Enhancing the Diagnosis of Diabetes Through Intelligent Machine Learning Models

[1] Muni Balaji Thumu
Lecturer,
Department of Information Technology,
University of Technology and Applied Sciences-Salalah,
Salalah,
Sultanate of Oman.
Email : muni.balaji@utas.edu.om

[2] T. Karthikeyan
Lecturer,
Department of Information Technology,
University of Technology and Applied Sciences-Salalah,
Salalah,
Sultanate of Oman.
Email: karthikeyan.t@utas.edu.om

[3] Mithun Vinayaka Kulkarni
Senior Lecturer,
Department of Engineering and Technology,
University of Technology and Applied Sciences-Salalah,
Salalah,
Sultanate of Oman.
Email: Mithun.Kulkarni@utas.edu.om

[4] Syed Khaja Mohideen
Lecturer,
Department of Information Technology,
University of Technology and Applied Sciences-Salalah,
Salalah, Sultanate of Oman
Email : syed.mohideen@utas.edu.om

***Abstract*:** Global diabetes prevalence is growing. Early diabetes detection can help avoid type 2 diabetes and prediabetes by prompting lifestyle adjustments. Thus, fast, accurate diagnostic tools are needed. Several studies are improving machine learning-based disease diagnosis speed, accuracy, and efficacy. Early disease diagnosis may benefit from machine learning. This study predicted diabetes using National Institute of Diabetes and Digestive and Kidney Diseases diagnostic (NIDDK) data. Eight patient-specific traits predicted diabetes. Of the 2000 people evaluated, 684 (34.2%) had diabetes and 1316 (65.8%) did not. The study used 10 machine learning algorithms to predict, and Random Forest was the most accurate at 98%. Other algorithms showed 81%–89% accuracy. This work presents an effective machine learning diabetes identification method that can help with early diagnosis.

***Keywords*:** *Diabetes, Detection, Machine learning, Diagnosis, Classification*.

I. INTRODUCTION

When insulin is either ineffective or the pancreas is unable to make enough, the result is chronic diabetes. Because the body can't metabolize glucose from food, blood sugar rises [1]. Chronically high blood sugar can damage the kidneys, heart, blood vessels, nerves, and eyes [2]. Monitor blood sugar to avoid problems. Although many don't recognize they have diabetes, early identification is essential for optimal treatment [3]. Diabetes can be Type I, Type II, or gestational. IDDM, often known as type I diabetes, requires insulin shots since the body cannot produce enough insulin [4]. Type II diabetes occurs when cells can't use insulin effectively. Gestational diabetes may occur in pregnant women with high blood sugar without a diagnosis [5]. Nearly 90% of cases have Type 2 diabetes, the most common type. As a quiet disease, its symptoms may go unnoticed for years [6]. Obesity and genetics increase Type 2 diabetes risk [7]. Many at-risk adults can postpone its onset with early diagnosis and lifestyle or medication changes [8-10]. For proactive diabetes management, early detection is crucial.

An increasing proportion of people worldwide are developing diabetes. Poor nutrition, obesity, aging, the rise in sedentary lives brought about by transportation and technological improvements, and the pervasiveness of computers, the internet, cell phones, and tablets in daily life are some of the factors contributing to this spike. The ongoing stress of contemporary work situations also contributes. 450 million individuals worldwide suffer with diabetes at the moment [11]. The situation is especially concerning in Turkey, where there have been over 10 million cases diagnosed. Turkey has almost twice the global average for the prevalence of diabetes per capita. In actuality, Turkey has Europe's highest diabetes growth rate. According to research from 2015, one in six people in Turkey suffers from this illness [11]. Diabetes treatment necessitates not just substantial financial resources but also careful consideration and care.

Data mining is becoming more widespread in various businesses as computers can handle more information [12], [13-21]. These procedures aid diagnosis and treatment, making them crucial to healthcare. Diagnoses are challenging and call for precise patient information, familiarity with medical literature, and clinical expertise. Medical situations are unpredictable, hence this diagnosis method requires more finesse than in other fields [22]. Clinical decisions sometimes depend on doctor and past knowledge. One issue is that people may misdescribe their symptoms. The amount of data accessible may also make decision-making harder.

Data mining and ML are becoming more important in diabetes research. Machine learning is becoming more and more popular as a tool for illness diagnosis prediction, particularly with diabetes. For example, Sowjanya et al [23] created an Android app that offers helpful information and guidance regarding diabetes while utilizing machine learning to estimate a user's risk of developing the condition. Using the decision tree algorithm, Orabi et al [24] developed a system to evaluate diabetes risk and achieved remarkable results. The Random Forest technique was shown to be the most effective

algorithm for diabetes risk prediction when Nongyao et al. [25] compared various methods. Humar et al. [26] created a hybrid system that merges artificial neural networks with fuzzy neural networks to achieve a diabetes diagnosis accuracy of 79.16%. Mohammed et al [27] achieved an accuracy of 86.13% in the detection of diabetes by combining SVR with the NSGA-II approach. With the AdaBoost classifier, Mujumdar and Vaidehi [28] reported a high accuracy of 98.8%. In their individual investigations, Faruque et al. [18] and Sonar and Jaya Malini [18] employed the C4.5 decision tree, achieved 73.5% and 85% accuracy. Using the random forest approach, Zou et al [29] showed an 80.8% accuracy rate in diabetes prediction. With the SVM-linear model, accuracy of Kaur and Kumari [21] was an astounding at 89%. Finally, Acar et al [30] achieved an accuracy rate of 87.06% in their investigation on biometric measures using the LS-SVM approach. In conclusion, a number of researchers are using sophisticated algorithms to improve the precision and effectiveness of diabetes prediction.

This research makes use of the Diabetes dataset that is available on Kaggle. All of the diabetes-related data is stored at the National Institutes of Health. The dataset was created with the express purpose of facilitating diabetes research and contains information on Pima Indian women residing in Phoenix, the fifth-largest city in Arizona, of 21 years old or older. Within the dataset, there are eight distinct numerical parameters and 2000 observations. Data on the target parameter result, which is defined as 1 denoting a positive result from a diabetes test and 0 as a negative result, is given in Table 1. In addition, the data type description, name, and role of each are listed in the table. The main objective of this research is to develop a predictive model that can reliably use clinical characteristics to identify women who are at risk of developing diabetes. The model should be very sensitive and specific. The study achieved this by comparing the accuracy of diabetes prediction using ten different machine learning algorithms. Here are the algorithms that were used: XGB, Gradient Boosting, Decision Tree, Logistic Regression, KNN, Support Vector Machine, Adaptive Boosting, Ensemble model, and Random Forest. The order of accuracy is as follows: LGBM, Adaptive Boosting, Decision Tree, Logistic Regression, KNN, and Ensemble model. The accuracy percentages and testing results of each of these algorithms were carefully compared in the study. This study differs from others in that it takes a comprehensive strategy that includes a wide range of algorithms in order to identify the most effective algorithm for predicting diabetes using clinical data.

II. EXPLORATORY ANALYSIS OF DATA

Exploratory data analysis (EDA) is a technique that focuses on visually examining datasets to highlight their key features and unearth information not found through conventional modelling or hypothesis testing [31]. Conducting early research on data using graphical representations and visualizations is the main objective of EDA. This helps in identifying of patterns, the formulation of hypotheses, and the verification of assumptions. In EDA, data visualizations are essential because they provide comparative and explanatory charts that efficiently communicate both tangible and abstract notions. EDA can identify possible problems by emphasizing important details and exposing hidden patterns in the data. By addressing these problems, diabetes diagnosis accuracy and other data-driven decisions can be improved.

A. Understanding and Visualizing Data

In this research, the data from the diabetic Kaggle dataset [32] was modelled and tested. NIDDK has a bigger dataset of which the selected dataset is a component. Many researchers [14], [17], [20], [21], and [29] have used this data set in predictive analyses. This dataset examines diabetes among women of Pima Indian heritage who live in Phoenix, the fifth-largest city in the U.S. state of Arizona, and are at least 21 years old. There are eight independent numerical parameters and 2000 observations in the data set. The result of a diabetes test is the goal variable; a positive result is 1 and a negative result is 0. Table 1 displays the name, type, and purpose of the data.

Table 2 provides an overview of the dataset with summary statistics. Both distributional measurements like the standard deviation and measures of central tendency like the mean and median are included in this. Realistically, there are between 0 and 17 pregnancies in the statistics. Some variables, however, have recorded values of 0, which are practically improbable. These attributes are Skin Thickness, Blood Pressure, Insulin, BMI, and Glucose. During the data pre-processing stage, these erroneous zero values were changed to reflect their corresponding mean values in order to remedy this. The 'DiabetesPedigreeFunction' gives a score between 0.08 and 2.42, which indicates a person's risk of getting diabetes based on their family history. In actuality, the age characteristic might be anywhere from 21 and 81 years old. For individuals with a diabetes diagnosis, the aim variable result has a value of 1, whereas for those without, it is 0.

Data visualization is the graphical depiction of quantitative data for analysis and communication [33]. As data becomes more varied and complicated, analytical and presentation approaches that can condense complex data into intelligible visuals are needed [34–35]. Analysts utilize charts to quickly spot patterns, trends, and anomalies without a model or hypothesis testing [36]. The data set's visualizations demonstrate variables' relationships, patterns, trends, and outliers, and simplify complex data in plain, concise pictures. Fig. 1 reveals 1316 (65.8%) of 2000 observations were healthy and 684 (34.2%) had diabetes. Fig. 2 displays glucose distribution and Fig. 3 provides dataset attribute boxplots.

Heatmaps are helpful when studying multivariate data. They are able to demonstrate differences between variables, point out commonalities between them, and establish connections. Such information is provided by the heatmap of the dataset in Fig. 4. From the heatmap we observe that, a variable's correlation with itself is shown by the diagonal line that extends from top-left to bottom-right, and it is always a perfect 1. The variables' direction and degree of correlation are shown by color intensity and values. Darker blues suggest unfavorable associations. Darker red hues are indicative of positive relationships. There is a significant positive association between our goal variable, the Outcome, and glucose. This implies that diabetes is more common in people with higher blood glucose levels. there is a positive link between age and pregnancies, suggesting that the likelihood of becoming pregnant rises with age. Most of the features show little association with one other, which is beneficial since it reduces issues about multicollinearity.

B. Data pre-processing

The process of arranging, purifying, and converting unprocessed data into a format suitable for modeling, analysis,

and decision-making is known as data preparation.. Encoding categorical variables, handling outliers, resolving missing values, normalizing or standardizing data scales, and choosing pertinent features are some of the methods used. The primary goal of data preprocessing is to improve the quality of the data since this ensures reliable and accurate results from subsequent data analysis or machine learning models.

A number of characteristics, including blood pressure, skin thickness, insulin, glucose, and BMI, have zero values after preprocessing, which is not consistent with the dataset. NaN for the corresponding properties is thus used in place of these zero values. The median was then subtracted from the Nan values in the corresponding characteristics. The data was given a mean of zero and a standard deviation of one.

C. Feature Scaling

The data's independent variables, or features, are standardized by the feature scaling method of data pre-processing. The following are two popular feature scaling methods: Z-score normalization, sometimes referred to as standardization, is used to scale features to match a normal distribution with a mean of 0 and a standard deviation of 1. Characteristics are typically scaled between 0 and 1 using the minimum-maximum scaling approach.

In this research, standardization approach is used for performing feature scaling. Mean and standard deviation for the scaled features is approximately equal to 0 and 1 respectively. Table 3 shows some attributes after scaling. The target attribute 'Outcome' was not scaled as it is a categorical variable.

D. Feature Engineering

Feature engineering creates new features from existing ones to improve machine learning models or extract more meaningful data. New characteristics after engineering are, Based on BMI, patients are classified as 'Underweight', 'Normal', 'Overweight', or 'Obese'. Patients are classified by age: 'Young' (Age < 25), 'Middle-aged' (25 <= Age < 60), and 'Senior' (Age >= 60). Patients are classified as 'Low' (below median) or 'High' (above median) based on their median insulin levels. The glucose-BMI connection is represented by glucose_BMI. The glucose-age interaction is Glucose_Age. Table 4 lists the new feature's name, category, and description. Finally, categorical features were quantified.

E. Handling Outliers

Outliers are observations that differ considerably from the remainder of a dataset. They may have unusually high or low values that depart from the data's distribution. Outliers can come from measurement, data entry, or variability errors. Even though they may provide useful data, outliers can distort statistics and confuse data analysis. Identifying and managing outliers in data pre-processing ensures the accuracy and reliability of subsequent studies or predictive modeling.

Interquartile Range (IQR) eliminated outliers in this study. IQR is the middle 50% of a dataset's distribution. Formula (1) subtracts the first quartile (Q1) from the third quartile (Q3). The 25th percentile of the dataset is Q1, meaning 25% of data points are below it. Q3: is the dataset's 75th percentile, meaning 75% of data points are below it.

$$IQR = Q3 - Q1 \qquad (1)$$

It was being identified that the dataset has 1085 observations as outliers. They are removed and the new dataset has 915 observations with 11 attributes. The missing values in the dataset are handled. In this research we used regression algorithm to handle missing values because it maintains the correlation between the dataset's attributes.

TABLE I. DATASET ATTRIBUTES AND DESCRIPTION USED FOR DETECTION OF DIABETES

S. No.	Attribute	Type	Description
1.	Pregnancies	Numerical	The number of pregnancies
2.	Glucose	Numerical	The oral glucose tolerance test's plasma glucose level two hours later
3.	BloodPressure	Numerical	Blood pressure Diastolic (mm Hg)
4.	SkinThickness	Numerical	Skin fold thickness in the triceps (mm)
5.	Insulin	Numerical	Two-hour serum insulin levels (mu U/ml)
6.	BMI	Numerical	Body mass index (weight in kg/(height in m)^2)
7.	Diabetes PedigreeFunction	Numerical	This function uses a descendant's family history to calculate their risk of developing diabetes.
8.	Age	Numerical	Age (years)
9.	Outcome	Categorical	Class variable (disease status (1) or absence (0))

TABLE II. DESCRIPTION OF THE DATA IN THE DATAFRAME

Attributes	count	mean	std	min	25 %	50 %	75 %	max
Pregnancies	2000	3.70	3.30	0	1	3	6	17
Glucose	2000	121.18	32.06	0	99	117	141	199
Blood Pressure	2000	69.14	19.18	0	63.50	72	80	122
Skin Thickness	2000.	20.93	16.10	0	0	23	32	110
Insulin	2000	80.25	111.18	0	0	40	130	744
BMI	2000	32.19	8.14	0	27.37	32.30	36.8	80.6
DiabetesPedigreeFunction	2000	0.47	0.32	0.07	0.24	0.37	0.62	2.42
Age	2000	33.09	11.78	21	24	29	40	81
Outcome	2000	0.34	0.47	0	0	0	1	1

TABLE III. SCALED FEATURES

Feature	Mean	Standard Deviation
DiabetesPedigreeFunction	1.23e-16	1.00025
BMI	6.39e-17	1.00025
Insulin	-1.78e-18	1.00025
Glucose	7.82e-17	1.00025
Age	1.14e-16	1.00025

TABLE IV. NEW FEATURES AFTER ENGINEERING

New Features	Type	Description
BMI_Category	category	0-18 : 'Under Weight', 18-24 : 'Normal', 24-29 : 'Overweight', >29 'Obese'

Age_Group	category	'Young' (Age < 25), 'Middle-aged' (25 <= Age < 60),
'Senior' (Age >= 60 Description)		
Insulin_Level	category	'Low' (Below median), 'High' (Above median)
Glucose_Age	float	Represents the interaction between Glucose and Age

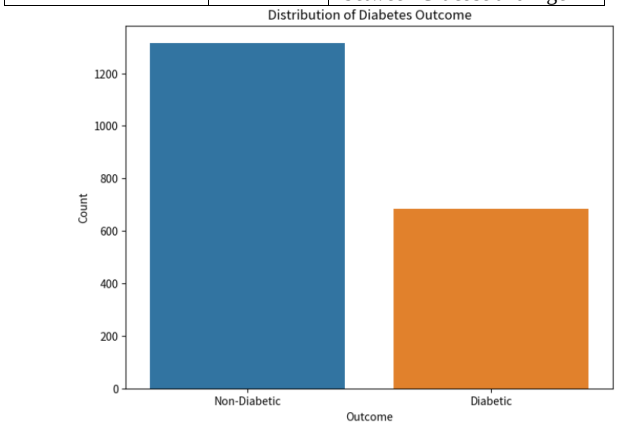

Fig. 1. Distribution of Outcomes Visualized

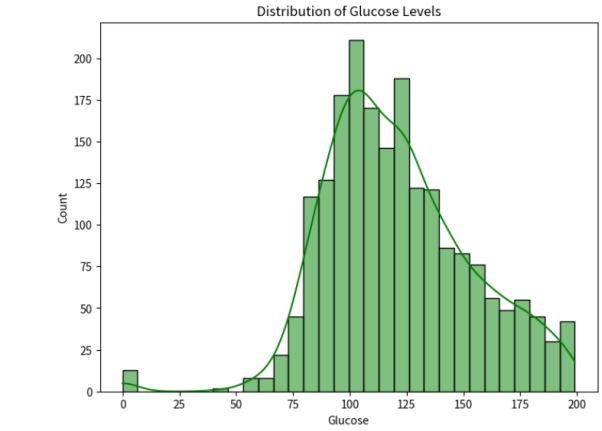

Fig. 2. Visualization of Glucose Distribution

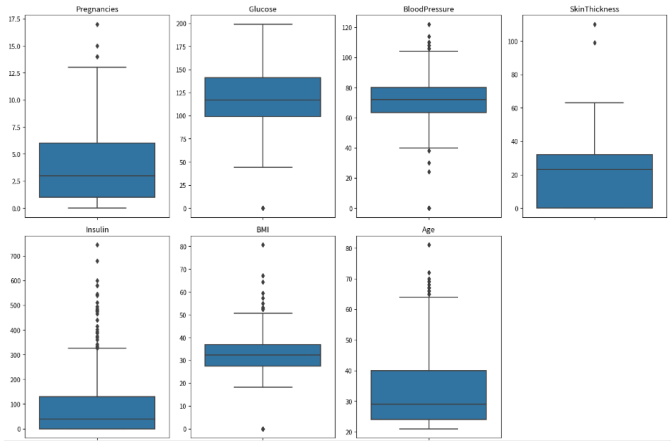

Fig. 3. Boxplots for key variables

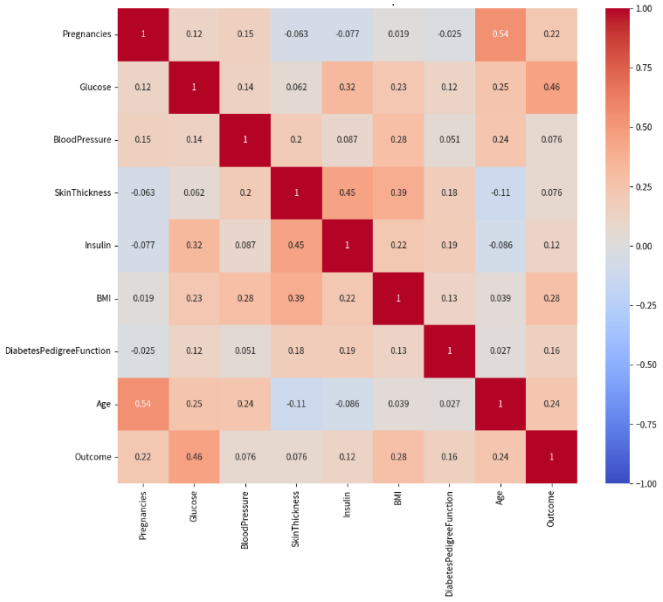

Fig. 4. Correlation heatmap for the dataset

III. EXPERIMENTAL ENVIRONMENT AND EVALUATION METRICS

Various methods, such as Random Forest (RF), XGB, Gradient Boosting (GB), LGBM, AdaBoost, Decision Tree, Logistic Regression, Support Vector Machine (SVM), KNN, and Ensemble model, were used to evaluate the dataset. These algorithms were chosen because they have been used extensively in the literature and have demonstrated the capacity to produce results on this particular dataset that are comparatively better than others.

A. Random Forest

Multi-decision tree Random Forest machine learning improves prediction accuracy [37]. During training, it constructs multiple trees to calculate classification mode or regression mean prediction. Random Forest corrects decision tree overfitting to training data. It performs effectively when dealing with both numerical and category data. Throughout the construction process, features are chosen at random to provide a variety of trees. Both regression and classification tasks can be performed with the algorithm. It provides information on the significance of features as well.

B. Decision Trees

A popular machine learning algorithm for classification and regression applications is the decision tree [38]. They produce a decision-tree-like model by iteratively dividing the dataset into subsets according to the most important features. Each internal node of the tree represents a decision taken in response to a feature's value, whereas each leaf node of the tree shows a result or prediction. There can be variations in the tree's depth and complexity; deeper trees generally capture more complicated patterns. However, deep trees may cause overfitting, where the model performs well on training data but poorly on new data. Techniques for pruning are frequently used to shape the tree and avoid overfitting.

C. Gradient Boosting

Gradient Boosting is an ensemble machine learning technique for regression and classification. Every tree is planted with knowledge from other trees that have already

developed. The fundamental concept is to reduce error and identify the desired results for upcoming model. This method depends on the development of predictions in the future by using lessons from the past mistakes [12]

D. eXtreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) improves gradient boosting. It is well-known for its speed and efficiency and builds trees using parallel processing. To avoid overfitting, it integrates L1 and L2 regularization and handles missing values naturally, obviating the requirement for imputation. Furthermore, XGBoost can be used to solve a wide range of predictive issues, such as regression, classification, and ranking [39-41].

E. Light Gradient Boosting Machine (LGBM)

LGBM is a fast and effective gradient boosting architecture based on tree-based algorithms. LGBM grows trees leaf-wise rather than level-wise, which can result in improved accuracy compared to other tree-based algorithms. It handles big datasets with ease and supports both numerical and categorical features. The framework is renowned for its increased efficiency and quicker training periods. With multiple hyperparameters for customization, LGBM can handle imbalanced datasets by varying the class weights. Because of its scalability and performance, it is frequently utilized in a variety of machine learning contests and real-world applications [42,43].

F. Adaptive Boosting (AdaBoost)

AdaBoost machine learning algorithm objective is to make a strong classifier by improving the performance of weak ones [44]. To make sure that erroneously categorized data points are given priority by later classifiers, iteratively modifies their weights. Based on accuracy, each classifier is given a weight, and the final prediction is the result of a weighted vote from all classifiers. AdaBoost's power is in its capacity to create models with lower bias and variance by combining several weak learners, most commonly decision trees. Because of its adaptive qualities, which enable it to adapt to intricate data patterns, it is an effective tool for a variety of categorization tasks.

G. Support Vector Machine (SVM)

The supervised machine learning algorithm SVM is typically used for regression and classification applications [45]. It works by identifying the hyperplane that most effectively partitions a dataset into classes. The hyperplane with the largest margin between the two classes is the optimum one. SVMs are flexible and efficient in high-dimensional spaces because they can handle both linear and non-linear relationships by utilizing different kernel functions [25]. The fundamental idea is to maximize the margin between the classes' closest data points, or support vectors. SVMs are a popular option for a variety of applications because they are resistant to overfitting, particularly in high-dimensional environments [26].

H. Logistic Regression

Logistic regression is a statistical technique that simulates the probability of a binary event depending on several predictor parameters [46]. Logistic regression forecasts the probability that a given instance will fall into a specific category, as opposed to linear regression, which forecasts continuous values. It squeezes a linear equation's output between 0 and 1 using the logistic function. Maximum likelihood estimation is used to estimate the predictor variable coefficients from the training data. In fields like economics, social sciences, and medicine, logistic regression is widely used due to its interpretability and simplicity of use. It functions as a foundational algorithm for binary classification-related machine learning tasks.

I. K-Nearest Neighbours (KNN)

By identifying the predefined number of training samples that are closest to a new point in terms of distance, KNN, a non-parametric, slow learning algorithm used for classification and regression applications [12], predicts the label based on the majority class of the neighbors. The user-defined constant 'K' denotes the number of neighbours. Typically, techniques like Manhattan distance or Euclidean distance are used to compute distance. Since KNN makes no assumptions on the distribution of the underlying data, it is adaptable. Its performance, however, can be greatly affected by the selection of 'K'. Since KNN is distance-sensitive, normalizing the data is crucial. KNN stores the complete dataset because it is a lazy learner, which can be computationally demanding for big datasets.

J. Ensemble Model

An ensemble model is a method that provides a single prediction output through the integration of various machine learning models. The main objective is to use the advantages of each unique model to increase prediction accuracy, stability, and performance. Ensemble techniques often yield better results than a single model alone, especially when multiple models are used, especially when each model has unique strengths and shortcomings. Techniques like bagging, boosting, and stacking are frequently used in groups. While boosting concentrates on training models sequentially to rectify the faults of prior ones, bagging decreases variance by training models on distinct subsets of data. In order to provide final predictions, stacking entails training a meta-model using the outputs of separate models. The parameters used in 10 machine learning algorithms are provided in Table 5.

TABLE V. MACHINE LEARNING ALGORITHMS WITH PARAMETERS

Algorithm	Parameters
Random Forest	{'max_depth': None, 'min_samples_leaf': 1,'min_samples_split': 2, 'n_estimators': 150}
Decision Tree	{'max_depth': 10, 'criterion': 'gini', 'min_samples_ leaf': 1, 'min_samples_split': 2, }
Gradient Boosting	{'learning_rate': 0.1, 'criterion': 'friedman_mse', 'subsample': 1.0, 'max_depth': 3, 'n_estimators': 100 }
XGB	{'learning_rate': 0.3, 'booster': 'gbtree', 'n_estimators': 100, 'max_ depth': 4 }
LGBM	{ 'boosting_type': 'gbdt', 'learning_rate': 0.1, 'max_ depth': -1, 'subsample': 1.0, 'n_estimators': 100}
AdaBoost	{'n_estimators': 100, 'learning_rate': 1 }
SVM	{'cache_size': 200, 'C': 1.0, 'coef0': 0.0, 'max_iter': -1, 'kernel': 'rbf',}
Logistic Regression	{'tol': 0.0001, 'C': 1.0, 'max_iter': 1000 }
KNN	{'metric': 'minkowski', 'n_neighbors': 5, 'leaf_size': 30}
ensemble_model	{estimators= [('lr', logreg_model), ('knn', knn_model), ('svm', svm_model), ('dt', dt_model), ('ada', ada_model), ('gb', gb_model), ('xgb', xgb_model), ('lgbm', lgbm_model)), voting='hard'}

K. Measurementl

The metrics used to measure classification performance are true positive (TP), true negative (TN), false negative (FN), and false positive (FP), following the procedure given in [47,48].

IV. Results And Discussion

Random forest group learning builds numerous decision trees during training. Tuning Random Forest with hyperparameters improves model performance. Random forest hyperparameters include: n_estimators: Number of trees, max_depth: Maximum depth, min_samples_split: Minus samples needed to split a node, min_samples_leaf: Minus samples at leaf nodes, max_features: The best split has "log2," "auto," or "sqrt." features. bootstrap: Develop a tree using bootstrap samples. oob_score: Criteria for accuracy estimation: Split measure quality—"mse" for regression, "gini" or "entropy" for classification. Grid Search: Methodically investigates hyperparameter combinations, Randomised Search: Tests any hyperparameter combination within limitations. cross validation: repeatedly splits data into training and validation sets. Overfitting: excels with training data but fails with unknown data.

In this research Random Forest model with hyperparameters is developed in which hyper parameters grid with 'n_estimators': [50, 100, 150], 'max_depth': [None, 10, 20, 30], 'min_samples_split': [2, 5, 10], 'min_samples_leaf': [1, 2, 4] are used. Calculations were made using cross-validation method. Accuracy of this model obtained was 98% Fig. 5 shows the framework of the suggested model.

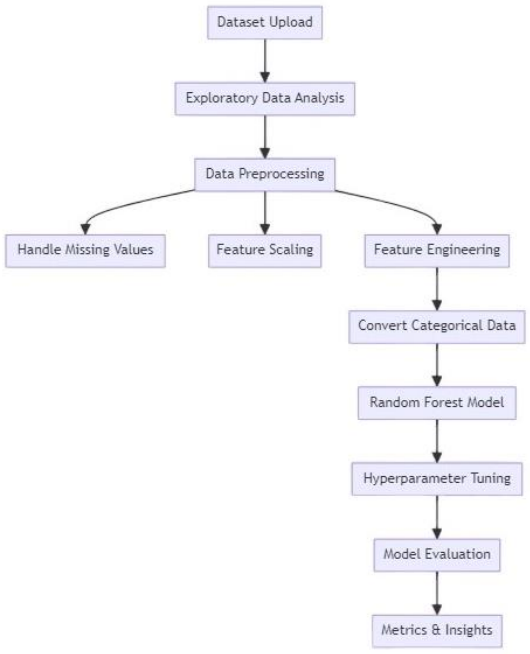

Fig. 5. Proposed Model

It is observed from the graph in Fig. 6 that, The F1 score tends to increase as the max_depth increases up to a certain point. After reaching an optimal depth, the F1 score starts to plateau or even decrease slightly. This indicates that there is an optimal max_depth value that gives the best performance for this dataset, and increasing it further might not lead to significant improvements.

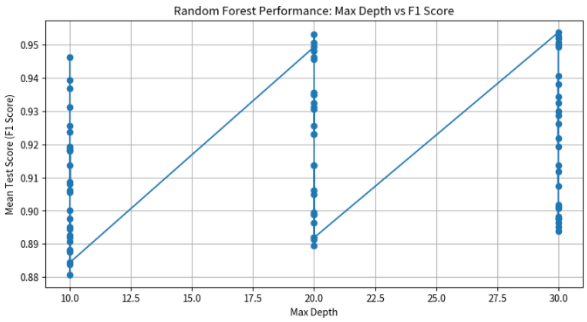

Fig. 6. Relationship between max_depth and the F1 Score

TABLE VI. PERFORMANCE METRICS OF 10 MODELS USING DATASET

Method	Accuracy	Recall	f1-score	precision
Random Forest	98	99	99	98
Decision Tree	93.44	91	95	99
Gradient Boosting	97.27	97	98	98
XGB	92.35	94	94	94
LGBM	97.81	98	98	98
AdaBoost	87.43	92	91	89
SVM	84.15	89	88	87
Logistic Regression	78.69	85	84	83
KNN	80.33	83	85	86
ensemble_model	96.17	99	97	95

Table 6 lists the performance metrics of this study's ten machine learning algorithms. The Random Forest algorithm is the most accurate of them all. A thorough analysis of each model's Accuracy, F1 score, Precision, and Recall can be seen in Table 7. Interestingly, the Random Forest algorithm turns out to be the most accurate and best-performing technique. Table 6 provides more insights that show the Random Forest algorithm's confusion matrix, which shows an amazing 98% accuracy rate. Ninety-eight percent of the time.

TABLE VII. CLASSIFICATION REPORT FOR THE RANDOM FOREST MODEL WITH HYPERPARAMETER TUNING

Method	Accuracy	Recall	f1-score
0	0.98	0.99	0.99
1	0.99	0.97	0.98
Accuracy	-	-	0.98
macro average	0.99	0.98	0.98
Weighted average	0.99	0.98	0.98

TABLE VIII. COMPARISION OF ACCURACIES OF CLASSIFIERS IN THE LITERATURE

Sl.No	Classifier	Accuracy (%)	Reference
1	Logistic Regression	96	[28]
2	Naive Bayes	79.5	[13]
3	Ensembling AB+XB	95	[14]
4	Decision Tree	78	[15]
5	GBM	84.1	[16]
6	SVM and KNN	77	[17]
7	C4.5 Decision Tree	73.5	[18]
8	Decision Tree	85	[19]
9	Deep Neural Network	77.8	[20]
10	Linear Kernel SVM	89	[21]
11	Random Forest	85.5	[5]
12	SVR using NSGA·ll	86.1	[27]
13	Random Forest	80.8	[29]
14	LS·SVM classifier	87.06	[30]
15	Random Forest	90.1	[47]

An overview of all of the research on diabetes prediction that have been published in the literature is provided in Table 8. The differences in the datasets used, the algorithms chosen, and the methodology used in each investigation can be used to explain the discrepancies in prediction accuracy among these studies.

Numerous studies have compared machine learning algorithms to determine their prediction power and effectiveness. It is clear from the analysed research that the majority obtain prediction accuracy of more than 80%. However, elements like dataset size and the quantity of features can have significant effects on how well an algorithm performs. As a result, a method that performs well on one dataset may not do so on another [49]. It is crucial to use machine learning to predict diabetes effectively because of the many complications that may arise from it. Early detection has the ability to save lives and improve management. Future study could be possible as the existing dataset does not allow for the prediction of diabetes type. Valuable next steps could include improving the forecast accuracy and identifying the particular kind of diabetes.

I. CONCLUSION

This study predicted diabetes diagnoses using machine learning. It included 2000 female participants aged 21 to 81, with an average age of 33.1316 (65.8%) were in good health, while 684 (34.2%) had diabetes. Diabetes risk was predicted using 10 machine learning algorithms. The main concern was forecast accuracy. The best prediction algorithm was Random Forest with 98% accuracy. The other algorithms were 81%–89% accurate. Future research could use more patient data to train the model.

REFERENCES

[1] K. G. M. M. Alberti, P. Zimmet, and J. Shaw, "International Diabetes Federation: A consensus on Type 2 diabetes prevention," Diabet. Med., vol. 24, no. 5, pp. 451–463, 2007, doi: 10.1111/j.1464-5491.2007.02157.x.

[2] D. O. F. Diabetes, "Diagnosis and classification of diabetes mellitus," Diabetes Care, vol. 33, no. SUPPL. 1, 2010, doi: 10.2337/dc10-S062.

[3] M. Franciosi et al., "Use of the Diabetes Risk Score for Opportunistic Screening of Undiagnosed Diabetes and Impaired Glucose Tolerance: The IGLOO (Impaired Glucose Tolerance and Long-Term Outcomes Observational) study," Diabetes Care, vol. 28, no. 5, pp. 1187–1194, May 2005, doi: 10.2337/diacare.28.5.1187.

[4] Z. Tao, A. Shi, and J. Zhao, "Epidemiological Perspectives of Diabetes," Cell Biochem. Biophys., vol. 73, no. 1, pp. 181– 185, Sep. 2015, doi: 10.1007/S12013-015-0598-4.

[5] N. Nai-Arun and R. Moungmai, "Comparison of Classifiers for the Risk of Diabetes Prediction," Procedia Comput. Sci., vol. 17 O. F. Akmese / Hittite J Sci Eng, 2022, 9(1) 09–18 69, pp. 132–142, 2015, doi: 10.1016/j.procs.2015.10.014.

[6] P. Hossain, B. Kawar, and M. El Nahas, "Obesity and Diabetes in the Developing World — A Growing Challenge," N. Engl. J. Med., vol. 356, no. 3, pp. 213–215, 2007, doi: 10.1056/ nejmp068177.

[7] F. Mercaldo, V. Nardone, and A. Santone, "Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques," Procedia Comput. Sci., vol. 112, pp. 2519–2528, 2017, doi: 10.1016/j. procs.2017.08.193.

[8] J. Tuomilehto et al., "Prevention of Type 2 Diabetes Mellitus by Changes in Lifestyle among Subjects with Impaired Glucose Tolerance," New England Journal of Medicine, vol. 344, no. 18. pp. 1343–1350, 2001, doi: 10.1056/ nejm200105033441801.

[9] J. L. Chiasson, R. G. Josse, R. Gomis, M. Hanefeld, A. Karasik, and M. Laakso, "Acarbose for prevention of type 2 diabetes mellitus: the STOP-NIDDM randomized trial," Lancet, vol. 359, no. 9323, pp. 2072–2077, Jun. 2002, doi: 10.1016/ S0140-6736(02)08905-5.

[10] A. Ramachandran, C. Snehalatha, S. Mary, B. Mukesh, A. D. Bhaskar, and V. Vijay, "The Indian Diabetes Prevention Programme shows that lifestyle modification and metformin prevent type 2 diabetes in Asian Indian subjects with impaired glucose tolerance (IDPP-1)," Diabetologia, vol. 49, no. 2, pp. 289–297, 2006, doi: 10.1007/s00125-005- 0097-z.

[11] T. Diyabet, V. Başkanı, and P. M. Temel, "DİYABET ORANI 10 YILDA YÜZDE 100 ARTTI," pp. 10–12, 2017.

[12] Ö. F. AKMEŞE, "Karın Ağrısı ile Acil Servise Başvuran Hastalarda Akut Apandisit Tanısı için Makine Öğrenmesi Yaklaşımlarının Kullanımı," Kırıkkale University, 2020.

[13] A. Iyer, J. S, and R. Sumbaly, "Diagnosis of Diabetes Using Classification Mining Techniques," Int. J. Data Min. Knowl. Manag. Process, vol. 5, no. 1, pp. 01–14, 2015, doi: 10.5121/ijdkp.2015.5101.

[14] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," IEEE Access, vol. 8, pp. 76516–76531, 2020, doi: 10.1109/ACCESS.2020.2989857.

[15] X. H. Meng, Y. X. Huang, D. P. Rao, Q. Zhang, and Q. Liu, "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors," Kaohsiung J. Med. Sci., vol. 29, no. 2, pp. 93–99, 2013, doi: 10.1016/j. kjms.2012.08.016.

[16] H. Lai, H. Huang, K. Keshavjee, A. Guergachi, and X. Gao, "Predictive models for diabetes mellitus using machine learning techniques," BMC Endocr. Disord., vol. 19, no. 1, pp. 1–9, 2019, doi: 10.1186/s12902-019-0436-6.

[17] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, "Prediction of diabetes using machine learning algorithms in healthcare," ICAC 2018 - 2018 24th IEEE Int. Conf. Autom. Comput. Improv. Product. through Autom. Comput., no. September, pp. 6–7, 2018, doi: 10.23919/ IConAC.2018.8748992.

[18] M. F. Faruque, Asaduzzaman, and I. H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus," 2nd Int. Conf. Electr. Comput. Commun. Eng. ECCE 2019, pp. 7–9, 2019, doi: 10.1109/ECACE.2019.8679365.

[19] P. Sonar and K. Jaya Malini, "Diabetes prediction using different machine learning approaches," Proc. 3rd Int. Conf. Comput. Methodol. Commun. ICCMC 2019, no. Iccmc, pp. 367–371, 2019, doi: 10.1109/ICCMC.2019.8819841

[20] S. Wei, X. Zhao, and C. Miao, "A comprehensive exploration to the machine learning techniques for diabetes identification," IEEE World Forum Internet Things, WF-IoT 2018 - Proc., vol. 2018-Janua, pp. 291–295, 2018, doi: 10.1109/WFIoT.2018.8355130.

[21] H. Kaur and V. Kumari, "Predictive modelling and analytics for diabetes using a machine learning approach," Appl. Comput. Informatics, 2019, doi: 10.1016/j.aci.2018.12.004.

[22] L. Parthiban and R. Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm," Int. J. Biol. Med. Sci., vol. 3, no. 3, pp. 157–160, 2008.

[23] K. Sowjanya, A. Singhal, and C. Choudhary, "MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices," Souvenir 2015 IEEE Int. Adv. Comput. Conf. IACC 2015, pp. 397–402, 2015, doi: 10.1109/ IADCC.2015.7154738.

[24] K. M. Orabi, Y. M. Kamal, and T. M. Rabah, "Early predictive system for diabetes mellitus disease," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016, vol. 9728, pp. 420–427, doi: 10.1007/978-3-319- 41561-1_31.

[25] AIZERMAN and M. A., "Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning," Autom. Remote Control, vol. 25, pp. 821–837, 1964, Accessed: Nov. 27, 2021. [Online]. Available: https://ci.nii. ac.jp/naid/10021200712.

[26] Boser Berhard E., G. I. M., and V. N. Vapnik, "A training algorithm for optimal margin classifiers," In Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, 1992, pp. 144–152.

[27] M. H. Zangooei, J. Habibi, and R. Alizadehsani, "Disease Diagnosis with a hybrid method SVR using NSGA-II," Neurocomputing, vol. 136, pp. 14–29, 2014, doi: 10.1016/j. neucom.2014.01.042.

[28] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," Procedia Comput. Sci., vol. 165, pp. 292–299, 2019, doi: 10.1016/j.procs.2020.01.047

[29] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting Diabetes Mellitus with Machine Learning Techniques," Front. Genet., vol. 9, no. November, pp. 1–10, 2018, doi: 10.3389/ fgene.2018.00515.

[30] V. ACAR, E, ÖZERDEM, M. AKPOLAT, "Forecasting Diabetes Mellitus with Biometric Measurements.," Int. Arch. Med. Res., vol. 1, no. 1, pp. 28–42, 2011.

[31] J. Tukey, "Exploratory data analysis," 1977, Accessed: Sep. 08, 2021. [Online]. Available: http://theta.edu.pl/ wp-content/uploads/2012/10/exploratorydataanalysis_tukey.pdf.

[32] R. S. Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, "Pima Indians Diabetes Database," 2016. https://www.kaggle.com/uciml/pima-indians-diabetes-database (accessed Aug. 01, 2021).

[33] Tufte: The visual display of quantitative information - Google Akademik." https://scholar.google.com/ scholar_lookup?title=The Visual Display of Quantitative Information&publication_year=2001&author=E. Tufte (accessed Sep. 08, 2021).

[34] S. Lavalle, E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz, "Big Data, Analytics and the Path from Insights to Value Big Data, Analytics and the Path from Insights to Value," no. 52205, 2011.

[35] R. Agrawal, A. Kadadi, X. Dai, and F. Andres, "Challenges and opportunities with big data visualization," 7th Int. ACM Conf. Manag. Comput. Collect. Intell. Digit. Ecosyst. MEDES 2015, pp. 169–173, Oct. 2015, doi: 10.1145/2857218.2857256.

[36] S. Nestorov, B. Jukić, N. Jukić, A. Sharma, and S. Rossi, "Generating insights through data preparation, visualization, and analysis: Framework for combining clustering and data visualization techniques for low-cardinality sequential data," Decis. Support Syst., vol. 125, no. March, p. 113119, 2019, doi: 10.1016/j.dss.2019.113119.

[37] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[38] K. S. Albayrak A., "VERİ MADENCİLİĞİ: KARAR AĞACI ALGORİTMALARI VE İMKB VERiLERİ ÜZERİNE BİR UYGULAMA * DATA MINING: DECISION TREE ALGORITHMS AND AN APPLICATION ON ISE DATA," no. May, 2014.

[39] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. 13-17-August-2016, pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785.

[40] J. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine Author (s): Jerome H. Friedman Source: The Annals of Statistics, Vol. 29, No. 5 (Oct., 2001), pp. 1189-1232 Published by: Institute of Mathematical Statistics Stable URL: http://www," Ann. Stat., vol. 29, no. 5, pp. 1189–1232, 2001.

[41] W. Zhao, J. Li, J. Zhao, D. Zhao, J. Lu, and X. Wang, "XGB model: Research on evaporation duct height prediction based on XGBoost algorithm," Radio engineering, vol. 29, no. 1, pp. 81–93, 2020, doi: 10.13164/re.2020.0081.

[42] W. Cai, R. Wei, L. Xu, and X. Ding, "A method for modelling greenhouse temperature using gradient boost decision tree," Inf. Process. Agric., Sep. 2021, doi: 10.1016/J. INPA.2021.08.004.

[43] M. Massaoudi, S. S. Refaat, I. Chihi, M. Trabelsi, F. S. Oueslati, and H. Abu-Rub, "A novel stacked generalization ensemble based hybrid LGBM-XGB-MLP model for Short-Term Load Forecasting," Energy, vol. 214, p. 118874, Jan. 2021, doi: 10.1016/J.ENERGY.2020.118874.

[44] R. E. Schapire, "Explaining AdaBoost," Empir. Inference Festschrift Honor Vladimir N. Vapnik, pp. 37–52, Jan. 2013, doi: 10.1007/978-3-642-41136-6_5.

[45] V. Vapnik, The Nature of Statistical Learning Theory. Springer science & business media, 2013.

[46] E. Ürük, "İstatistiksel Uygulamalarda Lojistik Regresyon Analizi," Marmara University, 2007.

[47] Omer Faruk Akmese, "Diagnosing Diabetes with Machine Learning Techniques." Hittite Journal of Science and Engineering, 2022, 9 (1) 09–18.

[48] Gorunescu Florin, Data Mining: Concepts, Models and Techniques. Berlin: Springer Science & Business Media, 2011.

[49] Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," Comput. Struct. Biotechnol. J., vol. 15, pp. 104–116, 2017, doi: 10.1016/j. csbj.2016.12.005

.