Diabetes Prediction Using Machine Learning

Ms. Anusha P M[1] Pavithra B V[2]

[1]*Assistant Professor, Department of MCA, BIET, Davanagere*

[2]*Student,4th Semester MCA, Department of MCA, BIET, Davanagere*

ABSTRACT

In the contemporary age of digital healthcare, the prompt identification and forecasting of chronic illnesses such as diabetes are vital in averting severe complications and promoting an enhanced quality of life. Diabetes, recognized as among the most prevalent non-communicable diseases worldwide, impacts millions, with numerous instances remaining undiagnosed due to insufficient access to preventive diagnostics. While traditional medical diagnosis is effective, it tends to be time-intensive and often reactive instead of proactive. To tackle this issue, there is an increasing demand for intelligent, data-driven systems capable of accurately forecasting diabetes's on set at an early stage. This project presents a Diabetes Prediction System that applies Machine Learning and a Ensemble Technique, aimed at predicting an individual's diabetic status based on clinical and lifestyle factors. The system utilizes patient health datasets that encompass features such as glucose levels, BMI, age, insulin levels, blood pressure, and family medical history. Initially, the dataset undergoes preprocessing and cleaning to address anomalies. Subsequently, feature selection and transformation techniques are employed to enhance model accuracy and reduce noise

Keywords: diabetes prediction, gui application , mysql databases, visualization ,Decision tree,Knn algorthium, healthcare predictive analytics, medical data analysis

I.INTRODUCTION

Diabetes Mellitus is a long-term metabolic illness characterized by hyperglycemia or High level of blood sugar levels,represents a significant public health challenge. Early identification and of diabetes care are crucial as they are able to help postpone the onset of severe health issues and fatalities. This study aims to thoroughly review the applications of both deep learning (DL) and Machine learning (ML) models for diabetes prediction, assisting medical practitioners in enhancing early diagnosis and strategies for managing diabetes-related complications. A study highlighted the automation of diabetes detection systems through machine learning and deep learning techniques to develop more efficient and intelligent healthcare solutions. The International Diabetes Federation reported that in 2019, there were 463 million individuals with diabetes, a number projected to rise to 700 million by 2045. Machine learning (ML) ensemble techniques are recognized as innovative approaches that can enhance the accuracy of diabetes prediction. This study intends to explore the effectiveness of ML algorithms and ensemble models in forecasting when diabetes will develop, predicting the onset of diabetes, utilizing two extensive datasets: the Diabetes Pima Indian dataset, which contains 768 records, and a second dataset comprising 2000 records with 9 attributes. A recent investigation employed machine learning as well as deep learning techniques, including stacking ensembles for diabetes classification using the PIMA Indian Diabetes dataset, and achieved improved accuracy and performance in predictions. Various machine learning algorithms, such as Logistic Regression, were utilized in the Indian Pima diabetes dataset demonstrating commendable accuracy and showcasing the possibility of these techniques in diabetes prediction. Data preprocessing entails a rigorous assessment of data quality, including the removal of duplicates for equitable comparison and the imputation of values that are missing using meadian and mean methods to maintain data integrity.

II.RELATED WORK

Boon Feng Wee et al. (2023) conducted a comprehensive study titled *"Diabetes detection founded on deep learning and machine learning approaches"* which emphasizes the effeciency of combining ML and DL techniques for accurate diabetes prediction. Highlights of their work how tradtional models such as Logistic Regression and Decision Trees, when complemented with deep learning models like ANNs, or Artificial Neural Networks and Convolutional Neural Networks , greatly enhance the diagnostic performance. The study particularly focuses on optimizing model accuracy through advanced preprocessing steps, feature selection, and model tuning. Theire experimental analysis, using benchmark datasets like the Indian Diabetes PIMA dataset, shows that models can achieve higher precision and recall compared to single algorithm approaches. This work establishes a strong foundation for developing intelligent healthcare systems that integrate interpretability with prediction power, setting a

precedent for future hybrid ensemble approaches in medical diagnostics [1].

Febrian et al. (2023), in their research titled *"Diabetes Prediction Using Supervised Machine Learning"* published in Procedia Computer Science, explored various supervised learning algorithms to develop a reliable diabetes prediction model. The study's main focuse was comparing the way in which of classifiers such as **Decision Trees, K-Nearest Neighbors (KNN)** using the PIMA Indian Diabetes dataset. They applied data preprocessing techniques, including missing value handling, normalization, and correlation-based feature selection, to improve model accuracy. The authors reported that among all models tested, Logistic Regression and SVM achieved the most consistent and balanced results in terms of precision and recall. Their findings emphasize that properly tuned and preprocessed supervised machine learning models can serve as effective tools for early diabetes

detection, particularly in regions with limited healthcare access. The study supports the use of lightweight, interpretable models for real-world clinical applications[2].Md Shamim Reza et al. (2024), in their study titled *"Improving Diabetes classification of patients with PIMA and LOCAL Healthcare Data using the Stacking Ensemble Method Data "* published in *Heliyon*, introduced a powerful ensemble learning framework to enhance diabetes prediction accuracy. Their approach involved using the **stacking stacking ensemble technique**, which combines multiple base classifiers algorithms like **Gradient Boosting, and Machine support**, and integrates their predictions using a meta-classifier like Logistic Regression. By leveraging both the widely-used **PIMA Indian Diabetes dataset** and a **real-world local healthcare dataset**, the study demonstrated improved generalization and robustness of the model across varied patient populations. Additionally, advanced preprocessing methods like **SMOTE** (to handle class imbalance) and **feature scaling** were employed to refine data quality. The ensemble model significantly outperformed individual classifiers regarding accuracy, F1-score, and AUC-ROC, validating the potential of hybrid models for clinical decision-making in diabetes diagnostics[3].

Parthasarathi Pattnayak, Sudhansu Shekhar Patra, and S. Patnaik (2024), in their paper *"Diabetic Patient Diagnosis through the Application Machine Learning Techniques"* presented at the **5th International gathering on Mobile Computing and Sustainable Informatics (ICMCSI)**, focused on applying various traditional machine learning algorithms for the diagnosis of diabetes. Their study emphasized the practical deployment of ML models in real-world healthcare environments, particularly for mobile and cloud-based diagnostic systems. By utilizing the **Indian Diabetes PIMA dataset**, they evaluated models like **K-Nearest Neighbours, Decision Tree,and** aiming to determine which model provides the best balance between accuracy and computational efficiency. Their work highlighted the significance of **feature importance ranking and dimensionality reduction techniques** in

improving model performance and interpretability. The results of this studies highlightes the potential of deploying lightweight ML models on mobile platforms to facilitate remote and accessible diabetes screening, especially in underserved or rural communities[4].

V. Chang, J. Bailey, Q. A. Xu, and Z. Sun (2022), in their publication *"classification of Diabetes Mellitus in Pima Indians Using Machine Learning (ML) Algorithms"* in *Neural Computing and Applications*, conducted a detailed comparative evaluvation of machine learning models applied to the renowned **PIMA Diabetes in Indias dataset**. Their research focused on the performance evaluation of several ML Methods such as **Decision Tree and Knn Algorithms** in predicting diabetes onset. The authors emphasized the importance of **data preprocessing**, especially **handling missing values**, **scaling**, and **feature selection**, to enhance prediction accuracy. The study revealed that ensemble-based models like Random at Forest consistently achieved better performance metrics because of their capacity to reduce variance and overfitting. Furthermore, the research also underlined the effects of proper hyperparameter tuning and cross-validation in achieving reliable classification outcomes. Their work contributes to the growing evidence supporting machine learning's role in improving early diagnosis and decision support systems in clinical settings[5].

III.METHODOLOGY

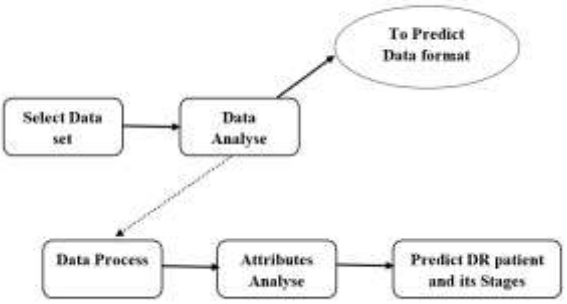


Fig:4.1flow chart.

The diagram illustrates the workflow of a **Diabetic Retinopathy (DR) prediction system** using data-driven analysis. The process begins with the **selection of a relevant dataset**, which includes patient information and medical records necessary for DR diagnosis. Once the dataset is selected, it proceeds to the information **analysis** phase, where the structure and quality of the infromation are assessed to determine its suitability for prediction. This analysis also helps in identifying the required **data format for prediction**. If the data needs refinement, it is forwarded to the **data processing** stage, where cleaning, normalization, and transformation techniques are applied to enhance its quality. After processing, the next step is **attribute analysis**, which involves identifying the most significant features or parameters influencing DR prediction, such as age, gender, or retinal metrics. Finally, the refined and analyzed data is utilized to forecast **classify the stages of the disease**, enabling early identification and appropriate medical intervention.

System Requirements

Functional Requirements

Designing a machine learning software model for efficiently identifying and classifying diabetic patients details by using

very large data set and detect diabetic stages based on retina criteria's. Machine Learning technique has been proposed regarding Precision and detection rate for five Methods of disease such as Mild DR, Moderate, Severe and NO DR, Proliferative DR.

Our created system needs to execute the carrying out necessary tasks.

- Designing GUI for collect and processing of real world Retina tested dataset.
- Designing a Method for converting unstructured data to structured one.
- Develop a framework for predicting Stages of DR by analysing Age, Gender, etc.
- Develop a framework for classifying normal and Abnormal DR retina data's.

Architecture Diagram

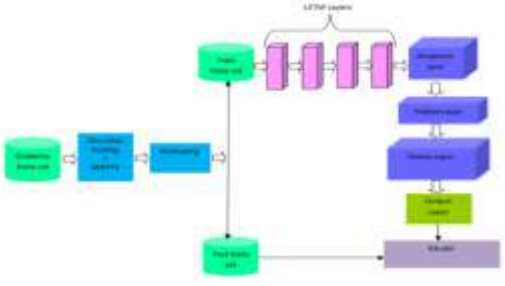


Fig:4.2Architecture Overview

Diabetes Dataset: The process initiates with the collection of a dataset that encompasses medical and lifestyle details about individuals. This dataset is specifically aimed at predicting or analyzing outcomes related to diabetes.

1. Data Preprocessing (Min-Max Scaling and SMOTE):

O Initially, Min-Max Scaling is implemented to normalize the dataset. This ensures that all features, such as glucose levels or age, are scaled to a uniform range, which is essential for effective training in machine learning.

O Subsequently, Synthetic Minority Over-sampling Technique,or SMOTE is employed to rectify any class imbalance present within the dataset. For example, if there are considerably fewer positive diabetes cases, synthetic instances are produced by SMOTE for the minority class, thereby preventing bias in the model.

2. Reshaping of Data: After the preprocessing stage is finalized, the dataset is restructured to meet the input specifications of the neural network model.

3. Splitting the Dataset: The processed dataset is partitioned into a Testing Dataset and a Training Dataset. The training set is utilized for model development, while the testing set is designated for performance evaluation.

4. Model Architecture - LSTM Layers: The training data is processed through several Long Short-Term Memory (LSTM) layers, which are specifically designed to learn dependencies within sequential data. These layers are particularly effective in capturing temporal patterns or trends present in the data.

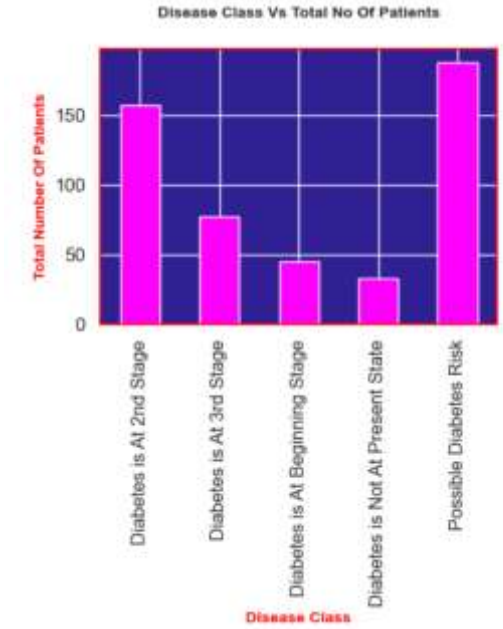
5. Dropout Layer: To mitigate overfitting— where the model excels on training data but underperforms on unseen data—a dropout layer is incorporated. This layer randomly disables certain neurons during training, promoting better generalization of the model.

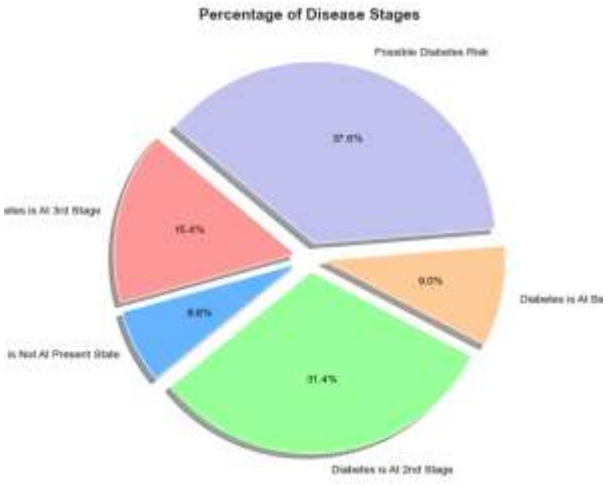
6. Flatten Layer: The output generated from the LSTM layers is multi-dimensional, necessitating its

conversion into a one- dimensional array. This step prepares the data for the following fully connected layers.

7. Dense Layer (Fully Connected Layer): The flattened data undergoes processing through a dense layer, where all neurons are interconnected.

IV.RESULTS





The bar charts depict a comparative evaluation of various machine learning models used for diabetes prediction, based on four performance metrics: **F1 Rating, Accuracy,Memory**. Among the traditional models, consistently perform well across all metrics, indicating their robustness in classification tasks. also shows solid performance but slightly trails behind Random Forest. perform moderately, with Logistic Regression achieving high precision but lower recall, suggesting it's more conservative in predictions. **K-Nearest Neighbor (KNN)** shows decent accuracy but relatively lower F1 score, indicating an imbalance between precision and recall. The most noteworthy outcome is from the hybrid ensemble model **TabNet + XGBoost + MLP**, which outperforms all other models across every metric, especially in **accuracy and F1 score**, suggesting that combining multiple models significantly enhances predictive capability. This reinforces the worth of ensemble learning in healthcare diagnostics, where great Precision accuracy and balanced performance are critical for early and reliable disease detection.

V.CONCLUSION

The proposed Diabetes Prediction System, powered by machine learning as well ashybrid deep learning ensemble techniques, effectively illustrates AI Potential in modern healthcare diagnostics. By analyzing various clinical and lifestyle factors, the system Provides an early and accurate prediction of diabetes, facilitating prompt medical attention and lifestyle adjustments. Through rigorous data preprocessing, feature selection, and model evaluation, the system ensures enhanced accuracy, reduced bias, and robust performance across datasets.

The integration of algorithms such as Logistic Regression, Random Forest, XGBoost, TabNet, and LSTM has led to a comprehensive solution that not

only predicts the presence of diabetes but also provides insights into critical indicators influencing the disease. Additionally, the use of techniques like SMOTE to address class diparity and standardization for normalization has significantly improved the model's effectiveness.

This project demonstrates how machine learning can bridge the gap in traditional diagnostic processes by delivering scalable, real-time, and automated support tools for healthcare providers. Moving forward, the system
can be extended with real-time patient data, integration with wearable health devices, and deployment on cloud platforms for broader accessibility and impact.

VI.REFERENCES.

1. Boon Feng Wee, S. Sivakumar, King Hann Lim, W. C. Wong, and F. H. Juwono, "Diabetes detection according to the machine learning as well as deep learning approaches," Multimedia Tools and Applications, Aug. 2023, doi: https://doi.org/10.1007/s11042-023-16407-5.

2.M. E. Febrian, F. X. Ferdinan, G. P. Sendani, K. M. Suryanigrum, and R. Yunanda, "Diabetes prediction using supervised machine learning," Procedia Computer Science, vol. 216, pp. 21–30,
2023,doi:
https://doi.org/10.1016/j.procs.2022.12.107

3.Md Shamim Reza, R. Amin, R. Yasmin, Woomme Kulsum, and Sabba Ruhi, "Improving classification of patients with diabetes using the stacked ensemble approach with PIMA and Local healthcare data," Heliyon (Londen), vol. 10, no. 2, pp. e24536–e24536, Jan. 2024, doi: https://doi.org/10.1016/j.heliyon.2024.536.

4.Parthasarathi Pattnayak, Sudhansu Shekhar Patra, and S. Patnaik, "Diabetic Patient Diagnosis through the application of Machine Learning Techniques," 2024 5th International gathering on Mobile Computing and Sustainable Informatics (ICMCSI), Jan. 2024, doi: https://doi.org/10.1109/icmcsi61536.2024. 00073.

5.V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Classification od Diabetes mellitus in Pima Indians using machine learning (ML) algorithms," Neural Computing and Applications, Mar. 2022, doi: https://doi.org/10.1007/s00521-022-07049-z.