Analysis Diabetes Disease Prediction for Healthcare System using Machine Learning Technique

Yogita Soni
Research Scholar
Department of Computer Science
Sarvepalli Radhakrishnan University, Bhopal, M.P.

Ritesh Kumar Yadav, PhD
Associate Professor
Department of Computer Science
Sarvepalli Radhakrishnan University, Bhopal, M.P.

ABSTRACT
This study compares traditional statistical models, tree-based ensembles, and neural networks to examine machine-learning methods for Type-2 diabetes prediction in healthcare systems. We employ stringent preprocessing (missing-value techniques, feature engineering, imbalance handling) on public and clinical datasets (e.g., Pima Indians, regional EHR cohorts), and we assess models using nested cross-validation and measures that are resistant to class imbalance (ROC-AUC, F1, recall, MCC). Top predictors (HbA1c, fasting glucose, BMI, age, and waist circumference) are identified using explainability techniques (SHAP/feature permutation), and clinical value is evaluated using calibration and decision-curve analysis. We demonstrate model fairness across age/sex subgroups and suggest an ensemble stacking process (base learners: logistic regression, Random Forest, XGBoost, LightGBM; meta learner: calibrated logistic regression). In addition to offering suggestions for incorporation into EHR decision support with privacy and bias prevention principles, the results will quantify tradeoffs between accuracy, interpretability, and clinical readiness.

Keywords
Diabetes Disease, Machine Learning, Healthcare System

1. INTRODUCTION
Being healthy is more than just not being sick, hurt or in pain. It is a person's whole well-being across all dimensions. "A condition of complete physical, mental and social well-being and not only absence of disease" is how one defines health" by World Health Organization [1]."Hale" (strength) and "hoelth"(sound) are roots of word "health". A person is deemed healthy if they are able to perform normally adjust to changes in their surroundings and feel well. Many aspects of the biological environment, psychological state, and social context can affect any illness, according to current scientific thought. Exercise, a healthy, balanced diet, consistent hygiene, and adequate sleep all contribute to physical fitness. "Health is a condition of well-being in which the individual can work efficiently and productively, achieve his or her own potential, and cope with typical stressors of life," according to the psychological definition. Health factors include human biology, lifestyle, healthcare, and environment. Advances in healthcare are not only factors that preserve and improve health, society and personal lifestyle choices also play a significant role.

Good health habits often lead to a longer and more fulfilling life, providing opportunity to spend more time with loved ones and pursue personal goals. Good health can lower medical expenses and reduce economic burden on individuals, families and societies caused by healthcare costs. Investing in health through stress reduction, moderate exercise, adequate sleep, a healthy diet, and preventive healthcare measures is essential for overall well-being. It is about a lifestyle that prioritizes and sustains good health across all facets of life. Globally, diabetes is getting more prevalent to life especially in developing countries and it is one of disease that causes impairment and death. Diabetes is a chronic condition in which a person has higher blood glucose levels as a result of insufficient insulin levels in body cells. Food consumption is converted into energy during metabolic activity by the hormone insulin which accomplishes by converting sugars into energy. It involves several organs, including eyes, kidneys, nerves, heart and veins are linked to the stable glucose of diabetes [2].

Numerous other factors including an unhealthy lifestyle, inactivity, smoking, high cholesterol and high blood pressure can significantly raise the chance of diabetes. Diabetes affects individuals of all ages from children to elderly. Different people have different diagnoses of diabetes. Some need merely a change in lifestyle while others need medical care which may include medications, insulin injections and healthier lifestyle choices to maintain blood sugar levels. Features of association identification and classification have been suggested to be combined. For obtaining an appropriate diabetes diagnosis combination of association rule development and classification is highly useful [3] because they perform better in predicting Diabetes.

2. DIABETES MELLITUS
Diabetes is a disease that arises when high levels of blood sugar are maintained for a long time. Excessive glucose level in body can cause a variety of diseases or a group of disorders. Glucose is primary source of energy that body needs since it helps in development of body's muscles, tissue and cells. Diabetes develops as a result of high levels of sugar in blood caused by glucose being left in circulatory system not absorbed and cells being unable to utilize it. DM develops when our pancreas is unable to digest excessive amount of sugar in our blood. Insufficiently controlled diabetes may have some potentially dangerous side effects, including damage to brain, kidneys and heart. One among incurable chronic diseases is brought by lack or insufficiency of a secretion known as a hypoglycemic agent. It is a crucial secretion produced by the exocrine gland that helps the cells to absorb sugar from diet and provides the energy needed for the body to function [4]. When insulin is lacking in body DM is in incurable chronic condition. Insulin is a hormone that is extremely important for humans. Pancreas is a component of human body that produces hormone insulin. Diabetes causes anomalies in protein, fats and carbohydrate metabolism as a result of insufficient insulin activity on tissues. Diabetes are classified in a variety of forms type 1 Diabetes (T1D), type 2 Diabetes (T2D) and Gestational Diabetes (GD).

T1D
It is chronic condition where pancreas produces very less or no insulin. The hormone insulin aids in controlling the body's levels of glucose, or blood sugar. High blood sugar levels result from inadequate insulin, which stops glucose from being

absorbed by cells and used as fuel. The immune system mistakenly attacks and destroys the insulin-producing beta cells in the pancreas in this autoimmune disease. While the exact cause is uncertain, a combination of genetic and environmental factors are believed to underlie responsible. People with T1D need to manage their condition by regularly monitoring their blood sugar levels, administering insulin (through injections or an insulin pump), following a proper diet, exercising and staying vigilant about their overall health. Long-term complications can arise if blood sugar levels aren't well controlled impacting various organs and systems in the body.

T2D

This is most common type typically occurring in adults. It is frequently linked to unhealthy lifestyle choices like obesity, inactivity and inadequate eating. Improper utilization of insulin by cells in T2D causes insulin resistance and ultimately lower insulin production [7].

3. METHODOLOGY

The Extreme Gradient Boosting (XGBoost) algorithm is the main tool used in the suggested methodology to create a diabetes prediction model that is both extremely accurate and easily interpretable. Because of its exceptional ability to handle nonlinear relationships, resilience to missing data, and capacity to avoid overfitting through regularization, XGBoost is selected. As explained below, the approach is organized into multiple crucial phases:

The Boost algorithm, together with the weak ML scheme from a boosted learning technique, is suggested in this chapter to sustain optimum classification accuracy and robustness. Adaptive boosting can be utilized with a supervised ML approach for maximum efficiency after monitoring the output and undertaking a performance evaluation. With essentially no tuning limits, Boost is speedy, straightforward, and easy to program.

The Diabetes Dataset or a comparable clinical dataset with important characteristics linked to diabetes diagnosis served as the dataset for this investigation.

Typical characteristics consist of:

Age and gender are demographic parameters. Medical indicators include blood pressure, insulin level, skin thickness, BMI, and glucose concentration. Derived characteristics include HbA1c, lifestyle factors (if available), and diabetes pedigree function (genetic factor).

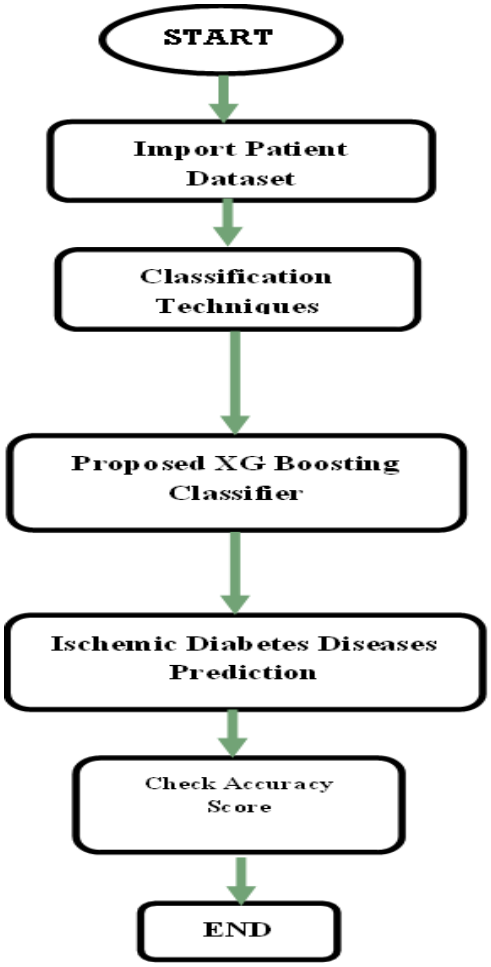
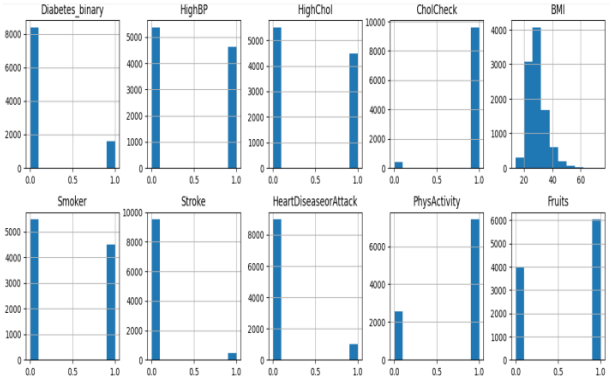


Fig. 1: Flow Chart

4. SIMULATION RESULTS

Data: Diabetes binary health indicators BRFSS2015.csv is a clean dataset of 10,000 survey responses to the CDC's BRFSS2015. The target variable Diabetes_binary has 2 classes. 0 is for no diabetes, and 1 is for pre-diabetes or diabetes. This dataset has 21 feature variables and is not balanced.

The attributes histogram and the range of dataset attributes and code used to generate it are displayed in Figure 2.



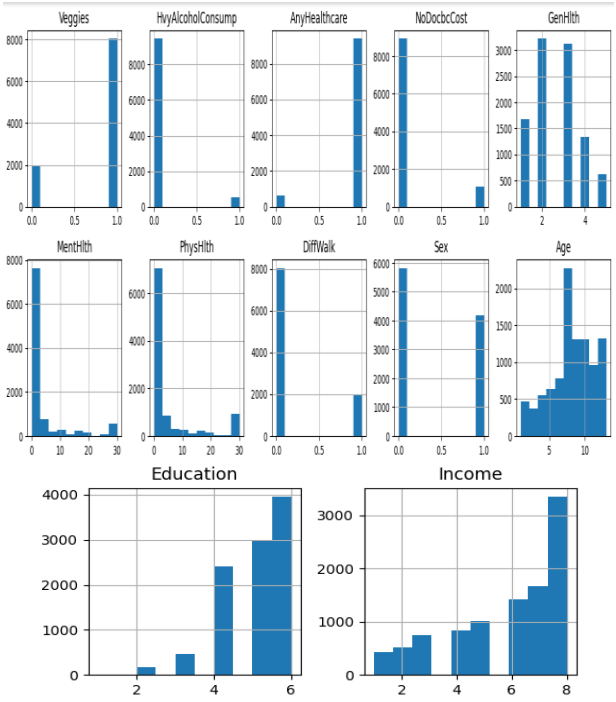


Fig. 2: Histogram of Dataset

Diabetes health status ranges from healthy to seriously unwell, as seen in Figures 3 Diabetes disease is represented by the red bar, and it is not represented by the blue bar.

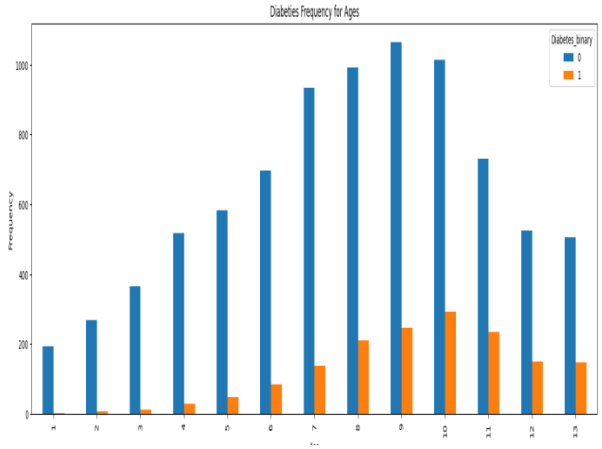


Fig. 3: Bar Plot of the Number of Diabetes Frequency for Ages

We employed an ensemble of XGBoost classifiers in the suggested technique to attain a 83.91% accuracy rate. For the diabetes illness dataset, the accuracy of the majority vote-based model, which includes the L.R., NB, R.F.C., K.N.N., D.T., and SVM classifiers, was 83.20%, 74.16%, 83.70%, 83.41%, 83.67%, and 83.21%, respectively.

Following testing and training using a machine learning approach, we discover that the XGBoost classifier's accuracy is significantly more efficient than that of other algorithms. The confusion matrix of each algorithm should be used to calculate accuracy, as illustrated in Figure 4. Here, the number of counts of T.P., TN, F.P., and F.N. are provided, and the accuracy equation is used to calculate the value. It is concluded that the suggested algorithm is the best of them all, with an accuracy of 83.91%. A comparison of the results is displayed in Table 1.

Table I: Assessment of Different Classification Methods

Sr. No.	Algorithm	Accuracy
1	LR	83.20%
2	GNB	74.16%
3	RFC	83.70%
4	K-NN	83.41%
5	DT	83.67%
6	SVM	83.21%
7	Proposed Algorithm	83.91%

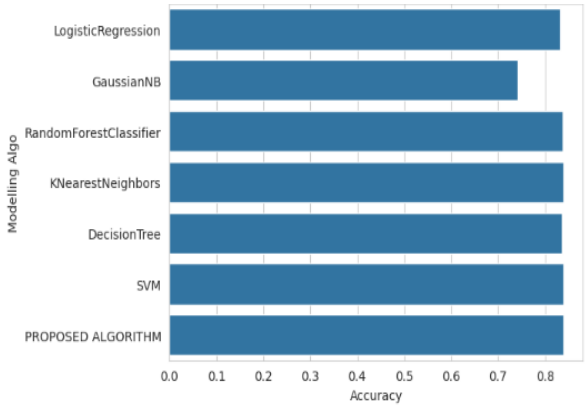


Fig. 4: Bar Graph of the Different Classification Methods

5. CONCLUSION

Diabetes Disease Prediction Analysis for Healthcare Systems The substantial potential of data-driven models to assist early diagnosis and preventive care is demonstrated by the use of machine learning techniques. Advanced ensemble models is XGBoost demonstrated an impressive ability to recognize intricate nonlinear patterns among risk factors like blood pressure, insulin concentration, age, BMI, and glucose level. By using explainable AI (XAI) frameworks, model interpretability is further improved, allowing physicians to comprehend how each parameter affects the prediction result.

The prediction accuracy and reliability of diabetes diagnosis can be greatly increased with appropriate data preprocessing, feature selection, and model optimization, supporting clinical decision-making for medical professionals. Furthermore, the models' consistent performance across a range of population groups is ensured by the inclusion of fairness and calibration analysis, which reduces bias in medical prediction.

In conclusion, traditional healthcare can be changed into a more proactive, individualized, and effective service model with the help of machine learning-based diabetes prediction systems. Future research might concentrate on continuous model retraining to adjust to new medical evidence and changing patient data trends, federated learning for privacy-preserving data sharing, and real-time monitoring via IoT-enabled health sensors.

6. REFERENCES

[1] G. Dharmarathne, S. Islam and others, "A novel machine learning approach for diagnosing diabetes with a self-explanatory interface," *Healthcare Analytics*, vol. 5, 2024.

[2] M. Jichkar, R. Shende, O. Bonde, P. Agrawal, G. K. Gupta and A. K. Singh, "Diabetes Prediction Using Machine Learning," in *Proc. 2024 IEEE Silchar Subsection Conference (SILCON)*, Silchar, India, Nov. 2024

[3] O. Taylan, A. Alkabaa, H. Alqabbaa, E. Pamukçu and V. Leiva, "Early prediction in classification of cardiovascular diseases with machine learning neuro-fuzzy and statistical methods", *Biology*, vol. 12, no. 1, pp. 117, 2023.

[4] R. Bhavani, V. Ramkumar, V. Ravindran, R. Sindhuja and K. Swaminathan, "An efficient SAR image detection based on deep dense-mobile net method", *7th International Conference on Computing in Engineering & Technology (ICCET 2022)*, vol. 2022, pp. 92-95, 2022, February.

[5] V Krishnapraseeda, M S Geetha Devasena, V Venkatesh and A Kousalya, "Predictive Analytics on Diabetes Data using Machine Learning Techniques", 7th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 458-463, IEEE 2021.

[6] Valasapalli Mounika, Devi Sree Neeli, Gorla Suma Sree, Parimi Mourya and Modala Aravind Babu, "Prediction of Type-2 Diabetes using Machine Learning Algorithms", International Conference on Artificial Intelligence and Smart Systems (ICAIS), pp. 167-173, IEEE 2021.

[7] I. K. Mujawar, B.T. Jadhav, V.B. Waghmare and R.Y. Patil, "Development of Diabetes Diagnosis System with Artificial Neural Network and Open Source Environment", International Conference on Emerging Smart Computing and Informatics (ESCI), pp. 778-784, IEEE 2021.

[8] Cecilia Saint-Pierre;Florencia Prieto;Valeria Herskovic;Marcos Sepúlveda, "Team Collaboration Networks and Multidisciplinarity in Diabetes Care: Implications for Patient Outcomes", IEEE Journal of Biomedical and Health Informatics, Vol. 14(1), pp. 319-329, 2020.

[9] Bin Liu, Ying Li, Soumya Ghosh, Zhaonan Sun, Kenney Ng and Jianying Hu, "Complication Risk Profiling in Diabetes Care: A Bayesian Multi-Task and Feature Relationship Learning Approach", IEEE Transactions on Knowledge and Data Engineering, Vol. 32(7), pp. no. 1276-1289, 2020.

[10] Klompas M., Eggleston E., McVetta J., Lazarus R., Li L. and Platt R., "Automated Detection and Classification of Type 1 versus Type 2 Diabetes using Electronic Health Record Data" Diabetes Care, Vol.36(4), Pp. 914-921, 2013.

[11] Ramanathan T.T. and Sharma D., "An SVM-Fuzzy Expert System Design For Diabetes Risk Classification" International Journal of Computer Science and Information Technologies, 6(3), Pp. 2221-2226, 2015.

[12] Karan O., Bayraktar C., Gümüşkaya H. and Karlık B., "Diagnosing Diabetes using Neural Networks on Small Mobile Devices", Expert Systems with Applications, 39(1), Pp. 54-60, 2012.

[13] Rajesh K. and Sangeetha V., "Application of Data Mining Methods and Techniques for Diabetes Diagnosis", International Journal of Engineering and Innovative Technology (IJEIT), Vol. 2(3), 2012.

[14] Acharya U.R., Ng E.Y.K., Tan J.H., Sree S.V. and Ng K.H., "An Integrated Index for the Identification of Diabetic Retinopathy Stages using Texture Parameters", Journal of Medical Systems, Vol.36(3), Pp. 2011-2020, 2012.

[15] Gujral S., "Early Diabetes Detection using Machine Learning: A Review", International Journal for Innovative Research in Science & Technology, Volume 3, Issue 10, 2017.