Diabetes Prediction Model Using Machine Learning

[1]Okardi, Biobele and [2]Henry-Uzor Okeoghene
[1,2]*Department of Computer Science and Informatics, Federal University Otuoke, Nigeria.*

***Abstract*—This diabetes detection project leveraged the Support Vector Machine (SVM) algorithm with a dataset from the National Institute of Diabetes and Digestive and Kidney Diseases, sourced from Kaggle's repository of community-published data for machine learning projects. Agile methodology, supported by essential Python libraries within the Anaconda IDE, facilitated flexible model development. A Streamlit web interface enabled easy data predictions and feature selection. Its significance lies in providing a reliable tool for early diabetes prediction, improving patient outcomes. The SVM's effectiveness with complex datasets and Agile's continuous refinement enhance accuracy. This approach combines Python, Streamlit's web interface, and SVM for effective early diabetes prediction, offering substantial benefits to healthcare efficiency and patient care.**

I. INTRODUCTION

Diabetes, a chronic metabolic disorder characterized by abnormal blood glucose levels, poses a significant global health challenge. With its prevalence on the rise, innovative approaches to predicting and managing diabetes are crucial for improving patient outcomes and reducing healthcare costs. In recent years, the fusion of advanced machine learning techniques with medical research has opened up new avenues for creating accurate and efficient diabetes prediction models. These models have the potential to transform how healthcare providers approach diabetes by enabling early intervention and personalized treatment strategies.

Traditional methods of diabetes prediction often rely on clinical markers and risk factors, which might not capture the full complexity of the disease's development. In contrast, machine learning algorithms have the capacity to analyse vast and intricate datasets, capturing intricate patterns and relationships that might otherwise go unnoticed. This empowers healthcare professionals with a more comprehensive understanding of the disease's underlying mechanisms and risk factors.

However, creating an effective diabetes prediction model is a complex endeavor. It requires the integration of diverse datasets from sources like electronic health records, patient surveys, genetic information, and even wearable devices. The data often contain a multitude of variables, both clinical and lifestyle-related, which necessitates robust data pre-processing and feature selection techniques. Additionally, the choice of machine learning algorithms is critical, as different algorithms have varying capabilities in capturing different types of patterns and relationships.

The success of a diabetes prediction model is ultimately determined by its ability to deliver accurate and reliable predictions. Evaluating these models involves employing a range of metrics that measure performance, such as accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC-ROC). These metrics provide insights into the model's ability to correctly identify individuals who are at risk of developing diabetes.

In conclusion, the intersection of machine learning and diabetes prediction holds immense potential to revolutionize healthcare approaches to diabetes management. By leveraging the power of data analysis and pattern recognition, these models enable earlier and more precise identification of individuals at risk. As technology continues to evolve and datasets grow larger and more comprehensive, the field of diabetes prediction using machine learning is poised to make substantial contributions to public health, personalized medicine, and the overall well-being of populations worldwide. Traditional diabetes prediction methods struggle with accuracy due to complex integration of genetics, clinical, lifestyle, and environmental factors. As a result, their predictions often lack accuracy, causing missed opportunities for early intervention and effective prevention strategies. This limitation hampers healthcare practitioners from identifying high-risk individuals in a timely manner. Machine learning, however, offers a solution by analysing diverse datasets to uncover hidden patterns and correlations within these complex factors. Despite its potential, effectively applying machine learning to diabetes prediction involves overcoming obstacles such as

data pre-processing, selecting relevant features, choosing appropriate algorithms, and validating the predictive models.

The goal is to develop a robust machine learning-based model that can accurately capture the intricate risk profiles associated with diabetes. By achieving this, healthcare practitioners can identify high-risk individuals earlier, leading to targeted interventions, personalized healthcare plans, and ultimately improved patient outcomes. This research seeks to enhance the accuracy of diabetes prediction and contribute to more effective preventive healthcare strategies. The potential impact includes early detection and prevention of diabetes, improved patient outcomes through timely interventions, cost-effective healthcare, personalized medicine based on individual risk factors, advancements in research and healthcare practices, enhanced clinical decision support, reduced burden on the healthcare system, and positive public health implications with global reach.

II. REVIEW OF RELATED LITERATURE

Machine learning algorithms can be combined, fine-tuned, and validated using cross-validation to create robust diabetes prediction models. By tailoring these algorithms to the specific features and characteristics of diabetes data, healthcare professionals can make more informed decisions for patient care and management.

In summary, machine learning algorithms are indispensable tools in diabetes prediction. They transform data into actionable insights, helping healthcare providers identify and manage individuals at risk of diabetes efficiently.

The paper "Diabetes Prediction using Machine Learning Techniques" by Kumar et al. (2020) focuses on addressing the significant challenge of Diabetes Mellitus, a growing fatal disease worldwide. The disease leads to various complications, including heart disease, stroke, nerve disease, and kidney damage. Given the severity of the condition, medical professionals require a reliable prediction system for early diagnosis. The paper highlights the importance of utilizing machine learning techniques to predict diabetes at an early stage. These techniques help in analyzing data from various sources and extracting valuable knowledge. The efficiency of mining diabetes-related data is emphasised as a crucial concern. The project described in the paper involves using a medical dataset to predict diabetes. R-Studio software is used as a statistical computing tool for diagnosis. The dataset utilized is the PIMA Indian database, obtained from the UCI repository, which is then analyzed to develop an effective model capable of predicting and diagnosing diabetes at an earlier stage.

The paper titled "Building Risk Prediction Models for Type 2 diabetes using machine learning techniques" by Xie et al. (2019) addresses the significant impact of diabetes, particularly type 2 diabetes, on both the health of individuals and the economic burden on the US economy. The primary objective of the study is to develop predictive models that can identify risk factors associated with type 2 diabetes. The aim is to facilitate early diagnosis, intervention, and cost reduction in healthcare. The authors conducted their study using cross-sectional data from the 2014 Behavioral Risk Factor Surveillance System, encompassing 138,146 participants, including 20,467 with type 2 diabetes. Various machine learning models were employed to predict type 2 diabetes. Univariable and multivariable weighted logistic regression models were used to investigate the associations between potential risk factors and type 2 diabetes. The study's results revealed that all predictive models achieved a high area under the curve (AUC) values, ranging from 0.7182 to 0.7949. The neural network model stood out with the highest accuracy (82.4%), specificity (90.2%), and AUC (0.7949). Meanwhile, the decision tree model exhibited the highest sensitivity (51.6%), making it suitable for initial screening and detection of type 2 diabetes. Interestingly, the study identified specific risk factors for type 2 diabetes. Individuals who slept for 9 or more hours per day or had a check-up frequency of less than 1 year were found to have higher risks for type 2 diabetes.

The paper "Voting Classification-Based diabetes mellitus prediction using Hyper tuned Machine-Learning techniques" by Mushtaq et al. (2022) focuses on addressing the challenges posed by diabetes mellitus, a chronic condition characterized by hyperglycemia. The prevalence of diabetes is increasing globally, and by 2040, it is projected to affect over 642 million individuals, highlighting the need for effective predictive models. The paper emphasizes the potential consequences of diabetes, including heart attacks, kidney damage, and blindness. Early prediction of diabetes using machine learning techniques is explored as a solution. The

authors sourced a dataset from an online repository for their study. However, they encountered an imbalanced dataset, which they addressed using techniques like Tomek and SMOTE to balance the data and manage outliers. The research employs a two-stage model selection approach. In the first stage, various machine learning algorithms such as logistic regression, Support Vector Machine, k-nearest neighbors, gradient boost, Naive Bayes, and Random Forests are applied to evaluate prediction efficiency based on patients' preconditioning. Random Forest was identified as the most effective with an 80.7% accuracy after utilizing the SMOTE oversampling technique. In the second stage, a voting algorithm is used to combine the predictions of three better-performing models: Naive Bayes Theorem, Gradient Boosting Classifier, and Random Forest. This approach yields encouraging results, achieving 82.0% accuracy with the default dataset and 81.7% accuracy with the balanced dataset.

The paper "Predictive Modelling and Analytics for Diabetes using a Machine Learning Approach" by Kaur et al. (2020) addresses the significance of diabetes as a major metabolic disorder that can have adverse effects on the entire body system. Undiagnosed diabetes poses risks such as cardiac stroke, diabetic nephropathy, and other disorders. Millions of people worldwide are affected by this disease, making early detection crucial for maintaining a healthy life. With diabetes cases on the rise globally, the disease has become a topic of global concern. The paper emphasizes the application of machine learning (ML) as a computational approach for learning from experience and improving predictive accuracy. The researchers have utilized a machine learning technique on the Pima Indian diabetes dataset to uncover trends and identify patterns with associated risk factors. They have employed the R data manipulation tool for this purpose. To classify patients into diabetic and non-diabetic categories, the study employs five distinct predictive models. These models are developed and analyzed using the R data manipulation tool. The supervised machine learning algorithms used include linear kernel support vector machine (SVM-linear), radial basis function (RBF) kernel support vector machine, k-nearest neighbour (k-NN), artificial neural network (ANN), and multifactor dimensionality reduction (MDR). Diabetes mellitus is a growing concern in Nigeria, with a high prevalence rate.

Moko and Onuodu (2021) in their study highlighted that feature selection is crucial for improving model performance and reducing computational costs, using different prediction techniques, such as logistic regression and support vector machine, the study results show that random forest predictive model achieved the highest accuracy score of 94%. The research focuses on the prevalence, hospitalization, and mortality of type 2 diabetes mellitus in Nigeria. It explores various feature selection methods for machine learning and their impact on predictive modelling. The research also reviews population-based studies on diabetes in Nigeria and highlights the importance of feature selection in enhancing model performance. The study concludes with an evaluation of different prediction techniques, including logistic regression and support vector machine, and their accuracy in predicting diabetes by adopting a hybrid methodology that combines qualitative and quantitative methods to collect and analyze data. The study also compares the performance of the proposed method with an existing method. The results show that the proposed method has good performance in predicting T2D risk.

The article "Development of various diabetes prediction models using machine learning techniques" by Shin et al. (2022) discusses the development of multiple diabetes prediction models to utilize easily accessible health screening test parameters. In this study, the authors aimed to address the need for reliable diabetes mellitus (DM) prediction models. They employed two sets of variables to create eight distinct DM prediction models. The first set of variables consisted of 62 examination results from commonly used parameters obtained from a tertiary university hospital. The second set consisted of 27 out of the 62 variables that are part of national routine health check-ups. The authors utilized gradient boosting and random forest algorithms to develop these prediction models. Internal validation was performed using the stratified 10-fold cross-validation method. The results showed that the DM model using the 62-variable set yielded the highest area under the receiver operating characteristic curve (ROC-AUC) of 0.928 among all the models. Even with the simplified 27-variable set, the ROC-AUC remained reasonably good, ranging between 0.842 and 0.880, with a decrease of around 0.04. Notably, the inclusion of fasting glucose increased accuracy by up to 11.5%, highlighting its significance in prediction. The study concludes that

the authors successfully developed diabetes prediction models that demonstrate strong performance. These models are based on parameters commonly assessed during health checkups in both tertiary university hospitals and national routine screenings. The authors express plans for prospective external validation, with the hope that these prediction models will find widespread utility in clinical practice.

The paper "Current Techniques for Diabetes Prediction: review and case study" authored by Marie-Sainte et al. (2019) delves into the prevalent nature of diabetes as a widespread disease. In recent years, a multitude of Machine Learning (ML) techniques have been employed to predict diabetes, with a shift towards exploring the robust array of Deep Learning (DL) algorithms due to the increasing complexity of the problem. The article highlights a significant achievement in diabetes prediction with a combined CNN-LSTM model, achieving an accuracy of 95.1%. While various ML algorithms have been applied to address this challenge, there's a subset of classifiers that have been underutilized or not used at all in this context. This creates a potential area of interest to assess the performance of these classifiers in diabetes prediction. Furthermore, the article points out that there's a lack of recent comprehensive surveys that review and compare the performance of all proposed ML, DL techniques, and combined models for diabetes prediction. To address this, the paper surveys ML and DL techniques-based diabetes prediction models published over the last six years. Additionally, a study was conducted to implement rarely and not commonly used ML classifiers on the Pima Indian Dataset to analyze their performance, resulting in classifiers achieving an accuracy range of 68%–74%. The authors recommend the incorporation of these underutilized classifiers in diabetes prediction, along with the potential for enhancing their performance through the development of combined models.

The research paper "Diabetes Prediction in Teenagers using Machine Learning Algorithms" authored by Vrindavanam et al. (2023) addresses the critical issue of Diabetes Mellitus (DM), a hazardous condition that can lead to widespread health problems due to elevated blood glucose levels, various factors such as age, lack of physical activity, high blood pressure, and poor nutrition. The primary goal of this study is to develop a predictive system focused on individuals aged 10 to 30, to merge outcomes from diverse machine-learning algorithms. Machine learning techniques are leveraged to enhance performance and improve prediction accuracy. The study employs several methods to detect diabetes at an early stage, including Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), K Nearest Neighbours (KNN), and XGBoost. Notably, the research concludes that the Random Forest and XGBoost Classifier algorithms outperform other machine learning techniques, achieving a maximum accuracy of 86%.

The paper "Prediction of Diabetes Using Hybridization-based Machine Learning Algorithm" authored by R. L. Naidu Boddeda et al. (2023), presents a comprehensive approach to predicting diabetes using machine learning algorithms. The authors address the challenge of early diabetes prediction, which can significantly reduce the risk and impact of the disease. Diabetes, characterized by prolonged elevated blood sugar levels, falls under the category of metabolic disorders. The paper highlights the complexity of accurate diabetes prediction due to limited labelled data and the presence of outliers or incomplete values in diabetes databases. To tackle these challenges, the authors propose a robust architecture for diabetes prediction. This architecture includes steps such as outlier exclusion, missing value imputation, data standardization, feature selection, and the utilization of various machine learning classifiers. In particular, the authors suggest the application of multiple machine learning classifiers, including k-nearest Neighbours, Decision Trees, Random Forest, Logistic Regression, and Support Vector Machine (SVM). To enhance prediction accuracy, a hybrid algorithm is introduced, involving a combination of different algorithms. The authors adopt a hybridization approach using various algorithms to achieve improved accuracy in diabetes prediction. The Pima Indian Diabetes Data Collection serves as the data source for the study. By combining outlier handling, data pre-processing, feature selection, and hybrid machine learning techniques, the proposed approach aims to provide more reliable and accurate predictions of diabetes.

The paper "A Machine learning approach to reduce the diabetes patient's readmission risk using a novel pre-processing technique" by Rajput and Alashetty (2022) addresses the critical issue of diabetes patients' readmission risk by utilizing a machine

learning approach combined with a novel pre-processing technique. The primary indicator of diabetes is elevated blood glucose levels, which can lead to various complications and result in multiple hospitalizations for diabetic patients. The impact on both the quality of treatment and the overall cost of medical services due to extended hospital stays necessitates a solution. The paper emphasizes the suitability of a machine learning-based approach to address this issue, considering that healthcare systems capture a wealth of data from clinical patient interactions. The study employs a dataset comprising one hundred thousand medical records for seventy thousand diabetic patients across 130 institutions in the United States. The novel pre-processing technique introduced in the paper plays a crucial role in reducing the dimensions of the dataset. By reducing the 55 characteristics of the United States diabetes dataset to 21, the technique streamlines the machine learning process. The study employs various machine learning algorithms, including K-Nearest Neighbour (KNN), support vector machines (SVM), decision trees, random forests (RF), logistic regression (LR), gradient boosting (GB), and Gaussian naive bayes (GaussianNB), to assess the accuracy of the data engineering technique. The results highlight that Gradient Boosting outperformed other machine learning methods, achieving an accuracy of 77.4% in predicting early readmission risk for diabetic patients.

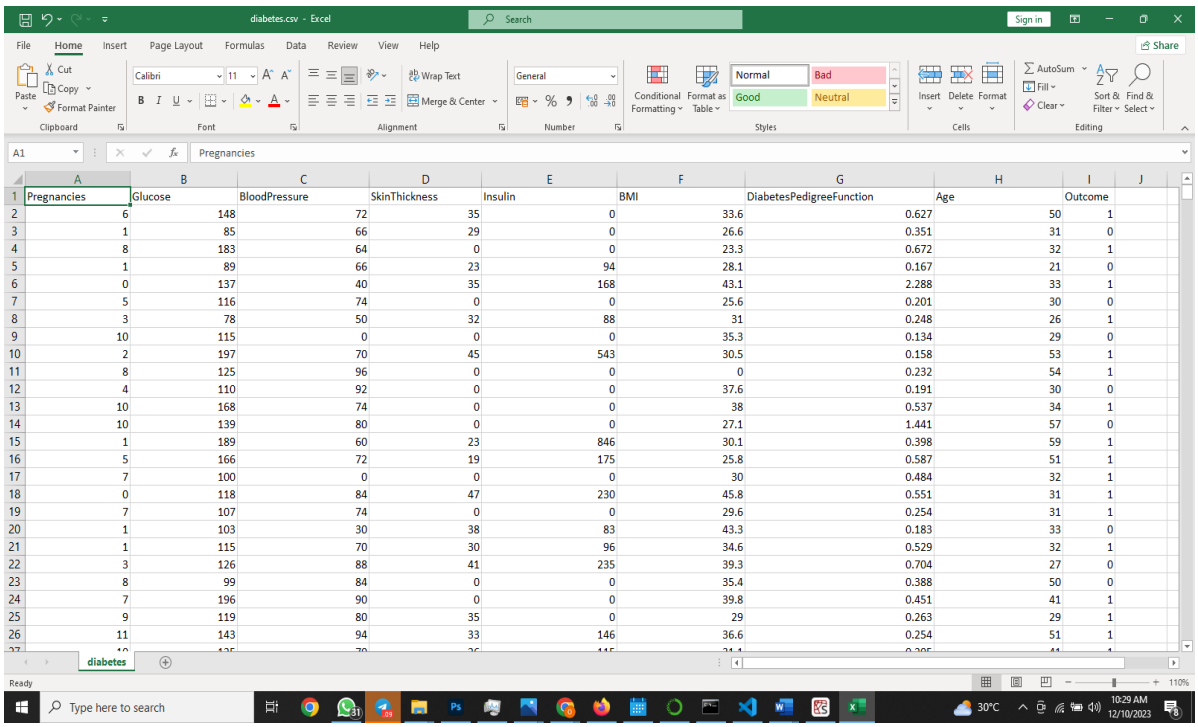
The paper "Detecting Diabetes Using Machine Learning Algorithms" by Thabit et al. (2022, September) addresses the urgency of modern techniques to detect and diagnose diabetes, which has become a rapidly spreading global concern. In this context, machine learning, particularly artificial intelligence (AI), has emerged as a promising approach for early diagnosis and prediction of diabetes. AI's efficiency in disease detection, especially for critical cases, has been well-established through its utilization of medically approved databases. The study focuses on estimating diabetes incidence by employing classification techniques that leverage various machine learning algorithms. These algorithms include Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbour (KNN), and Naive Bayes (NB) classifiers. These techniques are applied to assess diabetes prevalence, encompassing data pre-processing steps such as data cleaning, feature selection, data scaling, and data splitting. Using patient electronic health data from the Pima Diabetes Database of India, the research is conducted using the Python programming language within the ANACONDA environment, specifically under Jupyter Notebook (version 6.1.4) and Spyder (version 4.1.5) applications. The outcomes demonstrate the model's remarkable accuracy, capable of predicting the presence of diabetes with up to 100% accuracy.

The paper "A hybrid approach for predicting diseases using clustering and classification techniques" by Geluvaraj et al. (2022) highlights the significant role of machine learning in various domains such as image identification, data mining, natural language processing, and disease diagnosis. The paper explores different machine learning techniques applied to the diagnosis of diseases, specifically heart and diabetes diseases. The authors emphasize that machine learning algorithms for data analysis continually improve over time, relying on quality data to operate effectively. These techniques have been widely used by practitioners and researchers in the healthcare field for prediction purposes. Machine learning algorithms like KNN (K-Nearest Neighbors), K-Means, DSC (Decision Stump Classifier), RFC (Random Forest Classifier), and NB (Naive Bayes) have been employed to predict diseases with accuracy. The paper highlights that the DSC algorithm achieves the highest accuracy among the models used. Its strength lies in reducing ambiguity in complex decisions by assigning appropriate values to different action results. Following closely is the KNN algorithm, which offers highly accurate predictions and can be applied in applications requiring precision without the need for a human-readable model. The third-highest accuracy rate is attributed to the Naive Bayes algorithm. This algorithm is known for its speed and ease in forecasting test datasets. It performs well in multiclass forecasting when the assumption of independence is met, and it requires less training data. The Naive Bayes algorithm is highlighted as superior to other models such as logistic regression (LR).

III. METHODOLOGY

The project focuses on developing a diabetic disease prediction model using a machine learning technique, specifically the Support Vector Machine (SVM) classifier algorithm. The chosen methodology for this

project is the Agile methodology. The Agile methodology was chosen for developing the diabetic disease prediction model with the Support Vector Machine (SVM) algorithm due to its flexible and iterative nature.



Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1
3	126	88	41	235	39.3	0.704	27	0
8	99	84	0	0	35.4	0.388	50	0
7	196	90	0	0	39.8	0.451	41	1
9	119	80	35	0	29	0.263	29	1
11	143	94	33	146	36.6	0.254	51	1

Figure 1. Dataset Sample

This approach allows for continuous refinement of the model, incorporating feedback and insights as the project progresses. It encourages collaboration among diverse team members, including healthcare professionals and data scientists, ensuring that the model is accurate and relevant. Agile focus on delivering incremental value means that stakeholders can start benefiting from the model's insights early on, enabling timely intervention for diabetic patients. Additionally, Agile emphasis on early risk identification and mitigation aligns well with the sensitive nature of healthcare projects, ensuring ethical and technical challenges are addressed promptly. Overall, Agile provides a structured yet adaptable framework for developing a robust disease prediction model, prioritizing continuous improvement and value delivery.

i.Proposed System

The proposed system aims to develop a robust diabetic disease prediction model using the Support Vector Machine (SVM) classifier algorithm. It focuses on female patients, aged at least 21, of Pima Indian heritage, utilizing a dataset from the National Institute of Diabetes and Digestive and Kidney Diseases. This dataset encompasses key diagnostic measurements, including pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI, diabetes pedigree function, and age. Additionally, it features an outcome variable, labeled "Outcome," indicating the presence or absence of diabetes (0 or 1). The system's primary objectives are to develop a reliable prediction model, train it using SVM in Google Colab, implement it using Python, and deploy a user-friendly interface via Streamlit in Anaconda. Python will be the core programming language for this Endeavor.

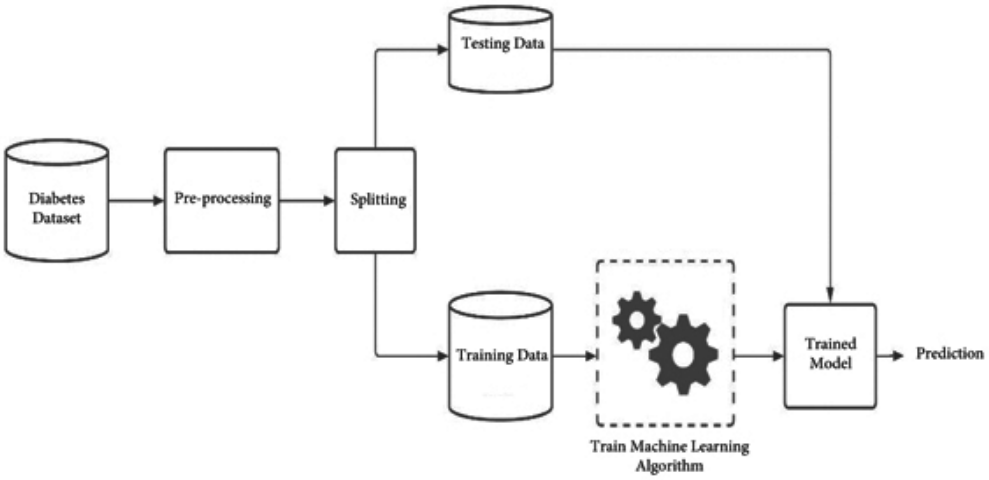


Figure 2. Proposed System Model

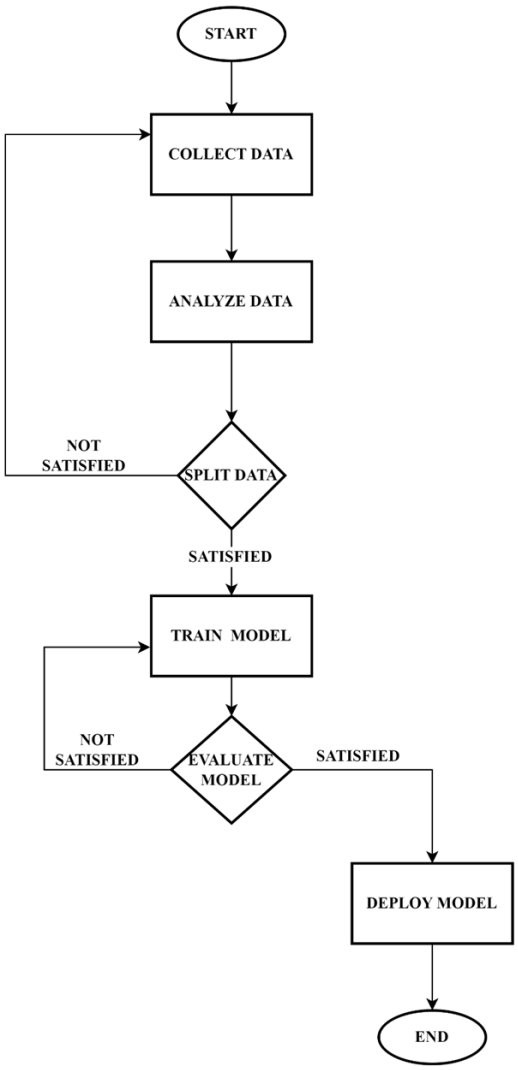


Figure 3. Flowchart

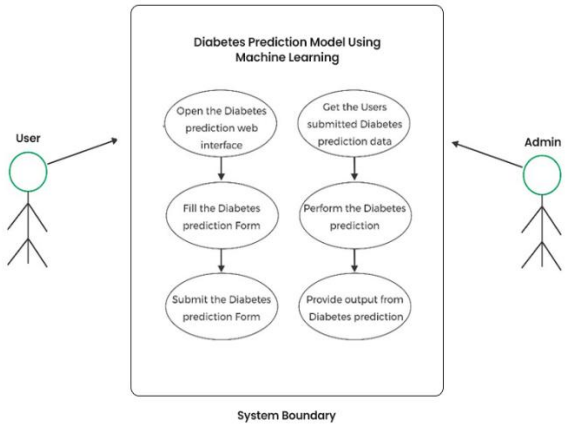


Figure 4. Use-case diagram

IV. RESULTS AND DISCUSSIONS

The system results refer to the outcomes and predictions generated by the Diabetic Disease Prediction System after processing input data. These results are based on the analysis and computations performed by the system using the trained Support Vector Machine (SVM) model. Specifically, the system evaluates the input data against the SVM model and provides a prediction regarding the likelihood of a person having diabetes.

Model Evaluation

Accuracy Score

```
[19] # accuracy score on the training data
     X_train_prediction = classifier.predict(X_train)
     training_data_accuracy = accuracy_score(X_train_prediction, Y_train)

[20] print('Accuracy score of the training data : ', training_data_accuracy)

     Accuracy score of the training data :  0.7833876221498371

[21] # accuracy score on the test data
     X_test_prediction = classifier.predict(X_test)
     test_data_accuracy = accuracy_score(X_test_prediction, Y_test)

[22] print('Accuracy score of the test data : ', test_data_accuracy)

     Accuracy score of the test data :  0.7727272727272727
```

Figure 5. Model Evaluation

Considering the need for a functional and practical web-based system to predict diabetes using a support vector machine algorithm, the following visuals below illustrate the proposed solution.

Figure 6. Output for a diabetic patient of the proposed system

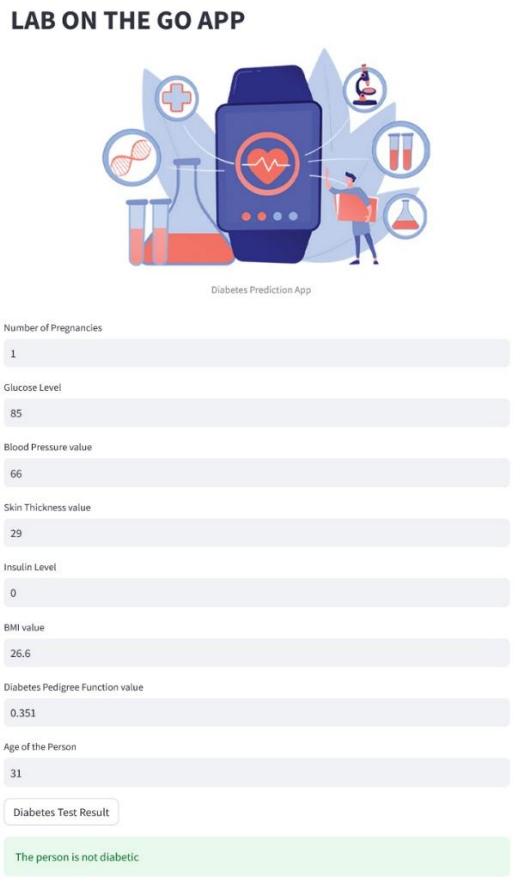


Figure 7. Output for a non-diabetic patient of the proposed system

V. CONCLUSION

The Diabetic Disease Prediction System represents a significant step toward enhancing healthcare

outcomes through early diabetes detection and proactive management. With a strong emphasis on data pre-processing and a robust SVM model, the system achieves reliable predictions. The system leverages the Support Vector Machine (SVM) classifier algorithm and is tailored to female patients, aged at least 21, of Pima Indian heritage. The dataset used for training and testing the model is sourced from the National Institute of Diabetes and Digestive and Kidney Diseases and includes essential features for accurate prediction.

The development process encompasses data acquisition, pre-processing, SVM model creation, model evaluation, and user interface design. The system offers potential benefits, including early diabetes detection, accessibility for users with varying technical proficiency, personalized predictions, and support for healthcare professionals in making informed decisions. It combines data science, machine learning, and user interface development to address a critical healthcare issue. The user-friendly interface facilitates easy data input, making it accessible to a wide range of users. The system's potential impact on healthcare is substantial, providing valuable insights into diabetes risk and promoting timely intervention.
However, it is essential to recognize that the system's accuracy is subject to data quality and model performance. Continuous monitoring and validation are necessary to ensure the system's effectiveness in real-world healthcare scenarios. Furthermore, ethical considerations regarding data privacy and responsible data usage must be upheld at all times.

VI. RECOMMENDATIONS

There is a need for the following:

- Continuous Improvement: Regularly update the SVM model and incorporate the latest advancements in machine learning to enhance prediction accuracy.
- Extensive Testing: Conduct rigorous testing with diverse datasets to evaluate the system's performance across various demographics and healthcare contexts.
- Data Privacy: Ensure strict adherence to data protection and privacy regulations to maintain patient trust and compliance.
- User Training: Provide user training to healthcare professionals and individuals using the system to maximize its benefits.
- Integration with Healthcare Systems: Explore opportunities for integrating the system with existing healthcare infrastructure for seamless patient care.
- Feedback Mechanism: Implement a feedback mechanism to gather user feedback and make necessary system improvements.
- Documentation: Maintain comprehensive documentation of the system's development and operation for transparency and future enhancements.
- Collaborative Research: Collaborate with healthcare institutions for further research and validation of the system's performance in clinical settings.

The Diabetic Disease Prediction System has the potential to make a significant contribution to healthcare by enabling early diabetes detection and empowering individuals and healthcare professionals with valuable predictive insights. Ongoing development, validation, and responsible usage are key to its long-term success.

REFERENCES

[1] American Diabetes Association. (2021). Diabetes Basics. Retrieved from: https://www.diabetes.org/resources/statistics/statistics-about-diabetes

[2] Diabetes Dataset. (2020, August 5). Kaggle. https://www.kaggle.com/datasets/mathchi/diabetes-data-set

[3] Geluvaraj, B., Santhosh, K., Prabhu, N., Reddy, A., Sandhya, T., & Bhaskar, S. V. (2022, January). A hybrid approach for predicting diseases using clustering and classification techniques. In 2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI) (pp. 1-6). IEEE. https://ieeexplore.ieee.org/abstract/document/9752552

[4] Kaur, H., & Kumari, V. (2020). Predictive modelling and analytics for diabetes using a machine learning approach. Applied Computing and Informatics, 18(1/2), 90–100. https://doi.org/10.1016/j.aci.2018.12.004

[5] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational

and Structural Biotechnology Journal, 15, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005

[6] Kumar, S., Obulesu, O., & Ramudu, B. V. (2020). Diabetes Prediction using Machine Learning Techniques. Helix, 10(2), 136–142. https://doi.org/10.29042/2020-10-2-136-142

[7] Kumar, Y., Koul, A., Singla, R., & Ijaz, M. F. (2022b). Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. Journal of Ambient Intelligence and Humanized Computing. https://doi.org/10.1007/s12652-021-03612-z

[8] Mahabub, A. (2019). A robust voting approach for diabetes prediction using traditional machine learning techniques. SN Applied Sciences, 1(12). https://doi.org/10.1007/s42452-019-1759-7

[9] Marie-Sainte, S. L., Aburahmah, L., Almohaini, R., & Saba, T. (2019). Current Techniques for Diabetes Prediction: review and case study. Applied Sciences, 9(21), 4604. https://doi.org/10.3390/app9214604

[10] Moko, A., & Onuodu, F.E. (2021). An Improved Model for Feature Selection on Type 2 Diabetes Risk Prediction in Nigeria. International Journal of Basic Science and Technology (IJBST), 7(2), pp 123 – 133.

[11] Mushtaq, Z., Ramzan, M. F., Ali, S., Baseer, S., Samad, A., & Husnain, M. (2022). Voting Classification-Based diabetes mellitus prediction using Hypertuned Machine-Learning techniques. Mobile Information Systems, 2022, 1–16. https://doi.org/10.1155/2022/6521532

[12] Rajput, G. G., & Alashetty, A. (2022, December). A Machine learning approach to reduce the diabetes patient's readmission risk using a novel preprocessing technique. In 2022 4th International Conference on Circuits, Control, Communication and Computing (I4C) (pp. 173-177). IEEE. https://ieeexplore.ieee.org/abstract/document/10057842

[13] Rishika Lakshmi Naidu Boddeda, Ritika Prasad, Shanmuk Srinivas Amiripalli, Mukkamala SNV Jitendra, "Prediction of Diabetes Using Hybridization based Machine learning algorithm", 2023 International Conference on Smart Computing and Application (ICSCA), pp.1-5, 2023. https://ieeexplore.ieee.org/document/10087491

[14] Shin, J., Kim, J., Lee, C., Yoon, J., Kim, S., Song, S., & Kim, H. (2022). Development of various diabetes prediction models using machine learning techniques. Diabetes & Metabolism Journal, 46(4), 650–657. https://doi.org/10.4093/dmj.2021.0115

[15] Thabit, Q. Q., Fahad, T. O., & Dawood, A. I. (2022, September). Detecting Diabetes Using Machine Learning Algorithms. In 2022 Iraqi International Conference on Communication and Information Technologies (IICCIT) (pp. 131-136). IEEE. https://ieeexplore.ieee.org/abstract/document/10010408

[16] Vrindavanam, J., Haarika, R., Sindhu, M. G., & Kumar, K. S. (2023, March). Diabetes Prediction in Teenagers using Machine Learning Algorithms. In 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 343-347). IEEE. https://ieeexplore.ieee.org/abstract/document/10112286

[17] World Health Organization. (2021). Global report on diabetes. Retrieved from: https://www.who.int/publications/i/item/9789241565257

[18] Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019). Building risk prediction models for Type 2 diabetes using machine learning techniques. Preventing Chronic Disease, 16. https://doi.org/10.5888/pcd16.

[19] Sasikumar, B., Naveenraju, D., Anand, K., Hariharan, S., Sudhakaran, P., & Bharathiraja, N. (2020). Diabetes prediction using sensors by analysing skin temperature. J. Eng. Sci. Technol, 15(2), 1357-1370. https://jestec.taylors.edu.my/Vol%2015%20issue%202%20April%202020/15_2_43.pdf