National Conference on Recent Trends in Artificial Intelligence and Data Science (NCRTAIDS-26)
Sri Vasavi College, Erode, Tamil Nadu, India

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

A Survey on Machine Learning Techniques for Diabetic Type Classification

[1]Dr.P.Suresh Babu, [2]Ms.M.Premavathi

[1]Assistant Professor Department of Computer Science Sri Vasavi College (Self FinanceWing), Erode
[2]Assistant Professor and Head Department of PG Computer Science Sri Vasavi College (Self FinanceWing), Erode

Abstract - Medical data mining analyzes health data to improve patient care, especially in diabetes management—a chronic disorder affecting over 500 million people, risking eyes, kidneys, heart, and nerves due to poor glucose regulation. It processes datasets like Pima Indians Diabetes for early prediction, classifying Type 1 and Type 2 based on glucose levels and family history. Key steps include data collection from records, secure storage, pre-processing (imputation, outlier removal, normalization, encoding, oversampling for imbalance), and modeling with ML (Random Forest, SVM) or DL (DNN). Optimized pipelines achieve up to 97% accuracy, outperforming traditional methods in speed and precision via imputation, tuning, and ensembles. Recent innovations reduce complexity and enable scalable diagnosis on diverse datasets.

Keywords - Medical data mining, diabetes, intervention and management, disease classification and prediction.

I. INTRODUCTION

Data mining on big data transforms healthcare, especially for diabetes—a common chronic condition in the elderly. The International Diabetes Federation reported 451 million global cases in 2017. This metabolic disorder stems from poor blood glucose regulation, via pancreatic issues (type 1) or insulin resistance (type 2). It requires lifelong management with treatments, meds, and lifestyle changes, often causing complications like cardiovascular disease, neuropathy, kidney failure, and vision loss. Risk factors: age, obesity, inactivity, genetics, poor diet, hypertension, dyslipidemia. Diabetics face higher comorbidity risks, highlighting predictive analytics' value.

II. LITERATURE SURVEY

Early detection and classification of diabetes types help to treat the patients with required treatments and prevent the complications which help in minimizing the risk of severe health problems. An optimized Light Gradient-Boosting Machine (Light GBM) and K-Nearest Neighbor (KNN) based ensemble algorithm was introduced in [1] for Type 2 Diabetic prediction. But the classification accuracy was not enhanced by Light GBM and KNN based ensemble algorithm. A data-driven approach was introduced in [2] to forecast the blood glucose level of Type 2 diabetes patients. But the prediction time was not reduced by data-driven approach. The machine learning algorithm was introduced in [3] with statistical analysis for normal and diabetic cases to distinguish influential features. Stacking ensemble model was introduced in [4] for diabetes risk identification at earlier stage. But the ensemble model failed to reduce the time consumption. Normalization and classifiers like Light GBM (LGBM), Gradient Boosting (GB), and Random Forest (RF) [5] was introduced for improved early diagnosis and classification of diabetic type with more accuracy but classification time was high. Fast Gradient Sign Method (FGSM) [6] was introduced to keep medical data private and for refining how we classify diabetes but also defends from intentionally manipulating inputted data that can cause incorrect decisions.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

III. DATASET DESCRIPTION

The first dataset used is the diabetes dataset. The URL of the dataset is given as https://www.kaggle.com/datasets/mathchi/diabetes-data-set. The dataset comprised the set of medical and demographic data of every patient. Every patient data has been classified whether Type 1 or Type 2 based on some attributes like Gender, Age, Insulin resistance, HbA1c_level, diabetes etc. The diabetes dataset contains the 95 features and 10000 data samples.

The second dataset used is data collected from laboratory of Medical City Hospital. The URL of the dataset is given as https://data.mendeley.com/datasets/wj9rwkp9c2/1. Patients' files were taken and data extracted from them and entered in to the database to construct the diabetes dataset. The data consist of medical information, laboratory analysis. The data attribute is: No. of Patient, Sugar Level Blood, Age, Gender, Creatinine ratio (Cr), Body Mass Index (BMI), Urea, Cholesterol (Chol), Fasting lipid profile, including total, LDL, VLDL, Triglycerides (TG) and HDL Cholesterol, HBA1C, Class (the patient's diabetes disease class may be Diabetic, Non-Diabetic, or Predict-Diabetic).

IV. METHODOLOGY

Diabetes mellitus is a major health issue affecting millions, characterized by factors such as age, obesity, lack of exercise, poor lifestyle, unhealthy diet, and high blood pressure. It significantly raises the risk of complications like heart disease, kidney failure, stroke, eye disorders, and nerve damage. Effective diagnosis and treatment are crucial, and big data analytics helps uncover hidden patterns in the data to predict diabetes outcomes accurately.

Gradient Boosting Machines (XGBoost, LightGBM, CatBoost) and Random Forest based data classification:

Big data processes vast information for better future predictions. In healthcare, analysts collect, integrate, and analyze diverse sources to improve patient outcomes. Accurate diagnosis relies on symptoms and signs. Gradient Boosting Machines (XGBoost, LightGBM, CatBoost) and Random Forest enable effective diabetes diagnosis and classification. GBMs build models sequentially to fix prior errors, often outperforming Random Forest (which uses independent bagging trees) in accuracy.

A Decision-Making Algorithm Approach (CART):

Decision Trees are a core supervised ML algorithm mimicking human decisions via flowchart structures. Used for classification and prediction, they recursively split datasets into homogeneous subsets based on features. They support Classification Trees (categorizing data) and Regression Trees (continuous outcomes like patient values). CART (Classification and Regression Trees) encompasses both. CART starts at the Root Node, selects the best feature using metrics like GINI Impurity or Information Gain for purest splits, branches by feature values, and repeats until optimal classification.

Multi-Model Ensemble Learning Approach

This study presents a multi-modal ensemble learning method integrating ANN, Random Forest (RF), and SVM for precise, reliable diabetes prediction. It employs Simple Ensemble Mean (SEM) and Weighted Ensemble Mean (WEM) for aggregation. The model combines RF, Extra Trees, and MLP to classify individuals as diabetic, non-diabetic, or pre-diabetic via prediction scores. Overall, it enhances accuracy, especially for diabetic cases.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Machine learning with AI technique

Diabetes, a top non-communicable disease, affected 537 million worldwide. Risks: excess weight, high cholesterol, family history, inactivity, poor diet. Complications: heart disease, kidney failure, nerve damage, retinopathy. Researchers built an automatic prediction system for Bangladeshi female patients using Pima Indian Diabetes Dataset (203 textile factory workers). It features semi-supervised Extreme Gradient Boosting for insulin prediction, plus SMOTE/ADASYN for class imbalance. Classifiers: Decision Tree, SVM, Random Forest, Logistic Regression, KNN, ensembles.

Performance Analysis

An empirical study compares four techniques—Gradient Boosting Machines (XGBoost, LightGBM, CatBoost), Random Forest (CART), Multi-Model Ensemble Learning, and Machine Learning with AI—on a medical dataset for accurate diagnosis and classification. Performance is evaluated using five metrics: diabetic diagnosis accuracy, precision, recall, F-score (efficiency), and diagnosis time. The performance is analyzed and calculated using the following functions:

- Significance of Precision: Calculated as, PE = TP/ TP + FP
- Significance of Accuracy: Calculated as, DDA =[TP+TN/TP+TN+FP+FN]*100
- Significance of Recall: Calculated as, RE = TP/TP+FN
- Significance of F1 Score: Calculated as, F1 Score = 2* [Recall * Precision/ Recall + Precision]
- Significance of Diagnosis Time: Calculated as, DDT = Sum of Ri * Time (DP)

Result and Future Work

This survey reviews ML for diabetes classification: traditional (SVM, Decision Trees, Random Forest, Logistic Regression) offer interpretable baselines; ensembles (XGBoost, LightGBM) and deep learning (MLP, DBN) excel in precision/recall/F1 on imbalanced data via SMOTE/ADASYN. Multi-modal ensembles (Random Forest, Extra Trees, and MLP) shine for diabetic/pre-diabetic/non-diabetic cases, applied to Bangladesh textile workers and Pima datasets. In future, XAI integration federated learning for privacy, longitudinal pre-diabetes studies, hybrid neuro-symbolic models for accuracy and trust.

TABLE II – Tabulation to Show Increased Accuracy, Precision ands Recall for Diabetic Dataset Using Proposed Techniques

Methods/ Parameters	Classification Accuracy	Precision	Recall	Classification Time (ms)
Gradient Boosting Machines (XGBoost, LightGBM, CatBoost) and Random Forest based data classification	97	96.1	96.8	32
A Decision Making Algorithm Approach (CART)	82.7	85	86.4	48
Multi-Model Ensemble Learning Approach	96.6	97	97.2	57
Machine learning with AI technique	91	92.2	92.8	55

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Table III – Tabulation to Show Increased Accuracy, Precision ands Recall for Medical City Hospital Dataset Using Proposed Techniques

Methods/ Parameters	Classification Accuracy	Precision	Recall	Classification Time (ms)
Gradient Boosting Machines (XGBoost, LightGBM, CatBoost) and Random Forest based data classification	96.4	97.2	98.8	29
A Decision Making Algorithm Approach (CART)	82.7	86.4	87.1	37
Multi-Model Ensemble Learning Approach	97.1	94	96.3	48
Machine learning with AI technique	92	93.3	97	53

Table II and Table III describe the performance results for eight different diabetes diagnosis methods with four performance metrics for two different datasets respectively. From above table, it is clear that Gradient Boosting Machines (XGBoost, LightGBM, CatBoost) and Random Forest based Data Classification technique gives best classification results than any other methods. For Diabetes Disease Prediction Dataset, the proposed classification technique attained 97% of diabetes classification accuracy, 96.1% of precision, 96.8% of recall and 29ms of classification time.

For Medical City Hospital Dataset, the proposed classification technique attained 96.4% of diabetes classification accuracy, 97.2% of precision, 98.8% of recall and 29ms of diabetes classification time. Figure 2 and Figure 3 shows the performance metric analysis of Diabetes Dataset and Medical City Hospital Dataset respectively.

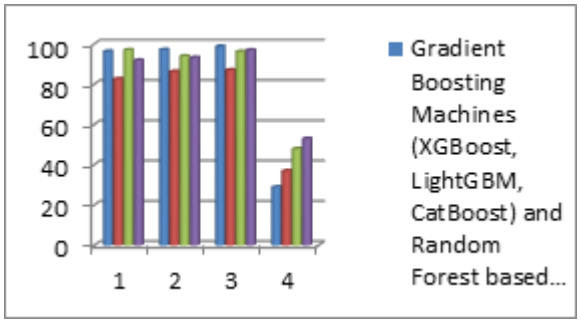

Fig. 2. Measurement Analysis Diabetic Dataset

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

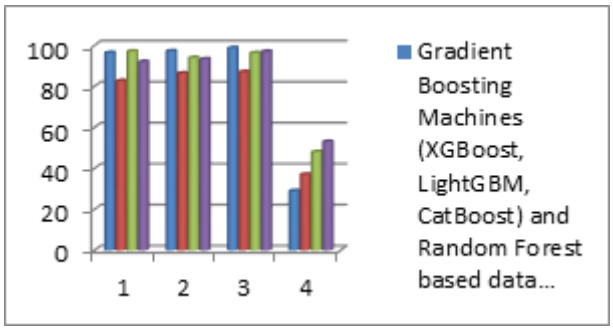

Fig. 2. Measurement Analysis of MCH Dataset

V. CONCLUSION

Machine learning techniques have demonstrated significant potential in improving the accuracy and efficiency of diabetic type classification. This survey highlights how traditional statistical approaches are increasingly being complemented or replaced by advanced machine learning algorithms such as Support Vector Machines (SVM), Decision Trees, Random Forests, k-Nearest Neighbors (k-NN), Naïve Bayes, and Artificial Neural Networks (ANNs). More recently, deep learning models have further enhanced classification performance, particularly when large and complex datasets are available.

The comparative analysis of these techniques shows that ensemble and deep learning methods often achieve higher accuracy due to their ability to capture complex nonlinear relationships within clinical and demographic data. However, simpler models still remain valuable in healthcare settings because of their interpretability, lower computational requirements, and ease of implementation.

Despite promising results, several challenges remain, including data imbalance, limited dataset sizes, feature selection issues, lack of model generalizability across populations, and concerns regarding privacy and ethical data usage. Moreover, clinical deployment requires models that are not only accurate but also transparent and explainable to healthcare professionals.

In conclusion, machine learning provides a powerful framework for diabetic type classification and early diagnosis support. Future research should focus on integrating multi-source medical data, improving explainability, enhancing model robustness, and conducting large-scale real-world validations to ensure reliable clinical adoption.

REFERENCES

1. Nur Farahaina Idris,Mohd Arfian Ismail ,Mohd Izham Mohd Jaya,Ashraf Osman Ibrahim,Anas W. Abulfaraj,Faisal Binzagr, "Stacking with Recursive Feature Elimination-Isolation Forest for classification of diabetes mellitus", PLoS ONE, Volume 19, Issue 5, 2024, Pages 1-18
2. Saumyadeep Chakrabarty, Susovan Jana, Pulak Baral, "A Data-Driven Approach for the Prediction of Medical Data", IEEE Access, Volume 04, October 2024.
3. V. K. Daliya and T. K. Ramesh, "A Cloud-Based Optimized Ensemble Model for Risk Prediction of Diabetic Progression— An Azure Machine Learning Perspective", IEEE Access, Volume 13, January 2025, Pages 11560 – 11575.
4. Alfredo Daza, Carlos Fidel Ponce Sanchez, Gonzalo Apaza-Perez, Juan Pinto, Karoline Zavaleta Ramos, "Stacking ensemble approach to diagnosing the disease of diabetes", Informatics in Medicine Unlocked, Elsevier, Volume 44, 2024, Pages 1-22.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

5. Xin Feng, Yihuai Cai Ruihao Xin, "Optimizing diabetes classification with a machine learning-based" (2023), BMC Bioinformatics, 2023 Nov 13;24:428. doi: 10.1186/s12859-023-05467-x.
6. Mohamed Elkawkagy, E. Elwan, Albandari Alsumayt, Heba Elbeh, Sumayh S. Aljameel, "Elevating Big Data Privacy: Innovative Strategies and Challenges in Data Abundance", IEEE Access, Volume 14, January 2024, Pages: 20931 – 20941.