Diabetes Detection

Ambika K [1],Ms. Kajal Jha [2]
[1] *Student, 4th Semester MCA, Department of MCA, EWIT, Bangalore*
[2] *Assistant Professor, Department of MCA, EWIT, Bangalore*
[1] ambikakrish13@gmail.com
[2] Kajaljha707@gmail.com

Abstract— The prevalence of diabetes is increasing globally, emphasizing the urgent need for early detection and preventive healthcare solutions. This project presents a **Diabetes Detection System** that leverages machine learning models integrated into a web-based application for efficient risk prediction. The system allows users to register securely, input their health data (such as glucose levels, BMI, age, and family history), and obtain real-time predictions through models including Support Vector Machine (SVM), Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN). Built using Django for authentication and backend operations and Streamlit for an interactive dashboard, the application provides predictions along with model accuracy, feature importance, and graphical insights. Although the current implementation is limited to static datasets and single-user access, future enhancements aim to include real-time health data integration, advanced visualization, and improved scalability. This project contributes to the growing field of **data-driven healthcare** by combining technology with preventive medicine to promote timely intervention and better health outcomes.

Keywords— Diabetes Detection is an important area of research in healthcare and preventive medicine. The process involves identifying patients who are at risk of diabetes mellitus through various clinical, biological, and lifestyle parameters. Important factors such as blood glucose levels, body mass index (BMI), age, family history, and lifestyle choices are widely considered in prediction models.

I. INTRODUCTION

Diabetes is one of the most pressing global health challenges, affecting millions of people worldwide and leading to severe complications if not diagnosed and managed early. The growing prevalence of this chronic condition highlights the need for effective, accessible, and technology-driven solutions for early detection. Traditional healthcare methods are often

instant feedback, leaving many individuals unaware of their potential risk factors. To address this gap, the Diabetes Detection Project has been developed as a user-friendly platform that integrates machine learning models with modern web technologies. The system enables users to securely register, input their health data such as glucose level, BMI, age, and family history, and receive real-time predictions regarding their diabetes risk. By incorporating models like Support Vector Machine (SVM), Logistic Regression, Random Forest, and K- Nearest Neighbors (KNN), the project ensures comparative analysis and reliable outcomes.

The application is built using Django for backend management and Streamlit for an interactive dashboard, making the system intuitive and accessible to both general users and healthcare professionals. Overall, this project demonstrates how data-driven healthcare solutions can empower individuals with timely knowledge, encourage preventive measures, and support medical practitioners in patient assessment.

II. RELATED WORK

The application of machine learning in healthcare, particularly in diabetes prediction, has gained significant attention in recent years. Researchers have developed various models and frameworks to identify risk factors and improve early diagnosis.

- **Support Vector Machines (SVM):** Widely applied in diabetes classification tasks. Known for high accuracy in binary classification problems.
- **Logistic Regression (LR):** One of the earliest and most interpretable models used for diabetes prediction. Diabetes Dataset. Limitation struggles with non-linear and complex relationships.
- **Random Forest (RF):** Effective ensemble learning model, combining multiple decision trees. Provides feature importance insights, aiding medical interpretability.
- **K-Nearest Neighbors (KNN):** Simple and intuitive algorithm for classification tasks. Performs well with small datasets, but efficiency decreases with large datasets.

Compared to these works, the proposed system distinguishes itself by integrating multiple machine learning models within a web-based platform built using Django and Streamlit**.** This Combination not only provides accurate predictions but also ensures user-friendly interaction, secure authentication, and visual insights into model performance**.**

III. METHODOLOGY

The methodology of the Diabetes Detection project involves several steps. First, health data such as glucose level, BMI, blood pressure, insulin, age, and family history are collected from the dataset. The data is preprocessed by handling missing values, removing outliers, and standardizing features to ensure accurate model training. Next, four machine learning models are applied for prediction: Support Vector Machine (SVM), Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN). The dataset is divided into training and testing sets, and each model is trained to learn patterns that distinguish between diabetic and non-diabetic cases. The models are then evaluated using performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to ensure reliability. Among these, Random Forest and SVM generally provide higher accuracy, while Logistic Regression and KNN add interpretability. For implementation, Django is used for user registration, login, and database management, while Streamlit provides an interactive dashboard. Users can log in, input their health data, select a machine learning model, and instantly receive predictions

along with accuracy scores and feature importance. Finally, the system is tested through unit testing, integration testing, and user acceptance testing to confirm functionality, usability, and prediction reliability. This structured methodology ensures that the system is both technically sound and user-friendly. Methodology adopted for the Diabetes Detection project is designed to combine machine learning models with a web-based interactive system to predict the risk of diabetes. It consists of multiple stages including data preparation, model training, evaluation, implementation, and testing.

The system uses a diabetes dataset containing medical attributes such as glucose level, blood pressure, BMI, insulin level, skin thickness, age, pregnancies, and diabetes pedigree function. Before applying the models, the data undergoes preprocessing steps such as handling missing values, removing noise and outliers, and standardizing features using scaling techniques. This ensures that the input data is clean and consistent for training machine learning models. The dataset is then divided into training and testing subsets, typically with an 80:20 ratio, to evaluate performance. The fourth stage is system implementation**.** The backend is built using Django framework**,** which manages user authentication, registration, and data storage (SQLite database). The frontend is developed using Streamlit**,** providing an interactive dashboard where users can log in, input health parameters, select a model, and receive predictions in real-time. Along with prediction results, the system also displays model accuracy and feature importance to enhance user understanding.

IV. RESULTS AND DISCUSSION

The Diabetes Detection system was successfully implemented using four machine learning models—Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), and K-Nearest Neighbors (KNN)**.** The models were trained on the diabetes dataset and evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC**.**

1. Support Vector Machine (SVM)

SVM achieved good accuracy in distinguishing between diabetic and non-diabetic patients. It performed well with linear separation but required proper scaling of input features. The model was effective for binary classification tasks with moderate.

2. Logistic Regression (LR)

Logistic Regression provided simple and interpretable results with acceptable accuracy. It worked effectively on structured health data such as glucose and BMI values. The model offered transparency in understanding the impact of features. Its limitation lies in handling complex non-linear relationships within medical data.

3. Random Forest (RF)

Random Forest produced the highest accuracy among the models tested. It provided robust predictions and highlighted key features like glucose and BMI. The ensemble approach reduced the risk of overfitting and improved reliability. This makes RF the most suitable model for real- world healthcare applications.

4. K-Nearest Neighbors (KNN)

KNN gave reasonable accuracy but was sensitive to noise in the dataset. Its performance depended heavily on

the value of 'k' and data distribution. The algorithm worked better for small datasets with fewer features. For larger datasets, computation became slower and less efficient.

5. System Performance and Usability

The integration of Django and Streamlit offered a secure and interactive platform. Users could easily log in, input health data, and get real-time predictions. Visual outputs such as confusion matrices and ROC curves improved clarity. Overall, the system balanced technical accuracy with user-friendliness.

6. Discussion and Insights

The results show that machine learning is effective for early diabetes detection. Random Forest emerged as the most accurate and reliable model. ZFuture enhancements should focus on real-time data integration and scalability.

Discussion

Diabetes is a chronic disease that arises when the body cannot effectively regulate blood sugar levels. It is becoming one of the fastest-growing global health concerns, leading to severe complications such as heart disease, kidney failure, blindness, and nerve damage if not detected early. The major challenge is that diabetes often develops silently, with many individuals remaining undiagnosed until serious health issues appear. This highlights the importance of early detection and preventive healthcare approaches**.** Traditional diagnostic methods, such as blood tests and clinical examinations, are accurate but often time- consuming, require laboratory facilities, and may not be accessible to everyone. With the growth of artificial intelligence and machine learning, new opportunities have emerged to detect diabetes at an earlier stage using predictive models. These models analyze health parameters such as glucose level, BMI, blood pressure, age, and family history to identify individuals at risk.

The use of machine learning in diabetes detection offers several advantages. It enables quick predictions, handles large volumes of health data, and provides insights into key risk factors. Models like Support Vector Machine (SVM), Logistic Regression, Random Forest, and K- Nearest Neighbors (KNN) have shown significant accuracy in predicting diabetes. In addition, advanced deep learning methods can capture more complex relationships, further improving performance. However, challenges still exist. The reliability of predictions depends heavily on the quality and diversity of datasets**,** as limited data may reduce model generalization.

V. CONCLUSION

The Diabetes Detection project successfully demonstrates the integration of machine learning models with a web-based application for early identification of diabetes risk. By using models such as Support Vector Machine (SVM), Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN), the system provides accurate predictions based on health parameters like glucose level, BMI, blood pressure, and age. Among

these, Random Forest proved to be the most reliable, offering both high accuracy and interpretability through feature importance.

The implementation using Django for backend authentication and Streamlit for the interactive dashboard created a secure and user-friendly environment. Users are able to register, log in, input health data, and receive instant predictions along with visual insights into model performance. This enhances accessibility for individuals seeking self-assessment as well as healthcare professionals looking for supportive diagnostic tools. The project highlights the potential of AI-driven healthcare solutions in improving early detection and prevention of chronic diseases. While the system achieves its objectives, certain limitations remain, such as reliance on a static dataset and manual data entry. Future improvements could include real- time integration with wearable devices, larger datasets, advanced deep learning models, and multi-user scalability. In conclusion, this project serves as a valuable step towards personalized, data-driven healthcare**,** empowering users with timely insights and supporting preventive action to reduce the long-term burden of diabetes.

REFERENCES

- Django Software Foundation. (2024). *Django documentation*. Retrieved from https://docs.djangoproject.com/en/stable/
- Streamlit Inc. (2024). *Streamlit documentation*. Retrieved from https://docs.streamlit.io/
- SQLite Consortium. (2024). *SQLite Python documentation*. Retrieved from https://docs.python.org/3/library/sqlite3.html
- Code With Harry. (2023). *Django tutorials for beginners* [YouTube channel]. Retrieved from https://youtu.be/JxzZxdht-XY
- World Health Organization (WHO). (2023). *Diabetes fact sheet*. Retrieved from https://www.who.int/news-room/fact-sheets/detail/diabetes
- American Diabetes Association. (2023). *Standards of medical care in diabetes*. Diabetes Care, 46(Supplement 1), S1-S154.
- 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.
- Sisodia, D., & Sisodia, D. S. (2018). *Prediction of diabetes using classification algorithms*. *Procedia Computer Science*, 132, 1578–1585.