6] Josep Vehí, Iván Contreras and Silvia Oviedo- Prediction and prevention of hypoglycaemic events in type-1 diabetic patients using machine learning, DOI: 10.1177/1460458219850682 [6]

[7] Georgia Fousteri1, Ely Montee Rodrigues2, Gian Maria Giamporcaro2 and Marika Falcone 2 Cellular & Molecular

Immunology (2021) 18:515–517; https://doi.org/10.1038/s41423-020-00594-4 [7]

[8] Quan Zou1,2*, Kaiyang Qu1, Yamei Luo3, Dehui Yin3, Ying Ju4 and Hua Tang - Predicting Diabetes Mellitus With Machine Learning Techniques, 10.3389/fqene.2018.0051E7 [8]

[9] KM Jyoti Rani- Diabetes Prediction Using Machine Learning, ISN : 2456-3307 (www.ijsrcseit.com)doi https://doi.org/10.32628/CSEIT206463 [9]

[10] Cooper Midroni, Peter J. Leimbigler, Gaurav Baruah, Maheedhar Kolla, Alfred J. Whitehead- Predicting Glycemia in Type 1 Diabetes Patients: Experiments with XGBoost, www.researchgate.net/puDlication[328174513] [10]

[11] Mitushi Soni, Dr. Sunita Varma- Diabetes Prediction using Machine Learning Technique, ISSN: 2278-0181 Vol. 9 Issue 09, September-2020 [11]

[12] Jingyu Xue1st,a, Fanchao Min2rd,b, Fengying Ma - Research on Diabetes Prediction Method Based on Machine Learning, doi:10.1088/1742-6596/1684/1/012062 [12]

[13] jensia Thomas1, Anumol Joseph2, Irene Johnson3, Jeena Thomas- Machine Learning Approach For Diabetes Prediction, https://doi.org/10.30534/ijiscs/2019/13822019 [13]

[14] Minakshi R. Rajput 1*, Sushant S. Khedgikar2 - Diabetes prediction and analysis using medical attributes: A Machine learning approach, ISSN No : 1006-7930, DOI: 10.37896/JXAT14.01/314405 [14]

[15] 1 Kucharlapati Manoj Varma, 2 Dr B S Panda- Comparative analysis of Predicting Diabetes Using Machine Learning Techniques, www.jetir.org (ISSN-2349-5162) [15]

[16] Tsehay Admassu Assegie, Pramod Sekharan Nair- The Performance Of Different Machine Learning Models On Diabetes Prediction, ISSN 2277-8616 [16]

[17] Umair Muneer Butt Fadratul Hafinaz Hassan ,1Sukumar Letchmunan ,1Anees Baqir Mubashir Ali- Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications, https://doi.org/10.1155/2021/9930985 [17]

[18] Aishwariya Dutta , Md. Kamrul Hasan 3 Md. Akhtarul Islam 6 , Mehedi Masud 7 , Mohiuddin Ahmad3- Early Prediction of Diabetes Using an Ensemble of Machine Learning Models, 2022, 19, 12378. https://doi.org/10.3390/ijerph191912378 [18]

[19] Khyati K. Gandhi1, Prof. Nilesh B.Prajapati- Diabetes prediction using feature selection and classification, 2014, e-ISSN: 2348 - 4470 , print-ISSN:2348-6406 [19]

[20] Monalisa Panda1, Debani Prashad Mishra1, Sopa Mousumi Patro1, Surender Reddy Salkuti2 - Prediction of diabetes disease using machine learning algorithms, 2022, pp. 284~290 ISSN: 2252-8938, DOI: 10.11591/ijai.v11.i1.pp284-290