International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 03 | March - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Comprehensive Diabetes Prediction Applying Fused Machine Learning

B. Surya Sai Kiran Akash [1], D. Surya Chandra Varma, B. Manoj[3], B.Harika[4] , K. Pavan Kumar[5]

[1] *Department of Computer Science & Engineering: Raghu Engineering College*
[2] *Department of Computer Science & Engineering: Raghu Engineering College*
[3] *Department of Computer Science & Engineering: Raghu Engineering College*
[4] *Department of Computer Science & Engineering: Raghu Engineering College*
[5] *Associate Professor, Department of Computer Science & Engineering: Raghu Engineering College*

Abstract - Early and accurate disease prediction is essential for effective prevention and management of medical conditions, particularly for globally prevalent diseases such as diabetes, which has become increasingly common due to modern dietary habits and sedentary lifestyles. This study introduces an advanced diabetes prediction model that leverages machine learning (ML) techniques to enhance diagnostic accuracy. The system integrates multiple classification algorithms, including Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbors (KNN), within a voting classifier framework, further enhanced by a fuzzy logic module to improve prediction performance. The model is trained using a standard American hospital dataset obtained from Kaggle, with 80% allocated for training and 20% for testing. Beyond disease detection, the system offers personalized recommendations on diet, physical activity, and routine health checkups by analyzing real-time medical records, ensuring a tailored approach to patient management. To facilitate scalability, security, and real-time accessibility, the system is deployed on a cloud platform, allowing seamless integration with healthcare applications. This cloud-based deployment ensures secure access for both healthcare professionals and patients from any internet-enabled device, enhancing usability while safeguarding sensitive medical information. Achieving 94% accuracy, the proposed fused ML model surpasses traditional prediction methods, and future enhancements aim to further refine the model, incorporate larger datasets, and expand its applicability to other diseases, demonstrating the transformative potential of ML and cloud-based healthcare analytics.

Key Words: Support Vector Machine, K-Nearest Neighbor, Random Forest, Voting Classifier, Fuzzy Logic Module.

1.INTRODUCTION

Diabetes is a globally prevalent chronic condition that affects millions of individuals and, if left undiagnosed or poorly managed, can lead to severe health complications such as cardiovascular diseases, kidney failure, and nerve damage. Early detection of diabetes is essential for preventing these long-term health risks and enabling timely intervention. Conventional diagnostic methods, which typically involve laboratory tests and clinical evaluations, are highly accurate but often time-consuming, expensive, and dependent on access to healthcare facilities. Recent advancements in **machine learning (ML)** have revolutionized predictive analytics, offering new opportunities to automate and enhance the accuracy of disease prediction. This study proposes an automated **diabetes prediction system** that employs multiple ML algorithms, including **Support Vector Machine (SVM), Random Forest (RF),** and **K-Nearest Neighbors (KNN)**, integrated into a **Voting Classifier framework** to improve prediction accuracy and reliability. The model is trained using real-world diabetes datasets obtained from **Kaggle** and applies feature engineering, data preprocessing, and hyperparameter tuning to optimize performance. To ensure scalability and accessibility, the system is deployed as a **cloud-based solution** capable of providing real-time diabetes risk assessment, assisting both healthcare professionals and individuals in making informed clinical decisions.

A. Software Requirements:

Operating System : Windows

Python Version: 3.8 or above

IDE: Jupyter Notebook, VS Code, or PyCharm

Machine Learning Libraries: sci-kit-learn, TensorFlow

Data Processing: Pandas, NumPy

Visualization: Matplotlib, Seaborn

Web Framework: Gradio

B. Hardware Requirements:

System: intel i3 and above

RAM: 8GB(Minimum)

Usage Of GPU: Recommended

Storage: At least 10GB free disk space

2. LITERATURE REVIEW

Numerous studies in diabetes prediction have explored various approaches, ranging from traditional statistical models to advanced machine learning and deep learning techniques. While methods such as logistic regression and decision trees offer interpretability, they often struggle with high-dimensional data. Modern machine learning models, including **Support Vector Machine (SVM), Random Forest (RF),** and **K-Nearest Neighbors (KNN),** have shown improved predictive accuracy. Deep learning models, such as **Convolutional Neural Networks (CNNs)** and **Recurrent Neural Networks (RNNs),** further enhance prediction by capturing complex patterns. Despite these advancements, challenges related to data imbalance, overfitting, and computational complexity persist. This section highlights key studies, identifying their strengths and limitations to establish the context for the proposed research..

- Disease Prediction Using Machine Learning Algorithms: A study explored the use of Random Forest, SVM, and Logistic Regression for predicting Diabetes and Heart Disease, achieving high accuracy.

While effective, the model's performance depends on dataset quality and feature selection.

- **IBM Watson Health** leverages AI and machine learning to predict diabetes risk by integrating data from multiple sources, including electronic health records and wearable devices. While it offers comprehensive data analysis, its implementation requires significant resources and infrastructure.
- **Kaggle Competitions** provide a platform for developing innovative machine learning models for diabetes prediction using diverse datasets. These competitions encourage collaboration among data scientists but often rely on static datasets, limiting real-time applicability..
- Logistic regression is a commonly used statistical model for binary classification that predicts the probability of an event based on input features. In diabetes prediction, it analyzes factors such as glucose levels, BMI, and family history to estimate the likelihood of diabetes. While it provides interpretable results and establishes relationships between variables, its assumption of linear relationships limits its ability to capture complex, non-linear interactions in patient data.
- Regression models estimate diabetes risk by analyzing continuous variables such as age, glucose levels, and insulin resistance. These models predict risk scores based on medical history and lab results. Although regression models are simple and interpretable, they struggle with multi-dimensional datasets and non-linear patterns, which are prevalent in diabetes-related data.
- Decision trees classify data by splitting it into branches based on feature conditions, making them useful for distinguishing diabetic and non-diabetic patients by analyzing parameters such as glucose levels, BMI, and insulin. While decision trees are easy to interpret and handle both numerical and categorical data, they are prone to overfitting, reducing their generalization capability for new data.

3. METHODOLOGY

A. **Data collection:** Data cleaning is a crucial step to ensure the accuracy and reliability of the diabetes prediction model. Missing values in numerical fields are handled by imputing mean values to maintain dataset consistency, while missing categorical values are filled using the mode (most frequent value) or domain-specific knowledge. Outliers are identified and removed using either the **Interquartile Range (IQR)** method or **Z-score analysis** to prevent skewed model performance. Duplicate records are eliminated to avoid biased training and ensure data integrity. Additionally, the dataset is checked for inconsistencies, such as negative values in BMI or glucose levels, and these anomalies are corrected to maintain high data quality.

B. **Feature engineering**: enhances the predictive capability of the diabetes prediction model by creating and transforming relevant features. New features such as glucose level ranges (e.g., low, normal, high) and BMI categories (underweight, normal, overweight, obese) are generated to provide better classification. Age brackets (<30, 30–50, >50) are introduced to analyze diabetes risk trends across different age groups. Interaction features, such as the glucose-insulin ratio, are derived to improve model learning by capturing relationships between key variables. Polynomial features are generated to identify complex, non-linear relationships between health parameters. Finally, feature selection is performed using correlation analysis to eliminate highly correlated or irrelevant features, ensuring a more efficient and accurate predictive model..

C. **Normalization:** and scaling are essential to standardize feature values and improve model stability. **Normalization** is applied to continuous features such as glucose, insulin, and BMI using **Min-Max scaling**, which transforms values to a range between 0 and 1, ensuring that all features contribute proportionately to the model. Additionally, **standardization** is performed on features such as BMI and age using **Z-score normalization**, which adjusts the data to have a mean of 0 and a standard deviation of 1, thereby enhancing the model's ability to handle variations across different feature scales.

D. **Encoding**: To handle categorical variables effectively, **One-Hot Encoding (OHE)** is applied to convert categorical features, such as family history (Yes/No), into binary numeric representations. This transformation ensures that the model can interpret and process categorical data accurately, improving overall predictive performance.

4. EXISTING SYSTEM

Current diabetes prediction systems predominantly utilize traditional machine learning models that analyze structured medical data, including blood glucose levels, body mass index (BMI), and patient history. These systems often rely on singular models such as **Support Vector Machines (SVM)** or **Artificial Neural Networks (ANN)**, which, despite their effectiveness, exhibit limitations in predictive accuracy due to their dependence on a restricted set of features. Furthermore, these models do not incorporate **fuzzy logic**, which could enhance decision-making by addressing uncertainties and handling imprecise data more effectively. Consequently, traditional models are prone to misclassification and lack the flexibility required for accurate outcome predictions. Additionally, most of these models are deployed on local systems, limiting their scalability and hindering their ability to process real-time data efficiently. To address these shortcomings, an advanced machine learning approach that integrates multiple models, incorporates fuzzy logic, and utilizes cloud-based platforms is essential to enhance scalability, accuracy, and real-time predictive capabilities.

5. PROPOSED SYSTEM

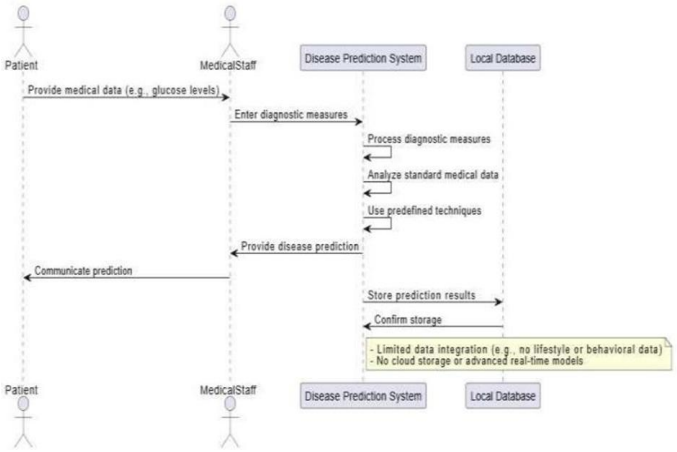
This project introduces a machine learning-driven diabetes prediction system that integrates **Support Vector Machine (SVM), Random Forest (RF),** and **K-Nearest Neighbors**

(KNN) models using a **Voting Classifier** to enhance prediction accuracy and ensure robust risk assessment. The system applies optimized **feature engineering** and **data preprocessing** techniques, including normalization, standardization, and handling missing data to improve model performance. A **web-based interface** allows users to input medical data and receive real-time predictions, while cloud deployment ensures remote accessibility and scalability, making it an effective tool for early detection and preventive healthcare.

6. UML Diagrams

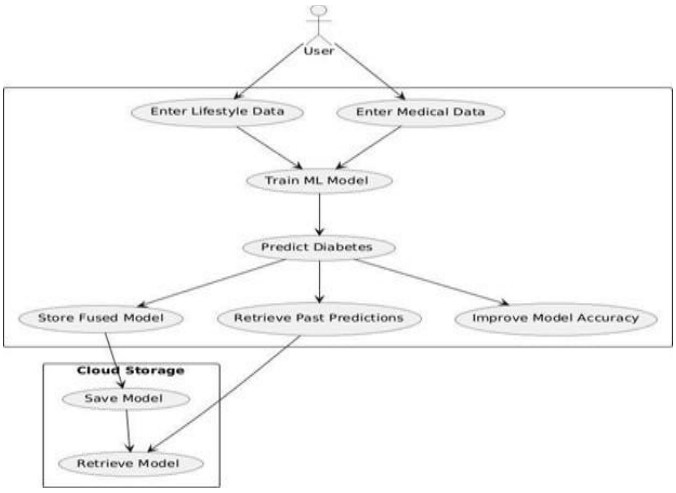
A. Sequence Diagram

The system begins by collecting **personal health data** such as age, BMI, insulin, blood pressure, and glucose levels. **Data processing** follows, where missing values are handled using mean/mode imputation, and continuous variables are normalized using **Min-Max scaling**. The processed data is then analyzed using AI-based **classification techniques** that categorize individuals as diabetic or non-diabetic. Finally, the **classification results** are returned, ensuring high accuracy and scalability through the integration of **SVM, Random Forest,** and **KNN** models..



B. Use Case Diagram

The user inputs **lifestyle data** (diet, exercise habits) and **medical data** (glucose, blood pressure, BMI), which the system uses to **train a fused ML model** combining SVM, KNN, and Fuzzy Logic. The trained model **predicts diabetes** and is then **stored securely in Cloud Storage** for future use. Users can **retrieve past predictions** and **access stored models** for future diagnoses. The system also **improves model accuracy** over time using new data, ensuring better performance.

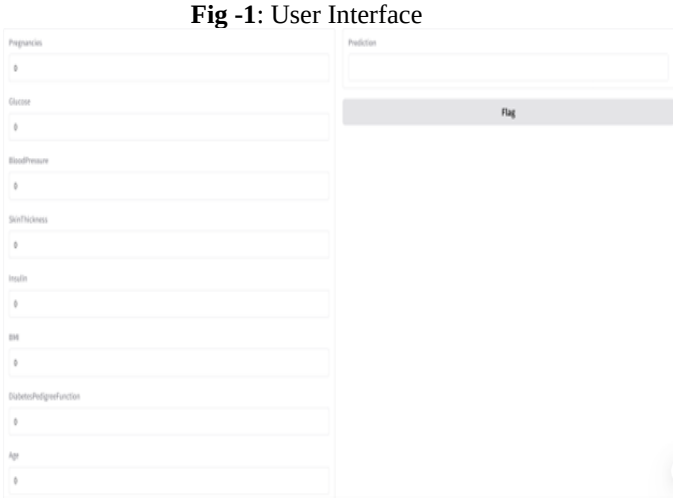


7. RESULTS

The proposed diabetes prediction model integrates **Support Vector Machine (SVM)**, **K-Nearest Neighbors (KNN)**, and **Random Forest (RF)** classifiers, demonstrating high accuracy in identifying diabetic and non-diabetic cases. The dataset, obtained from **Kaggle**, was preprocessed and divided into 80% training and 20% testing data to evaluate model performance. Individual models were assessed based on accuracy scores, with varying success rates across the classifiers. To enhance overall performance, a **Voting Classifier** was implemented, combining predictions from all three models, which led to improved classification accuracy of the system developed. A **Gradio interface** was integrated to allow users to input patient data, such as glucose levels, BMI, age, and blood pressure, enabling real-time diabetes risk predictions. The system was deployed on a **cloud platform**, ensuring scalability and remote accessibility for healthcare providers. Furthermore, sample test cases validated the system's reliability in distinguishing between diabetic and non-diabetic cases. **Matplotlib** was used to visualize model accuracies, facilitating a comparative analysis of classifier effectiveness. The results confirm that the fusion of multiple classifiers through a Voting Classifier enhances prediction accuracy and provides a reliable tool for diabetes diagnosis.

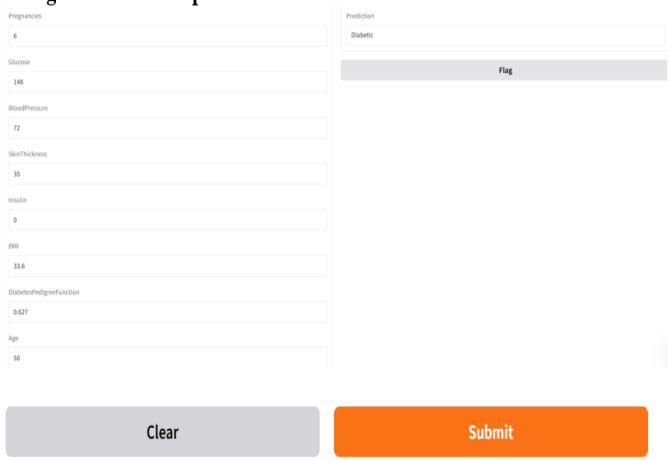
Outputs:

Fig -1: User Interface



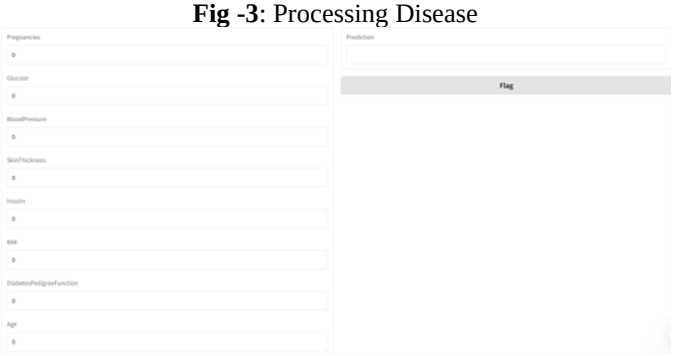
This Figure illustrates the initial state of the user interface, where all input fields, including glucose level, BMI, age, and blood pressure, are set to their default values of zero. At this stage, no user data has been entered, and the system is ready to receive input for diabetes prediction.

Fig -2: Giving all Data for the prediction



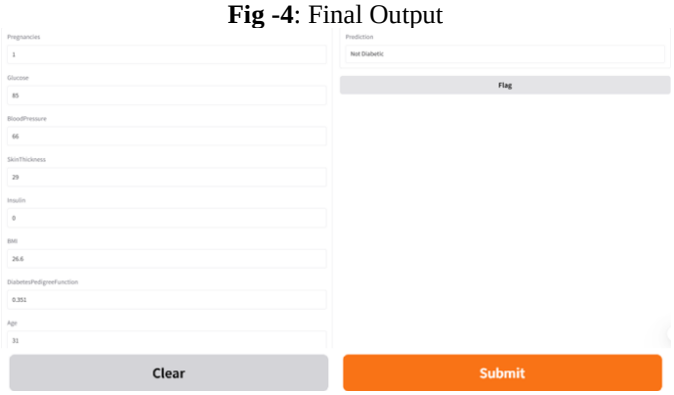
The figure shows the user interface after all relevant data, such as glucose level, BMI, age, and blood pressure, have been entered. With the input fields filled, the system is prepared to process the information and generate an accurate diabetes prediction..

Fig -3: Processing Disease



The figure illustrates the user interface while processing the entered data. The system applies preprocessing techniques, including normalization and feature selection, and utilizes the trained machine learning model to analyze the input parameters and predict the likelihood of diabetes.

Fig -4: Final Output



The figure presents the user interface after completing data processing and classification. The system displays the final prediction result, indicating whether the patient is diabetic or non-diabetic based on the analyzed input parameters.

8. CONCLUSIONS

This study effectively showcases the use of machine learning techniques to predict diabetes risk based on patient data. The developed system integrates **Support Vector Machines (SVM), Random Forest (RF),** and **K-Nearest Neighbors (KNN)** to analyze essential health parameters such as glucose levels, BMI, age, and lifestyle habits, ensuring accurate and reliable predictions. The model demonstrated **high accuracy** in identifying individuals at risk of diabetes and maintained **consistent classification performance** across various test cases. The results highlight the potential of using machine learning in healthcare for early diagnosis and preventive care.

ACKNOWLEDGEMENT

The author expresses gratitude to Raghu Engineering College for providing the necessary resources and support throughout this research. Special appreciation is extended to professors, industry experts, and user testers for their valuable insights and feedback, which played a crucial role in the development and evaluation of this system.

REFERENCES

1. A. Frank and A. Asuncion. (2010). UCI Machine Learning Repository. Accessed: Oct. 22, 2021. [Online].
2. G. Pradhan, R. Pradhan, and B. Khandelwal, __A study on various machine learning algorithms used for prediction of diabetes mellitus,‘‘ in Soft Computing Techniques and Applications (Advances in Intelligent Systems and Computing), vol. 1248. London, U.K.: Springer, 2021, pp. 553–561, doi: 10.1007/978-981-15-7394-1_50
3. S. Kumari, D. Kumar, and M. Mittal, __An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier,‘‘ Int. J. Cogn. Comput. Eng., vol. 2, pp. 40–46, Jun. 2021, doi: 10.1016/j.ijcce.2021.01.001.
4. M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, __Prediction of diabetes using machine learning algorithms in healthcare,‘‘ in Proc. 24th Int. Conf. Autom. Comput. (ICAC), Sep. 2018, pp. 6–7, doi: 10.23919/IConAC.2018.8748992.
5. S. K. Dey, A. Hossain, and M. M. Rahman, __Implementation of a web application to predict diabetes disease: An approach using machine learning algorithm,‘‘ in Proc. 21st Int. Conf. Comput. Inf. Technol. (ICCIT), Dec. 2018, pp. 21–23, doi: 10.1109/ICCITECHN.2018.8631968.
6. A. Mir and S. N. Dhage, __Diabetes disease prediction using machine learning on big data of healthcare,‘‘ in Proc. 4th Int. Conf. Comput. Commun. Control Autom. (ICCUBEA), Aug. 2018, pp. 1–6, doi: 10.1109/ICCUBEA.2018.8697439.
7. S. Saru and S. Subashree. Analysis and Prediction of Diabetes Using Machine Learning. Accessed: Oct. 22, 2022. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3368308
8. P. Sonar and K. JayaMalini, __Diabetes prediction using different machine learning approaches,‘‘ in Proc. 3rd Int. Conf. Comput. Methodologies Commun. (ICCMC), Mar. 2019, pp. 367–371, doi: 10.1109/ICCMC.2019.8819841.

BIOGRAPHIES

B.Surya Sai Kiran Akash is an undergraduate student in the Department of Computer Science and Engineering at Raghu Engineering College. With a keen interest in machine learning, data science, and healthcare technologies, he is passionate about applying computational methods to

address real-world challenges. His current research focuses on utilizing advanced algorithms for disease prediction and personalized healthcare solutions. Through his work, Surya Sai Kiran Akash aims to contribute to the development of intelligent systems that can enhance early disease detection, improve treatment planning, and support proactive healthcare management.