Grenze International Journal of Engineering and Technology, January Issue

An Insight of Diabetes Detection Performance using Machine Learning Techniques

Amita Gautam[1] and Shikha Verma[2]
[1-2]Sharda School of Engineering & Technology, Sharda University, Greater Noida, India
Email: 2024335393.amita@dr.sharda.ac.in, shikha.verma@sharda.ac.in

***Abstract*— Diabetes is a metabolic disorder that affects many people around the world. Every year, its incidence rates are rising rapidly. Diabetes-related problems in many cases have the potential to affect the body's vital organs & become lethal if left untreated. As this disease remains in asymptotic condition for long duration so its early detection can help. So, it must be detected early in order to be treated in a timely manner and prevent further complications.**
This paper presents a comprehensive review of machine learning approaches for early detection of Type 2 Diabetes Mellitus (T2DM). The study addresses the critical need for improved diabetes prediction methods and proposes automated diagnosis systems using various machine learning techniques. Multiple algorithms like Logistic regression, support vector machine, K-nearest neighbour & Random Forest were evaluated across different datasets, with Random Forest emerging as the most effective classifier, achieving up to 99% accuracy in some studies. The research examined diverse data sources including electronic health records and specialized datasets, focusing on key predictors such as fasting plasma glucose, HbA1c, and BMI. While the results demonstrate the transformative potential of AI-driven technologies in diabetes care, challenges remain regarding limited sample sizes and clinical validation. Future work aims to expand datasets, improve preprocessing methods, and explore applications in other diseases.

***Index Terms*— Machine learning, Ensemble machine learning models, Logistic Regression (LR), Support vector machine (SVM), K-nearest neighbour (KNN) and Random Forest (RF).**

I. Introduction

Diabetes is a chronic disease that harms the pancreas and compromises the body's ability to make insulin. Insulin is the main element in our body that keeps blood glucose levels stable. The three types of diabetes that must be detected are type 1, type 2, and gestational diabetes. People with type 1 diabetes have little or no insulin induced by the pancreas. Insulin therapy is one of the most common therapies for type 1 diabetes. Children and young adults (those under 30) are usually strained. While, type 2 diabetes is more usual among obese people and those over 65 and is normally caused by insulin resistance. Gestational diabetes, commonly known as hyperglycemia, is another problem that surfaces during pregnancy [15]. The hormone insulin regulates blood glucose levels and allows glucose to penetrate into cells for consumption as an energy source. Lowered insulin action leads to a accumulation of glucose in the blood, which shoots blood sugar levels. A rise in blood sugar levels is an indication of diabetes, which is scientifically known as diabetes mellitus (DM). It effects the kidneys, skin, nerves, and eyes with diabetics retinopathy (DR).

Grenze ID: 01.GIJET.12.1.213
© Grenze Scientific Society, 2026

As per World Health Organization (WHO), by 2040, about 600 million people across the globe are expected to have diabetes. This alarming forecast highlights how important it is to address the condition's rising chronicity. The number of newly diagnosed cases is continuously increasing [2], highlighting the dire need for strategies to address the disease and reduce its effect on the world's health. The treatment has higher success rate & less consequences if disease is diagnosed early.

II. Techniques used for Diabetes Detection

A. Supervised Learning

Supervised learning is one of the three type of machine learning. The task of this learning is mapping the function input to output data. We train the machine with labeled data i.e input-output pair data so that machine can learn and act further. A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples [4].
Supervised Learning Types:
Classification: It is supervised learning where output is a discrete value. It classifies the output value in to homogeneous group. It aims to predict discrete value belonging to specific class or group.
Regression: Regression model is used to predict the continuous value like predicting the salary of employee based on the age.

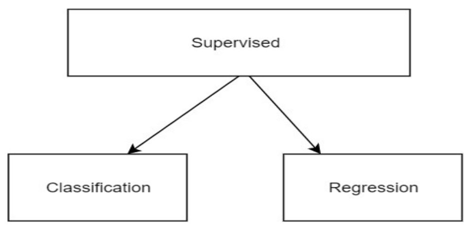

Fig. 1. Type of supervised learning

Brief description of Machine Learning Classification Techniques:

Logistic Regression Method (LR)
Logistic regression is a sort of supervised learning which estimates the connection between a binary dependent variable and at least one independent variable by evaluating probabilities with the help of sigmoid function. In contrary to its name, logistic regression is not used for regression problems rather is a type of machine learning classification problem where the dependent variable is dichotomous (0/1, -1/1, true/false) and independent variable can binomial, ordinal, interval or ratio level. The sigmoid/logistic function is given as [25]:

$$y = 1/ (1 + e\text{-}x) \qquad (1)$$

Where, y is the output which is the result of weighted sum of input variables x. If the output is more than 0.5, the output is 1 else the output is 0.

K- Nearest Neighbor Classifier
K-Nearest Neighbor (KNN) method can be used to solve problems pertaining to both regression as well as classification, though it is generally being used to solve classification problems in business. Its major advantage is simplicity of translation and low computation time. In figure 2, the points (2.5, 7) and (5.5, 4.5) will be allocated to any one of the clusters. The KNN uses Euclidean distance function to find distances between existing data points and any new data point. Thus, (2.5, 7) will belong to the green cluster, whereas, (5.5, 4.5) will belong to the red cluster [26].

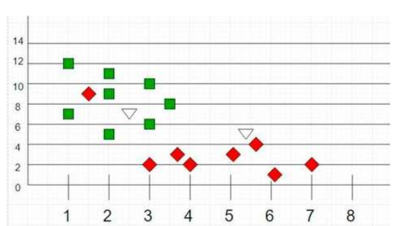

Fig. 2. KNN example [26]

Support Vector Machine (SVM)

SVM is a supervised classifier in machine learning algorithms that can be used both for regression and classification. It is majorly applied in solving classification problems. The goal of SVM is to classify data points by an appropriate hyperplane in a multidimensional space. A hyperplane is decision boundary to classify data points. The hyperplane classifies the data points with maximum margin between the classes and the hyperplane. Figure 3 shows support vector machine classification [27].

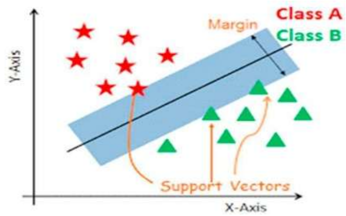

Fig. 3. Support Vector Machine [27]

Naïve Bayes(NB)Classification Method

Naïve bayes classification method is a probabilistic machine learning algorithm based on Bayes theorem described in probability. Even with its simplicity it outperforms other classifiers; hence, it is one of the best classifiers. The Bayes theorem for calculating posterior probability is given below [28]:

$$p(c \mid x) = \frac{p(x \mid c)\, p(c)}{p(x)} \tag{2}$$

Where,
P(c|x) = Posterior Probability
P(x|c) = Likelihood
P(c) = Class Prior Probability
P(x) = Predictor Prior Probability

Decision Tree (DT) Classification Method

A Decision Tree is works on the principle of decision making. It can be described in form of tree and provides high accuracy and stability. Figure 4 illustrates a decision tree [29].

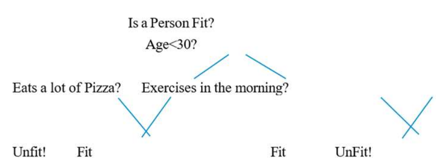

Fig. 4. Decision Tree Example [27]

Random Forest Classification

The Random Forest classifier creates multiple decision tree from randomly selected subset of training dataset shown in figure 5. Then it aggregates the votes from different decision trees to decide the final class of test objects [30].

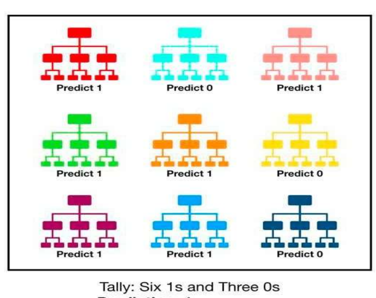

Fig. 5. Random Forest Classification

Confusion Matrix [31]

A confusion matrix is a matrix that summarizes the performance of a machine learning model on a set of test data. It is a means of displaying the number of accurate and inaccurate instances based on the model's predictions. It is often used to measure the performance of classification models, which aim to predict a categorical label for each input instance.

The matrix displays the number of instances produced by the model on the test data.

True Positive (TP): The model correctly predicted a positive outcome (the actual outcome was positive).

True Negative (TN): The model correctly predicted a negative outcome (the actual outcome was negative).

False Positive (FP): The model incorrectly predicted a positive outcome (the actual outcome was negative). Also known as a Type I error.

False Negative (FN): The model incorrectly predicted a negative outcome (the actual outcome was positive). Also known as a Type II error.

In order to find the exact accuracy of every method the following measures has been calculated by given formulas [32-34]:

$$\text{AccuracyRate} = \frac{TP + TN}{TP + TN + FN + FP} \tag{3}$$

$$\text{ErrorRate} = \frac{FN + FP}{TP + TN + FN + FP} \tag{4}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{6}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

$$\text{F-Measure} = \frac{2 * (\text{Precision} * \text{Sensitivity})}{\text{Precision} + \text{Sensitivity}} \tag{8}$$

B. Unsupervised Learning

Unsupervised learning is a class of machine learning techniques to find the patterns in data. The data given to unsupervised algorithm are not labelled, which means only the input are given, no corresponding output there. In unsupervised learning, the algorithms are left to themselves to discover pattern in the data.

Unsupervised learning can be of two types:

Clustering: In it data is divided into several groups with similar patterns [6]. Clustering make homogenous data group and each group datasets are similar to another.

Association: It is rule based machine learning technique which find out the relation between parameter of large datasets. Shopping store uses algorithm to find the relationship between products. If model trained well then it can be used to increase the sale of products.

Unsupervised algorithms are:

K-mean

Unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes. The objective of K-means is simple: group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset. A cluster refers to a collection of data points aggregated together because of certain similarities. We define a target number k, which refers to the number of centroids you need in the dataset. A centroid is the imaginary or real location representing the center of the cluster.

Hierarchical clustering

It is one of the popular and easy to understand clustering technique. This clustering technique is divided into two types:

- Agglomerative
- Divisive

Agglomerative Hierarchical clustering Technique

In this technique, initially each data point is considered as an individual cluster. At each iteration, the similar clusters merge with other clusters until one cluster or K clusters are formed, key operation is computation of the proximity of two clusters. It uses bottom-up approach.

Divisive Hierarchical clustering Technique

It is opposite of agglomerative Hierarchical clustering Technique. In it we consider all the data points as a single cluster and in each iteration, we separate the data points from the cluster which are not similar. Each data point which is separated is considered as an individual cluster. In the end, we'll be left with n clusters. It uses top-down approach.

Artificial Neural Network (ANN)

It can be used supervised and unsupervised as well. When we have known about the output class then it will be supervised learning otherwise it will be considering as unsupervised learning.

C. Reinforcement Learning

It teaches the model decide what to do and what not to do when the benefits are maximized, or the hazards are minimized. The algorithm develops new abilities through trial and error and feedback in the form of rewards or penalties.

D. Semi-Supervised Learning

Aspect of both supervised and uncontrolled learning are included in partially supervised learning. It uses both labeled and unlabeled data to improve model performance by utilizing the readily available labeled data and the larger pool of unlabeled data for generalization.

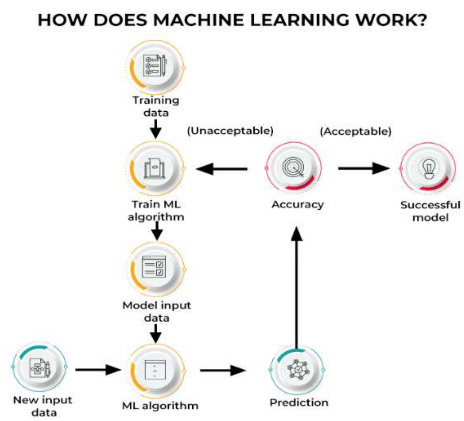

Fig. 6. Machine Learning Workflow

Mainly machine learning has three type viz. supervised learning, unsupervised learning and reinforcement learning. Supervised learning is widely used algorithms, and developer has to trained model to work with it with datasets and implement mapping function from input to output. Unsupervised learning need not to be trained unlike supervised learning. At the time of algorithm implementation developer set characteristics to find special type of pattern [11].

III. Quantitative Research

This study is based on the comparison done between various machine learning techniques used for diabetes disease detection. Here twenty-four researches of different authors have been concluded. This study in Table 1 shows the challenge in each approach and also displays its effect on detection accuracy. Firstly, comparison has been done based on various machine learning techniques used for early disease detection.

Second comparison is done on various deep learning techniques and Ensemble machine learning techniques for early disease detection. Table 1 demonstrates the contrast between the twenty-four researches about early diabetes detection and their research gaps. The study analysis results in various important factors that are relevant in review structure and also its respective domain. Another comparison has been done based on input and output parameters in various research papers as given in Table 1.

TABLE I. SYSTEMATIC COMPARISON OF MACHINE LEARNING APPROACHES IN DIABETES RESEARCH

#	Author	Domain	Input	Output	Review	Gaps
1	Swapna G.et al.[2018]	Deep learning	HRV signals employing convolutional neural network(CNN), Long short-term memory (LSTM) & its combinations	CNN 5-LSTM with SVM architecture gave highest accuracy of 95.7%	CNN and LSTM networks combined with SVM, to detect diabetes using HRV signals from ECG data	Limited sample size & absence of comparison with clinical diagnostic methods.
2	Gowthami S. et al. [2023]	Machine Learning	EHR data having attributes-age, gender, polyuria, polydipsia, sudden weight loss, and other relevant factors.	T Random Forest algorithm achieved the highest accuracy of 98% in detecting T2DM.	System for prediction of Type 2 Diabetes Mellitus (T2DM) using electronic health record data.	Small sample size & impact of additional variables -body size, height, and BMI.
3	Aishwarya Majumdar et al. [2019]	Machine Learning	records of 800 patients with attributes- Glucose, BMI, Age, Insulin etc	Classification of patients into diabetic or non-diabetic group	Model proposed to improve classification accuracy. Logistic Regression has accuracy of 96% &AdaBoost has 98.8%	lack of consideration for external factors beyond traditional ones.
4	Victor Chang et al. [2022]	Machine learning	BRFSS (Behavioral risk factor surveillance system)	work used 5 classifiers- DT, RFA, NB, KNN&LR out of which RFA has highest accuracy with 82.26%	Review of SVM, KNN, ANN, and decision trees, techniques & feature selection importance.	To develop accurate predictive models for early diagnosis
5	Nikos Fazakis et al. [2021]	Machine Learning	ELSA Dataset	predictions for Type 2 Diabetes Mellitus (T2DM) occurrence and appropriate lifestyle intervention guidance.	(ELSA) dataset analyzed for identifying of causal and temporal associations between elderly lifestyle and T2DM.	Small dataset size to give more accurate result increase variety of data.
6	Pronab Ghosh et al. [2021]	Machine learning	PID dataset & Attributes- glucose level, blood pressure, skin thickness, insulin, BMI etc.	Diabetes prediction system based on the Random Forest model of 99.35% accuracy	Analysis using SVM, AdaBoost, and Random Forest. Metrics of specificity, sensitivity and accuracy used. MRMR for feature selection.	Meticulous data selection and investigating pipeline designs for data pre-processing.
7	Bukola Badeji-Ajisafe, et al. [2024]	Machine learning	Pima Indian Diabetes Dataset	Random Forest Classifier gave highest accuracy of 87%	Machine learning algorithms are used to increase prediction accuracy of model and solve issues like imbalance data.	Clinical validation of these models and their incorporation into current healthcare system.
8	Atef Hadi Ataya et al.[2023]	Machine learning	Pima Indians Diabetes Dataset (PIDD)	The LightGBM model outperforms all other models, according to the results, reaching 88.5% of accuracy	Model for diabetes diagnosis, in pregnant women. Machine learning algorithms, such as SVM, K-means clustering & LightGBM, the latter showed higher accuracy of 88.5%	To enhance the accuracy of the predictive method used to diagnose diabetes.
9	Henock M. Deberneh et al. [2021]	Machine learning	The study used medical records of Hanaro Medical foundation in Seoul, South Korea	The output classified patients into three categories: normal, prediabetes, or diabetes	Key predictors such as fasting plasma glucose, HbA1c, and BMI were determined by feature importance ranking which was better than conventional T2D predictors.	HbA1c and the oral glucose tolerance test (OGTT) were neglected; while FPG level alone was used to distinguish between normal, prediabetes, and diabetes.

10	Roshan B irjaiset al.[2019]	Machine learning	Pima Indians Diabetes Dataset (PIDD)with features-age, BMI, and glucose etc.	Gradient boosting is one of the best machine learning methods for regression and classification issues with prediction accuracy of 86% .	Extracting data from healthcare & then . used techniques like Gradient Boosting, Logistic Regression and Naive Bayes with 86%,79% &77% prediction accuracy respectively.	Similar methods can be applied on different disease datasets like cardiovascular or oncology for prediction.
11	Zill E Humaet al.[2024]	Machine learning	Kaggle Diabetes Dataset with features- Glucose, Blood pressure ,Skin thickness Insulin, BMI etc	More accurate & early disease detection model being developed.	Used machine learning algorithms- XG Boost (XGB), Naïve Bayes (NB), and K-Nearest Neighbours (KNN). XGB had the finest accuracy, coming in at 90% & Federated learning's integration keeps patient data private.	To enhance the accuracy of the proposed model.
12	Hasan, D.A et al.[2021]	Machine learning	Labeled retinal images from datasets like Messidor, Kaggle, EyePACS, and IDRiD	classification accuracy, sensitivity, and specificity for detecting diabetic retinopathy	Review shows multiple datasets and evaluation metrics, reflecting the potential of machine learning in supporting physicians for DR diagnosis.	more updated datasets required & also reduce the pre-processing time.
13	J.J. Khanam et al.[2021]	Machine Learning & Deep learning	PID dataset with attributes- Pregnancy, BMI, Insulin level, Age, Blood pressure, Skin thickness, Glucose etc	NN with two hidden layers had highest accuracy rate of 86% in analysing diabetes.	Used ML algos on PID dataset with 768 patients and nine attributes. Research by Sisodia et al. achieved 76.30% accuracy using NB, while Tigga et al. found LR yielding 75.32% accuracy. Paper implemented seven ML algorithms and NN with varying hidden layers and epochs	It still remains a challenge to select logical features and appropriate classifiers for optimal accuracy.
14	Neha Prerna Tigga et al.[2019]	Machine Learning	PID Dataset & other collected dataset with attributes- gender, stress, exercise, BP etc	highest accuracy of 94.10% in RFA on both datasets.	It analyzed data from 952 participants through questionnaires and compared six different ML algorithms, with RF got highest accuracy of 94.10% .The study validates its findings by applying the same algorithms to the PIMA dataset.	This result can be used in future to predict any other ailment.
15	Oladimeji, O.O. et al.[2021]	Machine Learning	520 people data from Sylhet Diabetes Hospital with attributes- gender, age, alopecia, obesity etc.	highest accuracy of 98.30% obtained for RFA	Authors used WEKA software It used RF, J48, NB, and KNN. Feature selection proved crucial. RFA performed optimally, was cost-effective and accessible solution.	future research includes how- body size, height and BMI impact diabetes detection.
16	Lai et al.[2019]	Machine learning	CPCSSN(Canadian Primary Care Sentinel Surveillance Network) with attributes-FBS, SBP, TG, and BMI,	For GBM, LR, RFA & RPART the AROC is 84.7% , 84.0% , 83.4% &78.2% respectrively.	The study showed multiple approaches including (GBM) and LR comparing them with DT and RF methods. The study's strength lies in its	These models can be set up in a computer program online to help physicians in assessing Canadian patients'

			HDL,BMI and triglycerides, age, sex, blood pressure, and LDL etc		large Canadian dataset of 13,309 patients and robust validation methods.	risk of developing DM
17	Farajollahi B et al.[2021]	Machine Learning	(National Institute of Diabetes and Digestive and Kidney Disease) PID dataset-& Attributes-BMI, insulin level, age,	performance of 6 classifiers compared based on accuracy, F1-score, recall, precision and AUC. And Ad boost has the most accuracy 83%.	Taken PID dataset of female patients over 21 years old with various medical predictor variables. The research evaluated six classifiers performance through accuracy, F1-score, recall, precision and AUC metrics.	Earlier focus was on linear classifiers, this study addresses the gap by comparing both linear classifiers and ensemble learning algorithms for diabetes.
18	Aggarwal, S. et al.[2024]	Machine Learning	Sylhet diabetic patients' dataset.- - Age, Sex, Polyuria, Polydipsia,Weakness, Polyphagia, Delayed healing, Partial paresis etc	Out of DT, LR, SVM & RFA, RFA classifier turns out to be the most accurate technique with the accuracy of 99%.	The study evaluates four algorithms - DT, SVM, RF, and LR - using SDH dataset. It showed RF achieves the highest accuracy at 99%, followed by DT at 98%, SVM at 96.1%, and LR at 94.2%	Future research could expand the dataset to include more diverse populations and additional health parameters to improve prediction reliability
19	Katarya, R. et al.[2020]	Machine Learning	Attributes-Pregnancy, Glucose Concentration, BP, Skin Thickness, Diabetes Pedigree Function etc	Study tells that Random forest performs better than other with 84% accuracy, precision 83, recall 76, f1-score 86 and ROC-AUC score 83	Six ML algos analysed using the PIMA diabetes dataset. Evaluating models based on accuracy, precision, recall, f1-score, and AUC metrics.	Further work can be expanded using other ensemble machine learning methods.
20	S. M. nimmagadda et al.[2024]	Machine Learning	Age, gender, family history, glucose levels, BMI, Lifestyle factors etc	k-NN 81.3% accuracy, SVM 70.6% accuracy LR 72% accuracy, RF 88.31% accuracy	finding show that Supervised ML with classification algorithms is most effective for predicting type 2 diabetes.	Future work focuses on improving ML models through techniques like SHAP.
21	Hasan, S.M.M et al.[2020]	Machine Learning	Attributes-Pregnancies, Glucose, BP, Skin Thickness, BMI, Pedigree Function ,Age etc	Research showed that the Extra tree algorithm with the AdaBoost classifier outperforms other classifiers.	It handled missing values, imbalanced data, and feature selection. By using mutual information-based feature selection and Adaptive Boosting with Extra Trees as the base classifier, got 90.5% accuracy 1.	various feature selection, transfer learning and deep learning-based classification techniques can be explored.
22	Nuthakki, P. et al.[2024]	Machine Learning	Attributes-Pregnancies, The concentration of blood sugar, BPD, Serum insulin, BMI, The purpose of diabetes pedigree, Age etc	R F was best- F1 score 0.9864, Precision, Recall, and Sensitivity of 0.9864, & Accuracy of 0.99.	It aims on improving the accuracy of early diabetic forecasting through machine learning. Focusses on developing e-diagnostic system.	exploring new dimensions for improving predictive accuracy and applicability in diverse healthcare sections.
23	Daghistani, T. et al.[2020]	Machine Learning	Region, gender, age, BMI,BP etc.	Particularly Random Forest, significantly outperform statistical approaches for diabetes prediction	Previous research show machine learning's effectiveness in diabetes prediction. Data availability and dataset size were the challenges.	larger dataset can be used in future to evaluate the robustness of the models.
24	Uddin et al.[2019]	Machine Learning	Clinical & demographic data	SVM used in 29 studies, while NB was used in 23 studies. RF had highest accuracy	48 articles analysed that implemented supervised ML algorithms for disease	Future work can be expanded to Chronic disease surveillance &

				in 53% of the 17 studies, followed by SVM which showed top performance in 41% of its application.	prediction. It used databases - Scopus and PubMed. DT outperformed in most clinical and demographic data analyses.	disease risk prediction modelling.

The Table II further examines the various parameters relevant to diabetes detection and comparison done on various techniques used within the domains like machine learning and deep learning along with dataset and accuracy achieved by each of them. The Table II further compares the accuracy level based on different type of dataset.

TABLE II. ACCURACY RESULTS FOR EACH APPROACH

#	Author	Technique used	Dataset	Accuracy
1	Swapna G.et al.[2018]	CNN, LSTM, SVM	Electrocardiogram (ECG) of 40 participants – 20 with diabetes and 20 normal subjects	Among all models highest accuracy of 95.7%was obtained for for CNN 5-LSTM with SVM network.
2	Gowthami S. et al.[2023]	RFA, LR, K-NN, DT, and SVM	EHR Data	98%
3	Aishwarya Majumdar et al. [2019]	Logistic regression& pipeline model using AdaBoost classifier.	records of 800 patients	LR(96%) & AdaBoost classifier(98.8%)
4	Victor Chang et al.[2022]	In this work 5 classifiers used-DT,RF,NB,KNN &LR	BRFSS Dataset	RF has highest accuracy with 82.26%
5	Nikos Fazakis et al.[2021]	NB,DT,RF & LR are employed. Also ensemble learning techniques (stacking, weighted voting, and majority voting) used.	ELSA Dataset	The proposed Weighted Voting LRRFs ensemble model achieved the highest accuracy, with an Area Under the ROC Curve (AUC) of 0.884
6	Pronab Ghosh et al.[2021]	Random forest, AdaBoost, SVM, Gradient Boosting	The Prima Indians diabetes Dataset	Random forest shows highest accuracy of 99.35%
7	Bukola Badeji-Ajisafe, et al. [2024]	Random forest, SVM,LR & KNN	Pima Indian Diabetes Dataset	Random Forest Classifier gave highest accuracy of 87%
8	Atef Hadi Ataya et al.[2023]	LR, K-NN, SVM,RF, XGBoost, and LightGBM	Pima Indian Diabetes Dataset	The LightGBM model has 88.5%.
9	Henock M. Deberneh et al. [2021]	Prediction model like random forest, XGBoost, support vector machine are used	KNHANES dataset	The decision tree model (C5.0) had the best classification accuracy (77.87%)
10	Roshan Birjaiset al.[2019]	Gradient Boosting, Logistic Regression and Naive Bayes,	Pima Indians Diabetes Dataset (PIDD)	Accuracy as 86% for the Gradient Boosting
11	Zill E Humaet al.[2024]	XG Boost (XGB), Naïve Bayes (NB), and K-Nearest Neighbours (KNN).	Kaggle diabetes dataset	XGB(90%)
12	Hasan, D.A et al.[2021]	ANN,CNN,RF,SVM	Messidor, Kaggle, EyePACS, and IDRiD	Accuracies range from 86.22% to over 95% depending on the specific technique and dataset used
13	J.J. Khanam et al.[2021]	DT,KNN,RF,NB,AB,LR,SVM	Pima Indian diabetes (PID) dataset is collected from the UCI Machine Learning Repository	NN showed highest accuracy of 86%
14	Neha Prerna Tigga et al.[2019]	NB,RFA, LR, K-NN, DT, and SVM	Pima Indian diabetes & dataset collected from online questionnaire	highest accuracy of 94.10% in RFA on both datasets.
15	Oladimeji, O.O. et al.[2021]	RFA,NB & KNN both on cross validation & split method.	Record of 520 people from Sylhet Diabetes Hospital of Sylhet in Bangladesh.	highest accuracy of 98.30% obtained for RFA
16	Lai et al.[2019]	Gradient Boosting Machine and Logistic Re gression	CPCSSN- Canadian population	For GBM AROC is highest -84.7%
17	Farajollahi B et al.[2021]	LR, DT, (SVM), xgboost, RF and Adaboost.	(National Institute of Diabetes and Digestive and Kidney Diseases-)PID dataset	Adaboost has the most accuracy 83%.
18	Aggarwal, S. et al.[2024]	LR, SVM,RF &DT	Sylhet diabetic patients' dataset.	RFA classifier was most accurate technique with the accuracy of 99%

19	Katarya, R. et al.[2020]	K-NN, DT, NB, SVM, LR and RF.	PIMA diabetes dataset	RF highest accuracy of 84%
20	S. M. nimmagadda et al.[2024]	KNN,SVM,LR,RF	Kuwait Health Network (KHN) dataset,Pima Indian Diabetes Dataset (PIDD), Egyptian National Investigation Center database	RF 88.31% accuracy
21	Hasan, S.M.M et al.[2020]	DT,RF,ET	Pima Indians Diabetes Dataset (PIDD)	ADB classifier has highest accuracy of 88.3%
22	Nuthakki, P. et al.[2024]	RF, GB, DT,KNN,LR, and SVM	PID dataset-females	RF showed accuracy of 99%
23	Daghistani, T. et al.[2020]	LR & RF	MNGHA databases,	RF has-TPR 0.88, FNR 0.18, Precision 0.88, Recall 0.88, AUC0.94 F-measure 0.87
24	Uddin et al.[2019]	SVM,LR,KNN,NB,DT	databases - Scopus and PubMed	SVM showed highest accuracy in 41% of its applications

IV. Scope of Future Work

The study emphasizes on developing accurate predictive models for early diabetes diagnosis. The future scope of this research work includes increasing sample size & involve variety of data. Future work will also aim at comparing the predictive analysis with clinical diagnostic methods. Further research will concentrate on meticulous data selection, pipeline designs, and clinical validation. The method can be applied to various disease datasets, such as cardiovascular and oncology etc.

Fig.7 shows cumulative usage of different machine learning models like SVM, RF, KNN, NB, LR, ANN, GB etc.

Fig.8 compares the maximum accuracy achieved by different ML models such as SVM, RF, KNN, NB, LR, ANN, GB etc. where RF classifier achieved highest accuracy in most of the studies.

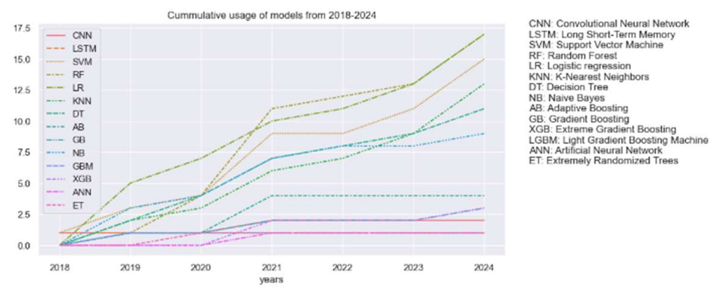

Fig 7 Cumulative usage of models from 2018-2024

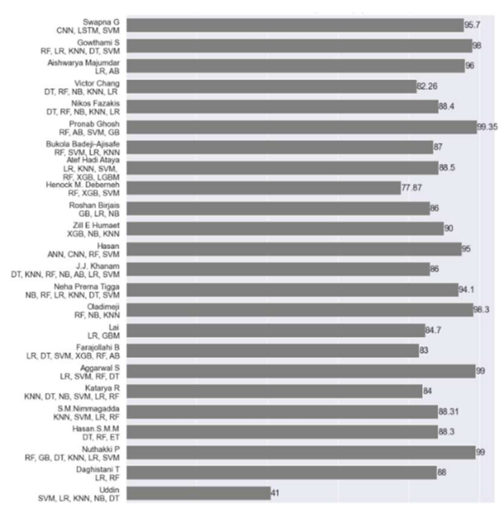

Fig 8 Maximum accuracy achieved by models

V. Conclusion

This comprehensive review of machine learning approaches for diabetes detection and prediction underscores the transformative potential of AI-driven technologies in diabetes care. The studies examined demonstrate that various machine learning algorithms can achieve high accuracy in predicting and detecting diabetes, often surpassing traditional diagnostic methods. Researchers have successfully utilized a wide range of data sources, from electronic health records to specialized datasets, showcasing the versatility of these approaches. Key predictors such as fasting plasma glucose, HbA1c, and BMI have consistently emerged as significant factors in accurate diabetes prediction.

The integration of deep learning techniques with traditional machine learning methods has shown particularly promising results in analyzing complex data like ECG signals. However, challenges remain, including limited sample sizes in many studies and the need for more extensive clinical validation. Future research should focus on expanding studies to include larger and more diverse patient populations, developing frameworks for seamlessly integrating these models into existing healthcare systems, and enhancing the interpretability of machine learning models. As these technologies continue to evolve and address current limitations, they have the potential to significantly improve early diagnosis, personalized treatment, and overall management of diabetes, ultimately leading to better patient outcomes and reduced healthcare costs.

References

[1] Swapna, G., Vinayakumar, R. and Soman, K.P., 2018. Diabetes detection using deep learning algorithms 1. ICT Express, 4(4), pp.243-246 1.

[2] S, G., Reddy, R.V.S. and Ahmed, M.R., 2024. Exploring the effectiveness of machine learning algorithms for early detection of Type-2 Diabetes Mellitus 1. Measurement: Sensors, 31, p.1009833 7 Available at: https://www.sciencedirect.com/journal/measurement-sensors 1. This open access article is published under the CC BY-NC-ND license 1. The paper discusses the application of machine learning techniques for early detection of Type 2 Diabetes Mellitus, highlighting the potential of these methods in improving healthcare outcomes and decision-making processes

[3] Mujumdar, A. et al., 2019. Diabetes Prediction using Machine Learning Algorithms 1. Procedia Computer Science, 165, pp.292-299 1. Available at: https://www.sciencedirect.com/science/article/pii/S1877050920300478

[4] Chang, V., Ganatra, M.A., Hall, K., Golightly, L. and Xu, Q.A., 2022. An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators. Healthcare Analytics, 2, p.100118.

[5] Fazakis, N. et al., 2021. ML Tools for Long-Term Type 2 Diabetes Risk Prediction. IEEE Access, 9, pp.103747-103763.

[6] Ghosh, P., Azam, S., Karim, A., Hassan, M., Roye, K. and Jonkman, M., 2021. 'A Comparative Study of Different Machine Learning Tools in Detecting Diabetes'. Procedia Computer Science, 192, pp. 467-477

[7] Badeji-Ajisafe, B., Obamiyi, S.E., Oguntimilehin, A., Idowu Victor, A. and Okebule, T., 2024. Early Detection of Diabetes Using Supervised Learning Approach. In: International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)

[8] Hasan, M.K., Alam, M.A., Das, D., Hossain, E. and Hasan, M., 2020. Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, pp.76516-76531.

[9] Deberneh, H.M., Kim, I., Park, J.H., Cha, E., Joung, K.H., Lee, J.S. and Lim, D.S., 2021. Machine Learning-Based Prediction of Type 2 Diabetes Mellitus Occurrence. International Journal of Environmental Research and Public Health, 18(6), p.3317.

[10] Birjais, R., Mourya, A.K., Chauhan, R. and Kaur, H., 2019. Prediction and diagnosis of future diabetes risk: a machine learning approach. SN Applied Sciences, 1(9), p.1112

[11] Huma, Z.E., Tariq, N. and Zaidi, S 2024. Predictive Machine Learning Models for

[12] Early Diabetes Diagnosis: Enhancing Accuracy and Privacy with Federated Learning. Journal of Computing & Biomedical Informatics, 8(1), pp.645-0801/2024

[13] Hasan, D.A. et al., 2021. Machine Learning-based Diabetic Retinopathy Early Detection and Classification Systems- A Survey. In 1st Babylon International Conference on Information Technology and Science (BICITS). IEEE, pp. 16-20

[14] Khanam, J.J. and Foo, S.Y., 2021. A comparison of machine learning algorithms for diabetes prediction. ICT Express, 7(4), pp.432-439

[15] Tigga, N.P. and Garg, S. (2020) 'Prediction of Type 2 Diabetes using Machine Learning Classification Methods', Procedia Computer Science, 167, pp. 706-716. doi: Available at: http://creativecommons.org/licenses/by-nc-nd/4.0

[16] Oladimeji, O.O., Oladimeji, A. and Oladimeji, O. (2024) 'Classification models for likelihood prediction of diabetes at early stage using feature selection', Applied Computing and Informatics, 20(3/4), pp. 279-286

[17] Lai, H., Huang, H., Keshavjee, K., Guergachi, A. and Gao, X. 1, 2019. 'Predictive models for diabetes mellitus using machine learning techniques', BMC Endocrine Disorders, 19(101)

[18] Farajollahi, B., Mehmannavaz, M., Mehrjoo, H., Moghbeli, F. and Sayadi, M. (2021) 'Diabetes Diagnosis Using Machine Learning', Frontiers in Health Informatics, 10(65), pp. 1-5

[19] Aggarwal, S., Mittal, A., Sarangi, P.K., Agrawal, A.K., Sahoo, A.K. and Pattanaik, B. (2024) 'Performance Analysis and Comparison of Machine Learning Models in Early Diabetes Prediction', 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Amity University, Noida, India, 14-15 March
[20] Katarya, R. and Jain, S. (2020) 'Comparison of Different Machine Learning Models for diabetes detection', IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering (ICADEE 2020)
[21] Nimmagadda, S.M., Suryanarayana, G., Kumar, G.B., Anudeep, G. and Sai, G.V. (2024) 'A Comprehensive Survey on Diabetes Type-2 (T2D) Forecast Using Machine Learning', Archives of Computational Methods in Engineering, 31, pp. 2905-2923
[22] Hasan, S.M.M., Rabbi, M.F., Champa, A.I. and Zaman, M.A., 2020. An Effective Diabetes Prediction System Using Machine Learning Techniques. In: 2nd International Conference on Advanced Information and Communication Technology (ICAICT)
[23] Nuthakki, P. and Kumar, T.P. (2024) 'Machine learning-based early detection of diabetes risk factors for improved health management', Multimedia Tools and Applications
[24] Daghistani, T. and Alshammari, R., 2020. Comparison of Statistical Logistic Regression and RandomForest Machine Learning Techniques in Predicting Diabetes. Journal of Advances in Information Technology, 11(2), pp.78-83
[25] Uddin, S. et al. (2019) 'BMC Medical Informatics and Decision Making', 19:281
[26] https://www.statisticssolutions.com/what-is-logistic-regression/
[27] https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/
[28] https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python
[29] https://www.saedsayad.com/naive_bayesian.htm
[30] https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb
[31] Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J. P. (2012) "An assessment of the effectiveness of a random forest classifier for land-cover classification." ISPRS Journal of Photogrammetry and Remote Sensing 67: 93-104.
[32] Understanding the Confusion Matrix in Machine Learning - GeeksforGeeks
[33] Käräjämäki, A.J., Bloigu, R., Kauma, H., Kesäniemi, Y. A., Koivurova, O. P., Perkiömäki, J., Huikuri, H., Ukkola, O. (2017) "Non- alcoholic fatty liver disease with and without metabolic syndrome: different long-term outcomes." Metabolism 66: 55-63.
[34] Gurka, M.J., Golden, S. H., Musani, S. K., Sim, M., Vishnu, A., Guo, Y., Cardel, M., Pearson, T.A., DeBoer, M.D. (2017) "Independent associations between a metabolic syndrome severity score and future diabetes by sex and race: the Atherosclerosis Risk In Communities Study and Jackson Heart Study." Diabetologia 60 (7): 1261-1270.
[35] Laakso, M., Kuusisto, J., Stančáková, A., Kuulasmaa, T., Pajukanta, P., Lusis, A. J., Collins, F. S., Mohlke, K. L., Boehnke, M. (2017) "The Metabolic Syndrome in Men study: a resource for studies of metabolic and cardiovascular diseases." Journal of lipid research 58 (3): 481- 493.