A Survey on Diabetes Prediction Using Machine Learning

[1]Rintu Raju, [2]Dr.Rahul Shajan

[1]PG Scholar, [2]Assistant Professor
[1&2]Department of MCA, St.Joseph's College of Engineering and Technology,palai,Kottayam,India

Abstract: Machine learning techniques, which teach computers to learn via experience, are used to analyze large databases. In the field of medical forecasting, ML approaches are currently used to estimate the probability that a patient will contract a disease in the future. One of the medical predictions that makes use of machine learning is in the situation of diabetes. This study's objective is to provide a comparative analysis of 10 papers through an examination of algorithm performance using various metrics.

Index Terms: Diabetes Prediction, Machine Learning (ML), Comparative Analysis

I. INTRODUCTION

Millions of individuals throughout the world struggle with diabetes mellitus, a chronic illness. It is brought on by the body's inability to create enough insulin or utilize it effectively, which leads to elevated blood glucose levels. The practise of training computers to learn from information and predict the future is described as machine learning, a branch of artificial intelligence. Machine learning techniques were utilized to forecast diabetes more and more lately. Type 1, type 2, and gestational diabetes are only a few of the several varieties of the disease. When the body's immune system assaults and kills the cells in the pancreas that make insulin, type 1 diabetes develops. When the body develops an insulin resistance or cannot create enough insulin to fulfil the body's needs, type 2 diabetes results.

II. LITERATURE REVIEW

Detecting diabetes using machine learning techniques and python GUI [1]: This model has been built using the SVM algorithm to forecast a patient's early-stage risk of developing diabetes.The UESD data set is A well-known dataset that is frequently utilised to generate models to forecast the onset of diabetes is the Pima Indians Diabetes dataset. It provides details on several facets of Pima Indian women's health, including their age, BMI, and blood sugar levels, among other aspects..Support Vector Machines are a powerful machine learning algorithm that is commonly utilised for classification tasks.SVMs segregate data into several groupings by locating the optimal hyperplane.It's also impressive that the model has achieved an accuracy of 80.5 . Diabetes prediction using supervised machine learning [2]: This study compares the performance of two popular machine learning algorithms, K-Nearest Neighbor and Naive Bayes, on the task of predicting diabetes utilizing the Pima Indians Diabetes Dataset.It suggests that Naive Bayes was better at correctly predicting cases of diabetes and non-diabetes in the test data. It's is a positive result for Naive Bayes and indicates that it may be a better choice for this particular task. A comparison of machine learning algorithms for diabetes prediction [3]: The effectiveness of several machine learning algorithms on a challenge to predict diabetes is examined in this article.. The authors found that combining logistic regression and support vector machine resulted in good prediction performance, and that a neural network model with two hidden layers achieved an accuracy of 88.6 percent.. A comparative analysis of early stage diabetes prediction using machine learning and deep learning approach [4]: In order to predict early diabetes disease, this paper compared and analyzed several ML and DL methods. Additionally, this model makes use of a diabetes data set with 17 attributes from the UCI repository, and it assesses the model's performance using a range of performance metrics.It is impressive that the XGBoost classifier algorithm achieved an accuracy of approximately 100% in predicting early diabetes disease, while the other algorithms had an accuracy of 90 percent .Classification and prediction of diabetes disease using machine learning paradigm [5]: This study was able to achieve a high accuracy rate in predicting diabetes disease. The use of multiple classifiers like decision tree, NB, adaboost, and random forest can help improve the overall performance of the machine learning system. Additionally, using logistic regression to pinpoint the peril factors for diabetes can provide valuable insights into the disease's underlying causes.The amalgamation of RF-based classifier and LR-based feature selection resulted in an even higher accuracy rate of 94.25 percent. This suggests that the two techniques complement each other well and may be effective in identifying important risk factors for diabetes.Prediction of onset diabetes using machine learning techniques [6]: This study compares several different types of classification algorithms, including SVM, Naive Bayes, and Logistic Regression. Among those Logistic Regression gives an accuracy of 78.01 percent. Analysis and prediction of diabetes mellitus using machine learning algorithm [7]: Proposed system includes four divergent machine learning algorithms: Naive net, Support Vector Machine, Decision Stump, and a proposed Ensemble method (PEM). The system has been tested on the data set, and the outcomes indicate that SVM has the highest accuracy at 88.8%, followed by Bayes Net at 88.54%, Adaboost M1 at 85.68%, and Decision Stump at 83.72%.However, the collaborative model (Ensemble) appears to offer the highest accuracy of 90.36%. This suggests that combining the outputs of multiple models can result in higher overall accuracy than any individual model alone. . Machine learning based diabetes prediction and development of smart web applications [8]: In this framework, numerous data sets are used to train machine learning algorithms such as k-nearest neighbour, decision tree, Naive Bayes, Random Forest, Logistic Regression, Support Vector Machine, and Gradient Boosting. Based on the given accuracy scores, SVM outperforms the other methods as it has the highest accuracy score of 80.26 percent. Random Forest also has a relatively high accuracy score of 80.25 percent, but SVM has a slightly better accuracy score, making it the best-performing model among the given methods.. A novel diabetes health care disease prediction using machine learning techniques [9]: Predictive analysis in the health care industry is examined in this study. This article analyzed the Pima Indian Diabetes Database and employed a variety of machine learning methodologies to anticipate diabetes. The study found that K Nearest

Neighbor, Support Vector Machine ,random forest and logistic regression techniques all provided reasonably accurate results, with an overall accuracy of 83 percent and low error rates. Prediction of diabetes empowered with fused machine learning [10]: Artificial Neural Network and Support Vector Machine models are the two algorithms that this model employs. This system decides if a diabetes diagnosis is favorable or unfavorable. This proposed machine learning model outperforms the previously published techniques with a prediction accuracy of 94.87 percent.

III. METHODOLOGY

A) *Support Vector Machine:* An algorithm for both classification and regression using supervised machine learning. This algorithm's main goal is to find the finest Hyper plane, or line, in an n-dimensional space for categorizing data points.[1]

B) *K Nearest Neighbor:* KNN is a type of the simplest learning approaches, primarily used for classification but also for regression as well. During the training phase, the KNN algorithm merely stores the data set, and when it receives new data, it classifies the old data into a category that is highly close to the new data..*[2]*

C) *Naive Bayes:* In Naive Bayes, the features are presumed to be conditionally independent, which indicates that the presence or absence of one trait has no bearing on the likelihood that another feature would be present or absent. This assumption simplifies the calculation of probabilities and makes the algorithm computationally efficient.. [5]

D) *Logistic Regression:* A method of supervised learning that addresses classification problems and predicts the likelihood of a binary (yes/no) event. The algorithm here determines whether or not a person is likely to have diabetes. [3]

E) *XG Boost Classifier:* A supervised learning technique, which is an is an extreme gradient boost algorithm for implementing gradient boosted decision trees. [8]

F) *Decision Tree:* It's indeed a popular supervised learning algorithm utilized for both regression and classification tasks. The tree-like structure of the algorithm allows it to recursively partition the feature space into smaller and smaller subsets based on the values of the input features, ultimately leading to a prediction or decision at the leaf nodes of the tree.. *[8]*

G) *AdaBoost:* The idea behind AdaBoost is to iteratively train a series of weak classifiers on the same data, each time assigning higher weights to the misclassified samples from the previous iteration. By doing this, the algorithm focuses on the difficult samples and allows the weak classifiers to learn from them more effectively. [5]

H) *Random Forest:*It is built on the concept of ensemble learning, where various models are blended to optimize the algorithm's overall performance.[9]

I) *Naive net:* A naive net is a Bayesian network with a single root, all other nodes are children of the root. [7]

J) *Decision Stump*:A decision stump is a type of decision tree model that comprise of single decision node and two or more leaf nodes. [7]

IV. ANALYSIS

Paper Title	Algorithm Used	Year	Accuracy
Detecting diabetes using machine learning techniques and python GUI [1]	Support Vector Machine [1]	2023	80.5
Diabetes prediction using supervised machine learning [2]	K Nearest Neighbor Algorithms and Naive Bayes Algorithm [2]	2019	76.07
Comparison of machine learning algorithms for diabetes prediction [3]	Support Vector Machine (SVM) and Logistics Regression [3]	2021	88.6
A Comparative Analysis of Early-Stage Diabetes Prediction using Machine Learning and Deep Learning Approach [4]	ML and DL Classification Algorithms [4]	2021	According to this experiment, the XG Boost classifier algorithms provides roughly 100.0%, while the rest of the algorithms provide90.0% accuracy.

Classification and prediction of diabetes disease using machine learning paradigm [5]	Decision Tree (DT),Random Forest(RF),Adaboost(AB), and ,Naïve Bayes (NB),[5]	2020	94.25
Prediction of Onset Diabetes using Machine Learning Techniques [6]	Logistic Regression,Support Vector Machine(SVM) and ,Naive Bayes[6]	2017	78.01
Analysis and Prediction of Diabetes Mellitus using Machine Learning Algorithm [7]	Naïve Net,SVM(Support Vector Machine)Ensemble method (PEM) and Decision Stump[7]	2022	90.36
Machine learning based diabetes prediction and development of smart web application [8]	Naive Bayes (NB), knearest neighbor (KNN), Logistic Regression (LR), Decision tree (DT) Support Vector Machine(SVM),Gradient Boosting (GB), and Random Forest (RF) [8]	2021	78.95,76.32,80.25,80.26,77.63,78.95 and 75 respectively
A Novel Diabetes Health care Disease Prediction Framework Using Machine Learning Techniques [9]	Random Forest (RF),support vector machine (SVM) and decision tree (DT)-based [9]	2022	83
Prediction of Diabetes Empowered With Fused Machine Learning [10]	Artificial Neural Network (ANN) and Support Vector Machine (SVM)[10]	2022	94.87

V. CONCLUSION

Through this paper a systematic study of 10 papers have been done.It is vital to notice that the accuracy of a model should not be the only factor considered when evaluating its performance. It's also worth taking into account other metrics notably sensitivity, specificity, accuracy, and recall. Additionally, the data utilized to develop and validate the models,as well as the features picked-out, can greatly impact their performance.It is also important to consider the context in which these models will be used. In a clinical setting, false negatives (when a patient with diabetes is incorrectly classified as not having diabetes) can have serious consequences. Therefore, a model with high sensitivity would be preferred, even if it may have a slightly lower overall accuracy.Overall, while the XGBoost classifier algorithm may have shown higher accuracy in these studies, it is important to thoroughly evaluate and validate any model before implementing it in a real-world setting.

VI. ACKNOWLEDGMENT

First and foremost, I give all glory, honor and praise to God Almighty who gave me wisdom and enabled me to finalize this work successfully.

I also desire to express my sincere gratitude to my parents for supporting me in this endeavor and in all of my other attempts.

I am incredibly grateful to Dr. V. P. Devasia, Principal of SJCET in Palai, for letting me use all of the facilities there as well as for his support. Words cannot adequately express how grateful I am.

I want to express my sincere appreciation. to Mr. Anish Augustine, HOD Incharge, Department of MCA, SJCET, Palai, who has served as a constant inspiration and without whose invaluable assistance and support this work would not have been possible.

I have a special debt of gratitude to Dr.Rahul Shajan, Asst. Professor, Department of Computer Science and Applications, SJCET, Palai, for all the necessary help and support that he has extended to me. His valuable suggestions, corrections, and sincere efforts to accomplish this work even under a tight time schedule were crucial to the successful completion of this work.

I want to sincerely thank all of our teachers and non-teaching staff at SJCET, Palai, for the knowledge they have imparted to me over the last three years.

Additionally,I want to thank everyone for their encouragement, advice, and support.

REFERENCES

1. Manivannan D and Manikandan N.K, "Detecting diabetes using machine learning techniques and python GUI" , https://aip.scitation.org/doi/abs/10.1063/5.0111455,30 January 2023
2. Muhammad Exell Febrian , Fransiskus Xaverius Ferdinan,"Diabetes prediction using supervised machine learning"Gustian Paul Sendani , Kristien Margi Suryanigrum and Rezki Yunanda, https://www.sciencedirect.com/science/article/pii/S1877050922021858 ,2023
3. Jobedha Jamal Khanam and Simon Y Foo ,"A comparison of machine learning algorithms for diabetes prediction", https://www.sciencedirect.com/science/article/pii/S2405959521000205,December ,2021
4. Md Abu Rumman Refat , Md. Al Amin , Chetna Kaushal , Mst. Nilufa Yeasmin and Md Khairul Islam, "A Comparative Analysis of Early Stage Diabetes Prediction using Machine Learning and Deep Learning Approach",https://ieeexplore.ieee.org/abstract/document/9609364
5. Md. Maniruzzaman1,2*, Md. Jahanur Rahman2, Benojir Ahammed1 and Md. Menhazul Abedin1,"Classifcation and prediction of diabetes disease using machine learning paradigm",https://link.springer.com/article/10.1007/s13755-0190095-z,3 January 2020
6. Md. Aminul Islam and Nusrat Jahan,"Prediction of Onset Diabetes using Machine Learning",https://www.researchgate.net/profile/Nusrat49/publication/321847368_Prediction_of_Onset_Diabetes_using_Machine_Learning_Techniques/links/5d2460b6458515c11c1f5911/Prediction-of-Onset-Diabetes-using-MachineLearning-Techniques.pdf Techniques,December 2017
7. Minyechil Alehegn , Rahul Joshi& Dr. Preeti Mulay,"Analysis and Prediction of Diabetes Mellitus using Machine Learning Algorithm",https://d1wqtxts1xzle7.cloudfront.net/77988080/87-libre.pdf?1641277676=&response-contentdisposition=inline%3B+filename%3DAnalysis_and_Prediction_of_Diabetes_Mell.pdf&Expires=1676738397&Signature =PyxIHkoEs6ubxpec9aJ~lkTKGGp5ZfpU4t8kkOYVhl6q4~tAsvqFLRO8sQNZNLADLP8n~geImEu28XdzgtmAHZOiIA4TlHmQPLfhsQvh5cM0vgKQMm1bJZ7kaQQCicarvM3E7WkXnoN5CLjasOmLbHOoy4ivHWyS1wLzEuxaTnWl6B7WfXzNcwACYo6rWm7ahBYDgelcxPy2oBoh9xa92dh9fKNxcZ19Umj12xrlf9xf0DQiOX2ww~9lRpwpz7jZqidQnFuvhYtxhDf8of0nLmNbvFVHowOnGZZ3wXuEC~1Q9zgeJLHcxie4bUNKB8SrRYKc--Ox5j5G8KWEkWVw__&KeyPair-Id=APKAJLOHF5GGSLRBV4ZA,2018
8. NazinAhmed,RayhanAhammed,Md.ManowarulIslam,Md.AshrafUddin,ArnishaAkhter,Md.AlaminTalukder,BikashKumarPaul,"Machine learning based diabetes prediction and development of smart web application",https://www.sciencedirect.com/science/article/pii/S2666307421000279,June 2021
9. Raja Krishnamoorthi,Shubham Joshi , Hatim Z. Almarzouki ,Piyush Kumar Shukla , Ali Rizwan , C. Kalpana,6 and Basant Tiwari,"https://www.hindawi.com/journals/jhe/2022/1684017/,2022
10. Usama ahmed, Ghassan f. issa, Muhammad adnan khan,Shabib aftab, Muhammad farhan khan, raed a. t. said,Taher m. ghazal , and Munir ahmad,"Prediction of Diabetes Empowered With Fused Machine Learning,https://ieeexplore.ieee.org/abstract/document/9676634,2022