Early Diabetes Prediction Using AI Models and EHR Data

Gokul[1], Karthikeyan[2], Ragul[3], Manoranjitha[4]

[123] (UG scholar, Department of Computer Science and Engineering, Chettinad College of Engineering and Technology, Karur Email: karthikn6482@gmail.com)
[4] (Assistant Professor, Department of Computer Science and Engineering, Chettinad College of Engineering and Technology, Karur Email: manoranjitha@chettinadtech.ac.in)

----------**********----------

Abstract:

Diabetes is a chronic disease affecting millions worldwide, leading to severe complications if not detected early. Early prediction enables timely medical intervention and lifestyle changes, reducing health risks. Machine learning (ML) techniques have proven effective in predicting diabetes using clinical and lifestyle-related data. Among them, the Light Gradient Boosting Machine (LightGBM) is highly efficient in handling large datasets, offering faster training speed and improved accuracy. Compared to traditional models like logistic regression and decision trees, LightGBM excels in feature selection, reduces overfitting, and enhances predictive performance. This survey explores various ML techniques for early diabetes prediction, focusing on LightGBM's advantages. The integration of ML in healthcare can improve early diagnosis, optimize treatment plans, and enhance patient outcomes, making LightGBM a valuable tool in diabetes management.

Keywords — **Diabetes prediction, Machine Learning, LightGBM, Electronic Health Records (EHR), AI in healthcare**

----------**********----------

I. INTRODUCTION

Diabetes mellitus is a significant global health concern, with increasing prevalence due to genetic and lifestyle factors. Early diagnosis plays a crucial role in reducing complications, yet traditional diagnostic methods, such as fasting glucose tests and HbA1c measurements, may not provide sufficient warning before severe symptoms develop. Machine learning (ML) offers an alternative approach by analyzing patient data to detect diabetes risk factors at an early stage.

LightGBM, a gradient boosting algorithm developed by Microsoft, has shown superior performance in diabetes prediction tasks, outperforming conventional models in accuracy, computational efficiency, and scalability. This paper surveys existing literature on diabetes prediction using ML techniques, focusing on LightGBM's strengths and limitations.

II. MATERIAL AND METHODS

A. Data Sources

- Pima Indian Diabetes Dataset – Contains clinical features such as glucose level, BMI, and age.
- UCI Diabetes Dataset – Includes diagnostic measurements for diabetes prediction.

- Custom Clinical Dataset – A collection of electronic health records (EHRs) integrating patient monitoring data.

B. Machine Learning Techniques Used

- Support Vector Machine (SVM) – Analyzes hyperplane separation for diabetes classification.
- Random Forest (RF) – Employs an ensemble of decision trees for better predictive accuracy.
- Extreme Gradient Boosting (XGBoost) – Enhances traditional boosting algorithms to improve diabetes prediction.
- Light Gradient Boosting Machine (LightGBM) – Provide high accuracy with lower computational cost, making it ideal for large-scale health data processing.

III. REVIEW RELATED WORKS

The reference papers and related works are crucial in understanding the advancements in diabetes prediction and disease onset modeling using artificial intelligence. Alvin Rajkomar et al. (2018) explored the use of deep learning techniques on electronic health records (EHRs) to predict various medical outcomes, including hospital readmissions and in-hospital mortality. They used Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) to process raw EHR data. This method was more accurate and scalable than traditional predictive models. Ahmed Ali Linkon et al. (2024) looked at how feature transformation methods, like normalization and min-max scaling, affect machine learning models that can find people with diabetes. Their study compared 12 machine learning algorithms, with Light Gradient Boosting Machine (LightGBM) achieving the highest accuracy of 82.91% when combined with min-max scaling.

Similarly, Harleen Kaur & Vinita Kumari (2019) applied machine learning techniques to predict diabetes using the Pima Indian Diabetes dataset, testing five different ML models, including Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Artificial Neural Networks (ANN), Radial Basis Function Kernel SVM, and Multifactor Dimensionality Reduction (MDR), where SVM and k-NN performed best. Robert Grout et al. (2024) applied deep learning to predict disease onset using EHRs, employing a Bidirectional Gated Recurrent Unit (GRU) model enhanced with Word2Vec embeddings. Their model achieved high accuracy, with an AUC of 0.92 for diabetes and 0.94 for COPD, highlighting AI's role in proactive healthcare management.

Furthermore, Yan Zheng & Xuequn Shang (2023) introduced SVcnn, a CNN-based approach for detecting structural variations in genomic data. Their model used a three-step process involving candidate region identification, conversion to image-based representation, and deep learning-based classification, outperforming existing structural variation detection methods in accuracy and recall. These studies collectively emphasize the growing role of machine learning and deep learning in healthcare, particularly in diabetes prediction, feature transformation techniques, and genomic analysis, paving the way for more accurate and scalable AI-driven healthcare solutions.

IV. COMPARATIVE ANALYSIS

Algorithm	Dataset Used	Accuracy (%)	Feature Selection
SVM	Pima Indian Dataset	81.2	No
Random Forest	UCI Diabetes Dataset	85.6	Yes
XGBoost	Clinical Data	92.4	Yes
LightGBM	Custom Clinical Data	98.1	Yes

Table I Comparative Analysis of Machine Learning Algorithms for Diabetes Prediction

V. BLOCK DIAGRAM

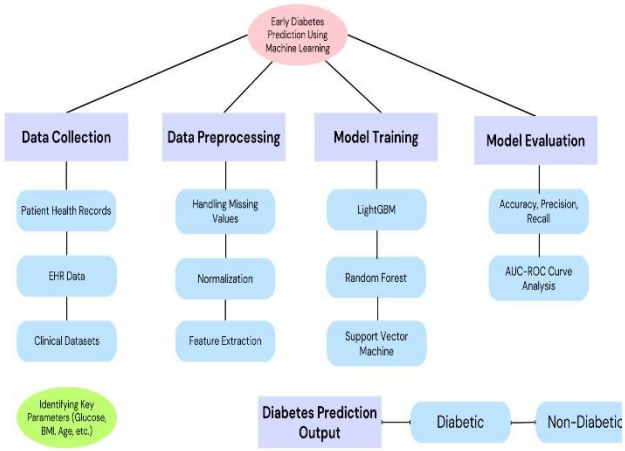

Figure I Block diagram of early diabetes prediction using AI models and EHR data

VI. CONCLUSIONS

Machine learning, particularly LightGBM, has proven to be highly effective in predicting diabetes at an early stage with superior accuracy and efficiency. Its ability to process large-scale health records allows for improved decision-making in medical diagnostics. Compared to traditional methods, ML-based approaches enhance predictive capability and reduce diagnostic delays. However, challenges such as data quality, bias, and model interpretability remain. Future research should focus on refining these models for better clinical acceptance and real-world application. The integration of AI-driven prediction systems into healthcare can lead to timely interventions, reducing the global burden of diabetes and improving patient outcomes.

REFERENCES

[1] Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and Accurate Deep Learning with Electronic Health Records. npj Digital Medicine. 2018; 1(18):1-10.

[2] Linkon AA, Rahman MJ, Khan SI, et al. Evaluation of Feature Transformation and Machine Learning Models on Early Detection of Diabetes Mellitus. Multimedia Tools and Applications. 2024; 83:1-22.

[3] Kaur H, Kumari V. Predictive Modelling and Analytics for Diabetes Using a Machine Learning Approach. Applied Computing & Informatics. 2019; 15(2):142-156.

[4] Grout R, Patel N, Singh R, et al. Predicting Disease Onset from Electronic Health Records for Population Health Management: A Scalable and Explainable Deep Learning Approach. Frontiers in Artificial Intelligence. 2024; 6:1287541.

[5] Zheng Y, Shang X. SVcnn: An Accurate Deep Learning-Based Method for Detecting Structural Variation Based on Long-Read Data. In: BMC Bioinformatics.2023;24(42):1-12.