Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 2 (2024)

An Extensive Examination of Machine Learning Methods for Identifying Diabetes

D. P. Singh
Amity University Uttar Pradesh Greater Noida Campus

Abstract: This paper presents a comprehensive examination of machine learning (ML) methods employed in the identification and management of diabetes. With the rising prevalence of diabetes globally, there is a pressing need for accurate and efficient diagnostic tools. ML techniques offer promising avenues for enhancing diagnostic accuracy, risk prediction, and personalized treatment approaches. This study reviews a range of ML algorithms applied to various datasets for diabetes detection, including supervised, unsupervised, and hybrid approaches. The examination of each algorithm's performance aims to identify the one displaying superior accuracy, precision, sensitivity, and specificity. Additionally, the decision-making process is evaluated to enhance the model. Through rigorous evaluation and comparison, insights are drawn regarding the effectiveness and applicability of each model in the context of diabetes prediction. The findings contribute to advancing the understanding of machine learning methodologies in healthcare and offer valuable guidance for developing robust predictive models for diabetes diagnosis and management. The study aims to contribute valuable insights into the application of machine learning techniques for diabetes prediction.

Keywords: *Machine learning, Diabetes Identification, Diagnostic Tools, Hybrid methods, Data quality, Healthcare, Feature selection and Predictive Models*

1. Introduction

Diabetes has emerged as a significant global health concern, with its prevalence steadily increasing across populations worldwide. The accurate identification and effective management of diabetes are crucial for mitigating its associated complications and improving patient outcomes. In recent years, machine learning (ML) techniques have garnered attention for their potential to revolutionize diabetes identification and management. ML algorithms offer a data-driven approach to analyzing complex datasets, thereby enhancing diagnostic accuracy, risk prediction, and treatment personalization.

Healthcare systems are organized to provide people with the necessary components to uphold good health and enhance the efficiency of identifying and diagnosing diseases and conditions, following conventional methods. Generally, patients often express significant concern about the quality of healthcare systems and the available facilities for delivering treatment. The positive outcomes of enhancements in healthcare systems tend to directly benefit individuals with existing health issues, comprising a substantial portion of the population affected by various ailments like diabetes, blood sugar problems, and hypertension. According to the National Diabetes Statistics Report for 2020, around one out of every ten individuals in the United States has been diagnosed with diabetes. Moreover, there is a significant increase in the occurrence of both type 1 and type 2 diabetes among young people[8].

Given that health and healthcare are fundamental pillars of a thriving society, leveraging the capabilities of computational methods and artificial intelligence [32] becomes essential for creating novel approaches applicable in healthcare systems. This is essential to promote a healthier community and reducing the chances of such illnesses among both present and upcoming generations, thus improving the general standard of living. The medical field has experienced a significant transformation with the progression of technology.

Detecting diabetes early allows for effective management. Adopting a balanced diet and regular exercise routine can aid in preventing diabetes[30]. For individuals with prediabetes, engaging in physical activity to shed excess weight can decrease the likelihood of developing Type 2 diabetes. The National Diabetes Prevention Program,

led by the Center for Disease Control and Prevention (CDC), offers a lifestyle modification program to assist those with prediabetes in altering their lifestyle and averting the onset of Type 2 diabetes[37].

The healthcare sector gathers vast quantities of data encompassing hospital records, patient medical histories, and medical test outcomes. Early disease detection traditionally relies on a doctor's expertise, yet this method can be prone to inaccuracies and biases. Consequently, manual decision-making poses risks. Hidden data patterns may go unnoticed, affecting treatment decisions and potentially depriving patients of suitable care. Improved accuracy through automated identification is crucial for early diabetes detection[4,17].

Technology plays a crucial role in overcoming barriers of distance and resources, benefiting those who have access to it. Innovations like magnetic resonance imaging machines in video technology and internet-based applications offer customizable services to patients, with telehealth becoming increasingly prominent for remote care after initial clinical visits. This advancement enables clinicians to address patient needs more effectively through online communication [32].For instance, video technology facilitates emergency care for trauma patients in both rural and urban areas lacking immediate clinical access [7]. Moreover, technology enhances home healthcare, improving productivity and safety [7].Hospitals face challenges related to data accuracy and availability, particularly in managing and analysing patient data. Machine learning and deep learning algorithms play a crucial role in addressing these challenges, offering state-of-the-art solutions like matching algorithms and natural language processing [3].Data mining emerges as a valuable approach to extract data directly, bypassing the need for expert knowledge. These methods yield unique patterns to develop personalized strategies for individual hospitals [6].

Over the past few years, data mining and machine learning have become essential tools in the medical field. Data mining is employed to pre-process and extract pertinent features from healthcare data, while machine learning techniques facilitate automated prediction of diabetes. [16,20]. These algorithms enable the identification of hidden patterns in data using advanced methods, leading to more reliable decision-making with greater accuracy. Nvidia describes machine learning as employing diverse algorithms to learn from processed data and make predictions[11].Data mining involves the application of various techniques, including machine learning, statistics, and database systems, to uncover patterns within large datasets[12].

This paper provides a comprehensive examination of ML methods utilized in the identification and management of diabetes. By reviewing a variety of ML algorithms applied to diverse datasets, including supervised, unsupervised, and hybrid approaches, this study aims to elucidate the strengths and limitations of these techniques in diabetes detection. Through synthesizing current research findings, this paper contributes to the ongoing discourse on leveraging advanced technology, particularly ML, to enhance diabetes care. By providing insights into the application of ML methods in diabetes identification, this study aims to inform clinicians, researchers, and healthcare policymakers about the potential of ML-driven approaches in improving diabetes management and patient outcomes

2. Literature review

In recent times, a substantial body of research has emerged focusing on the identification of diabetic patients through symptom analysis utilizing machine learning methodologies. In their study [30], researchers propose a model designed to ascertain whether a patient is diabetic or not. This model hinges on the predictive capabilities of robust machine learning algorithms, employing metrics like precision, recall, and F1-measure. Using the Pima Indian Diabetes (PIDD) dataset for diagnostic inference, the authors achieved predictive accuracies of 94%, 79%, and 69% respectively with Logistic Regression (LR), Naïve Bayes (NB), and K-nearest Neighbor (KNN) algorithms. Similarly, in [31], another study employed seven machine learning algorithms on the same dataset for diabetes prediction. They determined that Logistic Regression and Support Vector Machine (SVM) models outperformed others in this task. Furthermore, they constructed a neural network (NN) model with varying hidden layers, discovering that the NN with two hidden layers achieved an accuracy of 88.6%.

The study outlined in citation [35] utilizes a range of machine learning classification techniques, including Gaussian Naive Bayes, K-Nearest Neighbors, Artificial Neural Network, Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine, on the PIID dataset. Logistic Regression emerged with the highest

accuracy compared to the other algorithms. Another investigation by Sarwar et al. [26] delves into predictive analytics in healthcare, employing various machine-learning algorithms. A dataset containing medical records of patients is gathered for experimental purposes, and the performance and accuracy of these algorithms are analyzed and contrasted. In the paper referenced as [24], the authors present a novel diabetes prediction model that incorporates external factors alongside traditional ones like Glucose, BMI, Age, and Insulin. They demonstrate improved classification accuracy compared to existing models using this new dataset.

M. Rady et.al[33] assessed the effectiveness of eight different machine learning algorithms. These algorithms encompassed logistic regression, both linear and nonlinear kernel support vector machines, random forest, decision tree, adaptive boosting classifier, K-nearest neighbor, and naïve Bayes.The findings reveal that the Random Forest classifier outperformed the others, achieving an accuracy.

M. U. Emon [34], utilized various machine learning methods, including Logistic Regression, Gaussian Process, AdaBoost, Decision Tree, K-Nearest Neighbors, Multilayer Perceptron, Support Vector Machine, Bernoulli Naive Bayes, Bagging Classifier, Random Forest, and Quadratic Discriminant Analysis. Among these, the Random Forest classifier outperformed the others, achieving a 98% accuracy rate. This accuracy was notably higher than that of the other algorithms tested.

In their study on early-stage diabetes prediction outlined in [21], the authors introduce an innovative approach utilizing key attributes. They utilize various tools to identify significant attributes for both clustering and prediction purposes.

The results indicate a significant relationship between diabetes and factors such as body mass index (BMI) and glucose levels. Different methodologies, such as artificial neural networks (ANN), random forest, and K-means clustering, are utilized to predict diabetes, with ANN showing better predictive accuracy compared to alternative approaches [22]. The authors suggest an innovative method employing machine learning algorithms implemented on Hadoop-based clusters to forecast diabetes. They apply this technique to the Pima Indians Diabetes Database and Digestive Diseases dataset, achieving remarkably precise predictive outcomes. Another investigation [27] employs four machine learning algorithms—Random Forest, K-nearest neighbor, Support Vector Machine, and Linear Discriminant Analysis—to examine the initial phases of diabetes prediction through experimentation. Numerous researchers employed machine learning techniques to forecast diabetes utilizing the Pima Indian diabetes dataset (PIDD), which comprises 9 attributes and 768 records detailing the patients.

Alam, T.M. and colleagues demonstrated [23] a 75.7% accuracy rate by employing Artificial Neural Network (ANN) techniques on Primary Immunodeficiency Diseases (PIDD). Sajida Perveen and co-researchers utilized a dataset sourced from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) in their study. This dataset encompassed variables such as systolic blood pressure (sBP), diastolic blood pressure (dBP), HDL, triglycerides (TG), BMI, fasting blood sugar (FBS), and gender. Their methodology involved the application of Bootstrap aggregating, Adaptive Boosting, and decision tree models.

For increased precision, Adaboost can be utilized to forecast illnesses such as diabetes, coronary heart disease, and hypertension. In their study, Sisodia et al. [17] determined that the NB classifier demonstrated superior accuracy of 76.30% compared to SVM and DT when applied to PIDD. Additionally, Tigga et al. [29] employed logistic regression for predicting diabetes in PIDD.

In a study by Amour Diwani et al. [25], patient data underwent training and testing through 10 cross-validations using Naive Bayes and decision tree algorithms. Subsequently, performance was assessed, scrutinized, and compared with other classification algorithms using WEKA.

In the research by Zou et al. [18], various classification algorithms including Random Forest, Decision Tree, and ANN were applied to classify PIDD after employing feature reduction techniques such as Principal Component Analysis (PCA) and Minimum Redundancy Maximum Relevance (mRMR). Their study revealed that the highest accuracy of 77.21% for Pima Indians was achieved using random forest with the mRMR feature reduction method. Selecting relevant features and choosing an appropriate classifier are identified as significant challenges in the realm of machine learning.

In our study, we employed Pearson's correlation technique to identify relevant characteristics. Our focus was on predicting whether a patient has diabetes. To accomplish this, we applied various machine learning classification

algorithms such as Naive Bayes (NB), Support Vector Machine (SVM), Linear Regression (LR), Adaboost, Random Forest, K Nearest Neighbour (KNN), and Neural Network (NN) with varying hidden layer configurations. We assessed the performance of these methods using different evaluation metrics.

3. Machine Learning algorithms for classification and prediction of Diabetes:

Choosing the appropriate machine learning algorithms for classification and prediction of diabetes based on available datasets.The tasks are contingent on various factors such as the characteristics of the data, dataset scale, available computational resources, interpretability needs, and the project's particular objectives. To achieve the optimal results from the model, various machine learning algorithms were applied and evaluated[18,23,29]. We continuously sought to improve performance during the model development phase, leading us to conduct Hyper parameter tuning to identify the most effective parameters. Here are some commonly used algorithms for classification and prediction tasks[37,38]:

3.1 Logistic Regression: Logistic regression, a type of supervised machine learning, is employed to classify data by estimating the likelihood of an observation belonging to a specific category. Drawing from statistical principles, this technique examines the correlation between pairs of variables in the data.

Logistic regression is utilized for forecasting the result of a categorical dependent variable, meaning the result is categorical or discrete. This concept could be framed as a choice between Yes or No, 0 or 1, true or False, and so forth. However, rather than supplying absolute values such as 0 and 1, logistic regression presents probabilities that fall within the range of 0 to 1. Instead of fitting a linear regression line, logistic regression employs a sigmoid-shaped logistic function, which predicts outcomes as either 0 or 1. The logistic regression model employs the sigmoid function to make predictions for both positive and negative classes.

The formula of the sigmoid activation function is: $(z) = \frac{1}{1-e^{-z}}$, and Logistic regression equation will be: $P(X, b, w) = \frac{1}{1+e^{-wX+b}}$

Consequently, when z tends towards positive infinity, its projected value approaches 1, and when it approaches negative infinity, it tends towards 0. Furthermore, we categorize a label as part of class 1 (the positive class) if the output of the sigmoid function exceeds 0.5, and as part of class 0 (the negative class) if it is below 0.5.

3.2 Decision Trees: A decision tree stands out as a highly effective instrument among supervised learning algorithms, serving in both classification and regression endeavors. Its creation yields a flowchart resembling a tree, where internal nodes represent attribute tests, branches represent outcomes, and leaf nodes contain class labels. This tree evolves by repeatedly dividing training data into subsets based on attribute values, stopping based on predetermined conditions such as maximum tree depth or minimum samples required for further division at a node.

It performs a classification process utilizing input characteristics, wherein every leaf node represents a result, while inner nodes retain dataset properties and decision rules. The feature demonstrating the highest information gain is selected for outcome prediction, with information gain assessed at each node across all attributes. Hyperparameter tuning was conducted through Grid Search CV and Randomized Search CV, setting minimum leaf samples to 10, maximum depth to 6, and employing the 'gini' criterion.

3.3 Random Forest: Random Forest, a widely-used ensemble learning method rooted in tree structures, forms a 'forest' consisting of decision trees. These trees are typically trained using the 'bagging' method, which combines multiple models to enhance the final outcome. Random Forest enhances the performance of Decision Trees by minimizing variance. It achieves this by increasing model randomness through the growth of tree numbers. Instead of solely emphasizing the most influential feature during node division, the algorithm selects the best feature from a random subset of features. This approach contributes to a more effective model.

In the Random Forest algorithm, several decision trees act as the fundamental learning components. The procedure entails randomly choosing rows and features from the dataset to form sample datasets for each model, a process referred to as Bootstrap sampling. Instead of relying on single decision trees, this method combines numerous trees to produce the ultimate result.

3.4 Support Vector Machines (SVM): The Support Vector Machine (SVM) is a powerful method in machine learning applied to tasks encompassing linear or nonlinear classification, regression, and identifying outliers. Its versatility extends to a wide array of applications including text and image categorization, spam identification, handwriting and facial recognition, gene expression analysis, and detecting anomalies. SVMs excel in managing high-dimensional data and nonlinear correlations, making them versatile and efficient across a range of tasks. Their effectiveness lies in their ability to determine the maximum separating hyperplane between different classes within the target feature.

To find the optimal hyperplane, one needs to calculate the margin, which is the distance between the closest points of each class and the decision boundary. SVM aims to select the hyperplane with the greatest margin. However, there are instances where SVM prioritizes accurate class prediction over maximizing the margin to determine the best hyperplane. In SVM implementation, careful selection of hyperparameters is crucial as they significantly impact accuracy. To optimize these hyperparameters, Randomized Search CV was employed. Among various kernels, the Radial Basis Function Kernel was chosen. The equation for this kernel function was utilized in the analysis.

$$K(t,t') = e^{-\frac{\|t-t'\|^2}{2\sigma^2}}$$

In this context, the term $\|t - t'\|^2$ represents the squared Euclidean distance between two vectors. The SVM with an RBF kernel was configured with specific parameter values: a gamma value of 1 and a penalty parameter of 1000 for the error term. Fine-tuning these parameters led to enhanced accuracy.

3.5 K-Nearest Neighbors (KNN): KNN stands as a fundamental but crucial classification technique within machine learning, operating within the realm of supervised learning. It holds significant relevance in various fields including pattern recognition, data mining, and intrusion detection.

The K-NN algorithm, renowned for its simplicity and adaptability, is widely utilized in machine learning. Unlike many other methods, it doesn't rely on presumptions about data distribution, making it applicable across diverse datasets. Its versatility extends to handling both numerical and categorical data, rendering it a pragmatic choice for classification and regression tasks. As a non-parametric technique, K-NN predicts outcomes by assessing the similarity between data points, demonstrating robustness against outliers. It functions by locating the K closest neighbors to a specified data point using a selected distance metric such as Euclidean distance. Afterwards, the data point's class or value is established through either majority voting or averaging among its nearest neighbors. By leveraging local data structures, this approach enables the algorithm to adapt to varying patterns effectively.

3.6 Naive Bayes: Naïve Bayes methods are classification algorithms which utilize Bayes' Theorem, operating under the assumption of predictor independence. Bayes' Theorem computes the probability of an event occurring based on the probability of another event that has already taken place.

The mathematical expression of Bayes' Theorem is as follows: $P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right)P(A)}{P(B)}$, A & B represent events, with P(A) & P(B) denoting the separate probabilities of A & B occurring independently.

P(A|B) represents the probability of event A occurring if event B is known to have occurred, whereas P(B|A) represents the probability of event B occurring if event A is known to have occurred.

3.7 Gaussian Naive Bayes

Gaussian Naive Bayes, a version of the Naive Bayes method, handles continuous attributes by assuming that the features of the data follow a Gaussian distribution throughout the dataset.

In the context of the Sklearn library, Gaussian Naive Bayes is classified as an algorithm for categorization, specifically designed for continuous features following a normal distribution, and it operates on the principles of the Naive Bayes algorithm. Before delving further into this subject, it's essential to grasp the fundamental workings of Gaussian Naive Bayes.

Gaussian Naive Bayes involves implementing the Naive Bayes algorithm on data that follows a normal distribution. It assumes that the probability of observing each feature given a class follows a Gaussian distribution. This is expressed as $P\left(\frac{x_i}{y}\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)}$, where σ represents the standard deviation and μ represents the mean.

In classification, the algorithm calculates the posterior probability for each class given a new data point x, and assigns the data point to the class with the maximum posterior probability.

3.8 Gradient Boosting: Gradient boosting is a method that builds a final model by integrating several weak learning algorithms trained on the same dataset. It operates by gradually introducing weak learners in stages. Initially, the first weak learner in the gradient boosting algorithm doesn't directly train on the dataset; rather, it predicts the mean of a pertinent column. Subsequently, the residuals stemming from this prediction are computed and act as the target column for the next weak learner. This iterative process continues with subsequent weak learners until the residuals are minimized. For gradient boosting to work, the dataset needs to comprise numerical or categorical data, and the loss function employed to compute residuals must be continuously differentiable.

3.8.1 XGBoost: XGBoost, which stands for Extreme Gradient Boosting, is an advanced machine learning approach that extends the principles of gradient boosting.Unlike traditional gradient boosting, XGBoost incorporates regularization, making it a more advanced variant. This regularization aspect contributes to its superior performance and speed compared to standard gradient boosting. Moreover, XGBoost is particularly effective when dealing with datasets containing a mix of numerical and categorical variables. The prediction procedure utilized the XGBoost regression model, which underwent refinement by tweaking its hyperparameters through a combination of grid search and threefold cross-validation. The hyperparameter grid encompassed variables such as 'n_estimators', 'max_depth', 'learning rate', 'colsample_bytree', and 'gamma', with the aim of minimizing squared error. Following the identification of optimal hyperparameters, the model underwent training on the complete training dataset and subsequent evaluation on the test set.

Evaluation of performance involved metrics such as R-squared value, root mean squared error (RMSE), and magnitude relative error (MRE):

$$RMSE = \sqrt{\frac{1}{n}\sum_i^n (y_i - \check{y}_i)^2} \quad \text{and } MRE = \frac{|y_i - \check{y}_i|}{|y_i|}$$

n represents the total count of observations within the dataset. Each observation, denoted as y_i, corresponds to the actual value for the i-th entry, while $\check{y}_i$ stands for the predicted value for the same i-th observation.

3.8.2 AdaBoost: AdaBoost is a boosting technique that operates by iteratively adding weak learners to form a strong learner. The alpha parameter's value is inversely correlated with the weak learner's error, differing from XGBoost's Gradient Boosting, where the alpha parameter is computed based on weak learner errors.

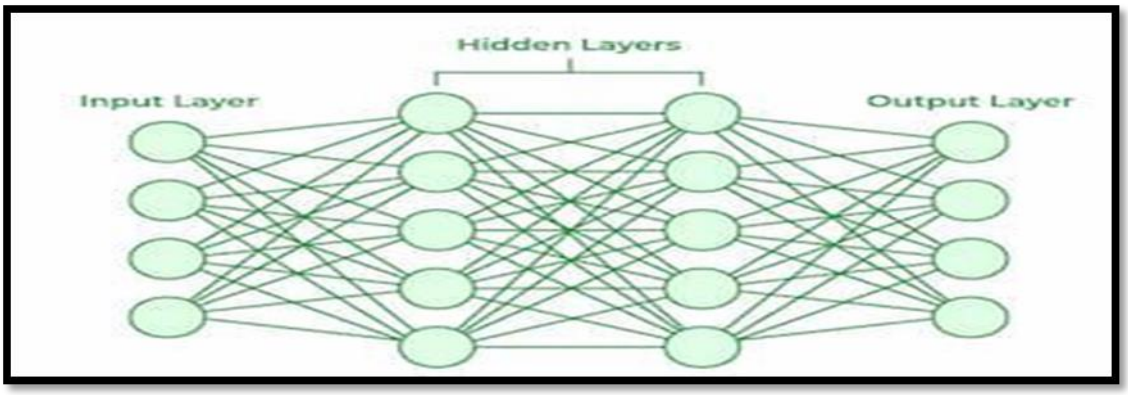
3.8.3 CatBoost: CatBoost stands out from its competitors primarily due to its distinctive feature of fostering symmetric growth of decision trees. This characteristic enhances its performance compared to other algorithms. CatBoost particularly excels with categorical datasets, thanks to its unique approach in handling them. Categorical features are encoded based on the output columns, thus leveraging the output column's weight during training or encoding. This methodology significantly boosts accuracy when dealing with categorical data.

3.9 Neural Networks: Neural networks discern distinctive characteristics from data without relying on pre-set comprehension. Constituent elements of these networks comprise neurons, connections, weights, biases, propagation functions, and a learning mechanism. Neurons process inputs, guided by thresholds and activation functions, while connections manage information transmission through weights and biases. Learning unfolds across three phases: input processing, output creation, and iterative enhancement, which bolsters the network's effectiveness across a range of tasks.

The initial stage receives various features such as the email's content, sender details, and subject. These features undergo multiplication with adjusted weights and are then processed through concealed layers. Over time, through training, the network learns to identify specific patterns that signify whether an email qualifies as spam or not. The output layer, utilizing a binary activation function, provides a prediction regarding the email's spam status, either indicating spam (1) or not spam (0). As the network continually adjusts its weights via backpropagation, it

improves its ability to differentiate between spam and legitimate emails, demonstrating the practical utility of neural networks in applications like email filtration.

Neural networks are complex systems crafted to mimic specific functions of the human brain. They comprise various layers, including an input layer, one or more hidden layers, and an output layer composed of interconnected artificial neurons. The primary processes involved are termed forward propagation and backpropagation.



When applying these algorithms to diabetes prediction, it's essential to preprocess the data, handle missing values, scale features, and possibly perform feature selection or dimensionality reduction techniques to improve model performance. Additionally, evaluating models using metrics like accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC) is crucial to assess their performance accurately.

4. Confusion Matrix in Machine Learning:

A confusion matrix serves as a summary of how well a machine learning model performs on a given test dataset. It provides a visual representation of the accurate and inaccurate predictions made by the model. Typically utilized in assessing the effectiveness of classification models, which predict categorical labels for input instances. When evaluating the performance of a classification model, the use of a confusion matrix is crucial. This matrix offers a comprehensive breakdown of correct positives, correct negatives, incorrect positives, and incorrect negatives. It allows for a more profound comprehension of the model's recall, accuracy, precision, and overall capability to differentiate between classes. The table displays the frequency of occurrences predicted by the model on the test dataset.

4.1 Accuracy: Accuracy functions as a measurement for assessing model effectiveness, determined by the ratio of accurately classified instances to the overall number of instances.

$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$, where TP= True positives, TN= True negatives, FP= False positives and FN= False negatives.

4.2 Precision: Precision denotes the correctness of a model's positive forecasts and is gauged by the proportion of true positive forecasts relative to the total number of positive forecasts generated by the model.

$$Precision = \frac{TP}{TP + FP}$$

4.3 Recall: Recall assesses the effectiveness of a classification model in identifying all relevant instances in a dataset. It is determined by dividing the number of true positive (TP) instances by the sum of true positive and false negative (FN) instances.

$$Recall = \frac{TP}{TP + FN}$$

4.4 Specificity: Specificity, a crucial metric in assessing classification models, especially in binary scenarios, gauges a model's accuracy in correctly recognizing negative instances, also referred to as the True Negative Rate.

$$Specificity = \frac{TN}{TP + FP}$$

4.5 Type 1 and Type 2 error: Type 1 error arises when the model incorrectly predicts a positive instance as negative. Precision is influenced by false positives since it reflects the ratio of true positives to the sum of true positives and false positives.

$$Type\ 1\ Error = \frac{FP}{TN + FP}$$

Type 2 error happens when the model doesn't forecast a positive case. Recall is impacted by false negatives as it represents the proportion of true positives relative to the sum of true positives and false negatives.

$$Type\ 2\ Error = \frac{FN}{TP + FN}$$

4.6 F1-Score: The F1-score assesses a classification model's overall performance by calculating the harmonic mean of its precision and recall.

$$F1Score\ = 2 \times \frac{Precision\ \times\ Recall}{Precision\ +\ Recall}$$

A tool used to assess classification tasks across all levels is the AUC–ROC curve. AUC indicates the efficacy of differentiation, while ROC represents a probability curve. Higher AUC values indicate a model's improved ability to distinguish between diabetic and non-diabetic cases. The ROC curve illustrates the relationship between the True Positive Rate and the False Positive [10, 38].

5. Information and Approach:

5.1 Exploring and preparing the dataset: Exploring and preparing the dataset is a critical phase in the data analysis and modeling pipeline, laying the foundation for accurate predictions and meaningful insights. By gaining insights into the data's structure, quality, and characteristics, researchers and practitioners can make informed decisions and derive actionable insights to address real-world problems effectively. Effective dataset exploration and preparation significantly contribute to the success of data-driven projects and enable the development of robust and reliable machine learning models.

In this study, we aim to analyze the Pima Indian Dataset using advanced algorithms for effective utilization in the Internet of Medical Things (IoMT). The dataset was obtained from Kaggle (https://www.kaggle.com/datasets/mathchi/diabetes-data-set database). It has been appropriately anonymized, ensuring the absence of any identifiable patient information. Table 1 presents a summary of the dataset features and their descriptions.

Table:1

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1

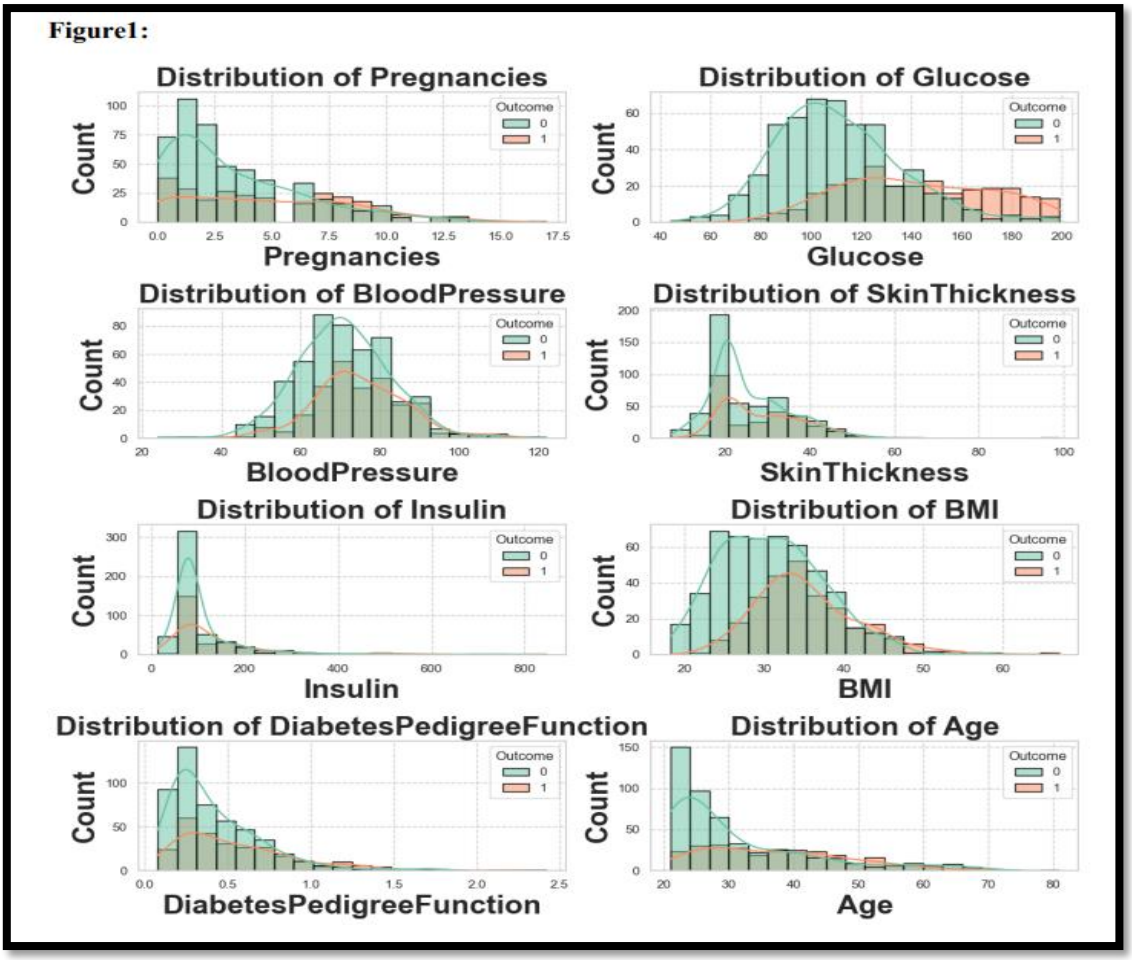
The dataset does not contain any null or missing values, as shown in Table 2.

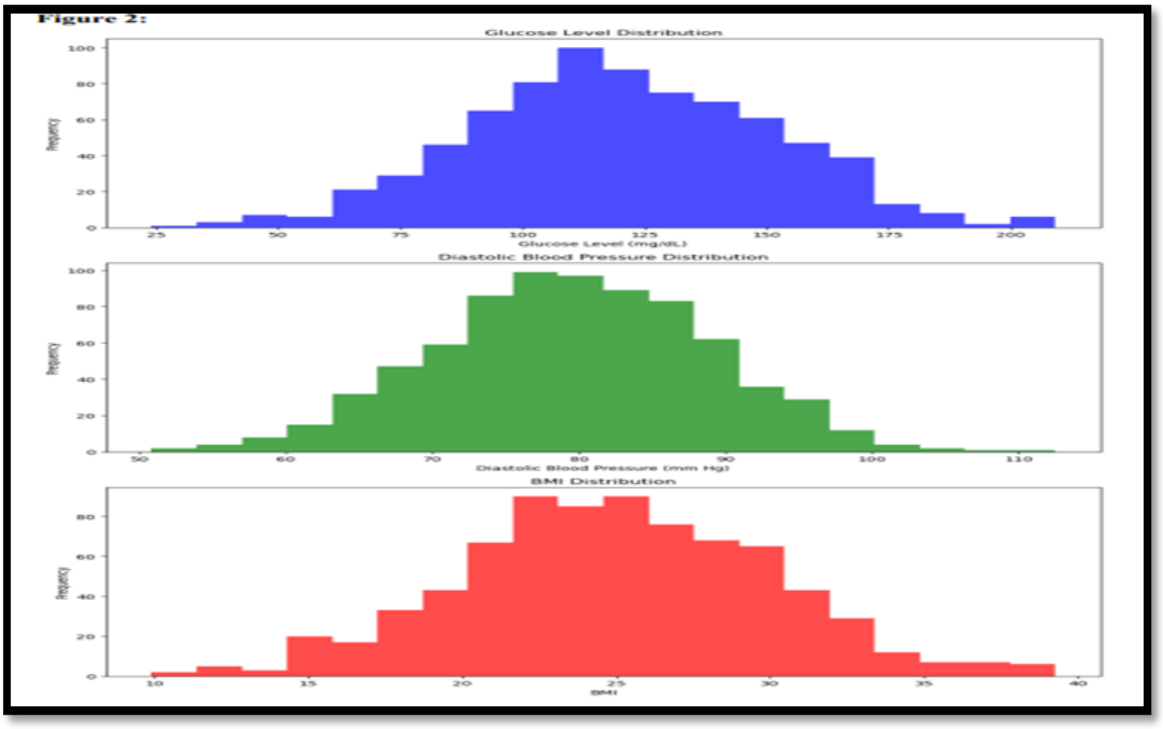
Table 2:

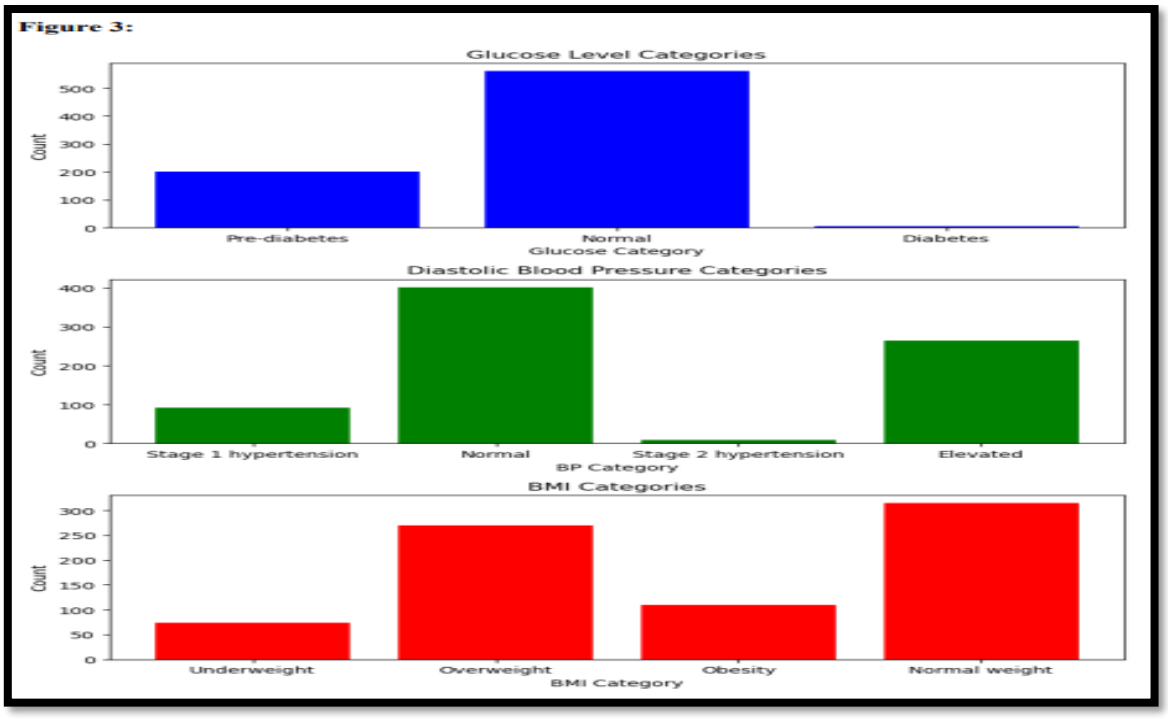
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Zia UA and Khan N [13], utilized machine learning methods to anticipate diabetes in medical datasets. They observed disparities in specific attribute values like glucose concentration (Gluc), blood pressure (BP), skin fold thickness (Skin), insulin, and BMI. They noted that zero values in these attributes are deemed abnormal and unreliable since they deviate from the anticipated range (refer to Figures 1, 2, and 3).

Figure1:

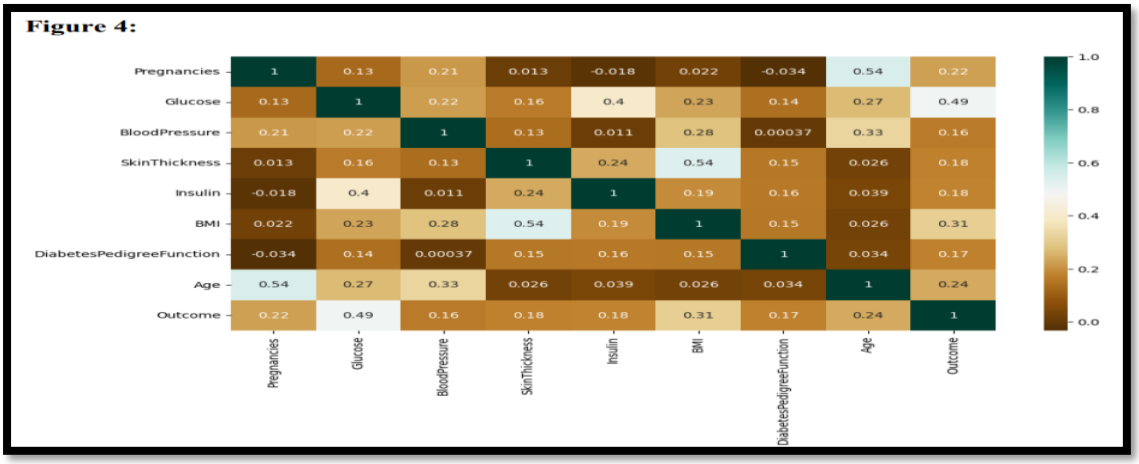






The scatterplot matrix serves as a useful tool for initially assessing the pairwise connections between features. If data points appear scattered, it suggests no clear relationship, whereas a roughly linear arrangement indicates a linear relationship. Analyzing the scatterplot matrix illustrated in Figure 4 reveals that the attributes showing the strongest correlation or proportionality consist of [pregnancy and age], [skin thickness and BMI], and [glucose and insulin], as evidenced by their scatterplots all exhibiting a positive correlation.

Figure 4:



5.2 Pre-processing data by using Quantile Transformer: Quantile Transformer is employed to convert the data distribution into a Gaussian distribution and normalize the outcome, aligning the values around a mean of 0 and a standard deviation of 1This technique alters the features so that they adhere to either a uniform or a normal distribution. Consequently, it causes the common values within a feature to become more dispersed while lessening the influence of outliers, making it a resilient pre-processing method. The quantile transformation offers an automated method for reshaping a numeric input variable to conform to a different data distribution, consequently enabling its utilization as input for predictive modeling. The Quantile Transformer has been utilized to pre-process the data in Table 3.

Table:3

Pre-processing data by using Quantile Transformer

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	0.666564	0.873569	0.037988	0.852495	-5.199338	0.239380	0.678858	1.220640
1	-0.731217	-1.275817	-0.416878	0.362241	-5.199338	-0.753452	-0.059586	0.139710
2	1.096804	1.708508	-0.589456	-5.199338	-5.199338	-1.330050	0.788098	0.203961
3	-0.731217	-1.120205	-0.416878	0.012660	0.426762	-0.535083	-1.258282	-5.199338
4	-5.199338	0.574460	-1.634019	0.852495	0.965625	1.447963	2.903113	0.269066

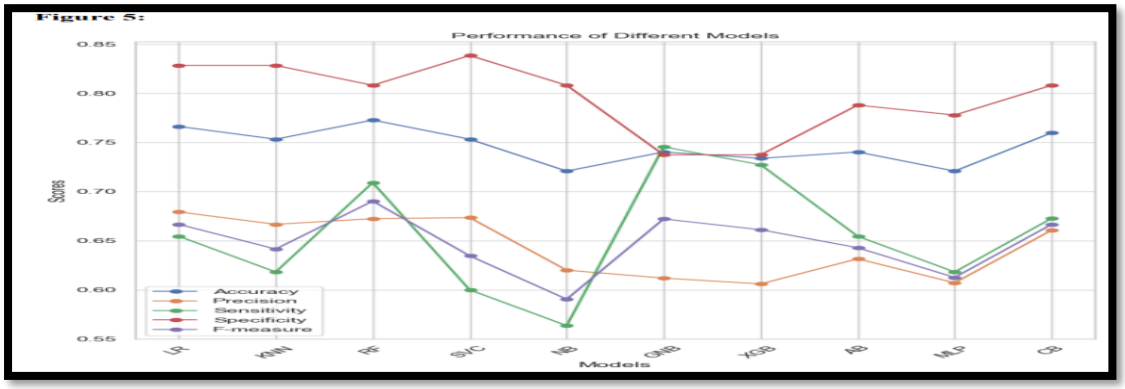
All the selected machine learning algorithms intended for application in this research paper are models designed for classification tasks.Evaluation of model performance commonly involved metrics like accuracy, precision, sensitivity, specificity, F-measure (F-score), and mean square error (MSE).

Table 4

	Accuracy	Precision	Sensitivity	Specificity	F-measure	MSE
SVC	0.753247	0.673469	0.600000	0.838384	0.634615	0.246753
NB	0.720779	0.620000	0.563636	0.808081	0.590476	0.279221
GNB	0.740260	0.611940	0.745455	0.737374	0.672131	0.259740
XGB	0.733766	0.606061	0.727273	0.737374	0.661157	0.266234
LR	0.766234	0.679245	0.654545	0.828283	0.666667	0.233766
KNN	0.753247	0.666667	0.618182	0.828283	0.641509	0.246753
RF	0.753247	0.649123	0.672727	0.797980	0.660714	0.246753
AB	0.740260	0.631579	0.654545	0.787879	0.642857	0.259740
MLP	0.740260	0.641509	0.618182	0.808081	0.629630	0.259740
CB	0.759740	0.660714	0.672727	0.808081	0.666667	0.240260

Explainable AI (XAI) revolves around the concept within artificial intelligence where users can understand the decisions made by a machine learning model[28]. Interpretability, on the contrary, refers to the ability to identify cause and effect relationships within such a model. While interpretability may be inherent, as evident in decision trees, it can also be integrated into a model after training by employing techniques to produce explanations.

Despite progress, limited research has concentrated on developing explainable machine learning models for non-communicable diseases[28].



It's worth noting that interpretable models may not always offer explanations to a degree where humans fully understand the process leading to a model's decision.

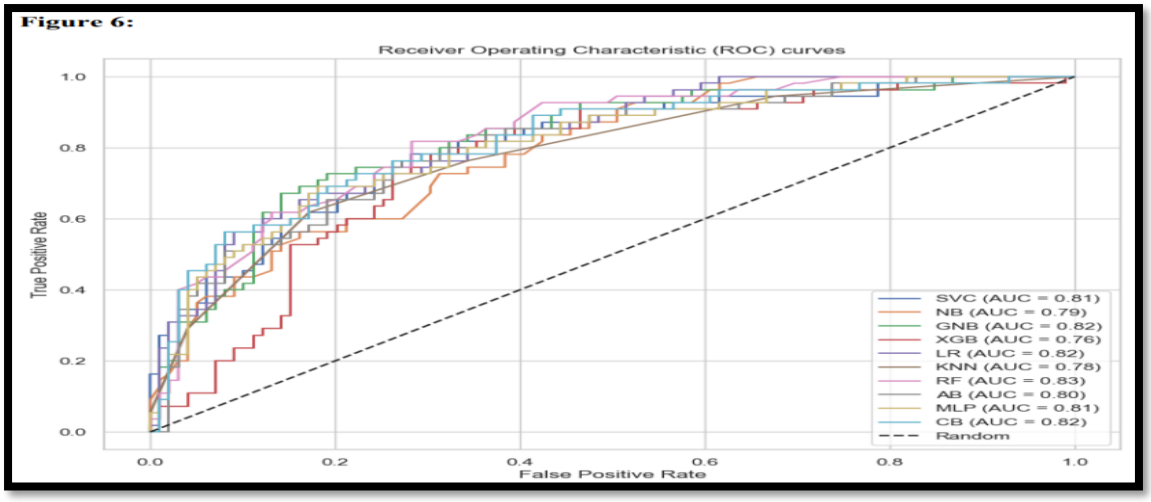
6. Result: The examination of machine learning (ML) methods for identifying and managing diabetes reveals promising advancements in diagnostic accuracy, risk prediction, and personalized treatment. Various ML algorithms, including supervised, unsupervised, and hybrid approaches, have been applied to diverse datasets, showcasing their potential in diabetes detection.

Through the scrutiny and comparison of 10 distinct algorithms, we ascertain the most proficient approach, based on a variety of evaluation metrics, for early-stage prediction of diabetes mellitus.

The Logistic Regression Machine algorithm (accuracy:76.63%,ROC–AUC score:0.82),

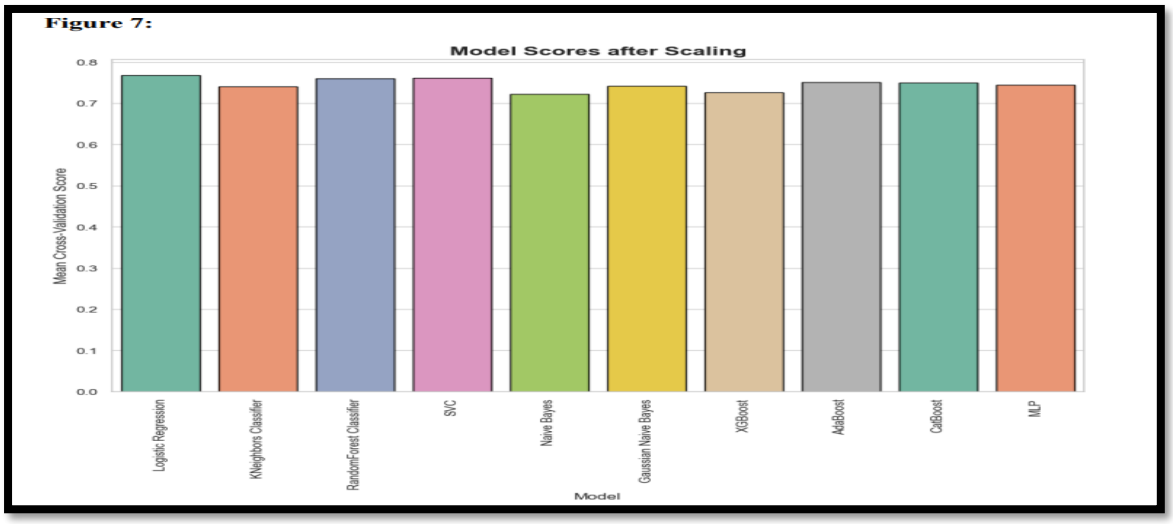
Cat Boost Machine algorithm (accuracy:75.97%,ROC–AUC score:0.82), and

Random Forest Machine algorithm (accuracy:75.32%,ROC–AUC score:0.83) emerges as the frontrunner, surpassing other models.



So, either LR or CB or RF could be considered as the best model for predicting diabetes, depending on whether you prioritize accuracy or overall performance across classes.

Model score after scaling is given in figure7:



7. Conclusion and prospective endeavours:

This paper has provided a comprehensive examination of machine learning (ML) methods in the identification and management of diabetes. The rising prevalence of diabetes globally underscores the importance of accurate and efficient diagnostic tools, and ML techniques offer promising avenues for addressing this need. This study has highlighted the potential of ML in enhancing diagnostic accuracy, risk prediction, and personalized treatment approaches.

However, the application of ML in diabetes identification is not without its challenges and limitations. Factors such as data quality, feature selection, interpretability, and clinical integration pose significant hurdles in leveraging ML effectively for diabetes care. Addressing these challenges and advancing ML techniques tailored to the specific needs of diabetes management are critical for realizing the full potential of ML-driven approaches in improving patient outcomes.

Moving forward, future research should focus on overcoming these challenges and exploring innovative ML techniques that prioritize data quality, interpretability, and clinical relevance. By addressing these gaps, ML-driven approaches have the potential to transform diabetes care by providing clinicians with accurate diagnostic tools and personalized treatment strategies. Overall, this paper contributes to the ongoing efforts in improving diabetes care through advanced technology and underscores the importance of continued research and innovation in this field.

References:

[1] M.W. Craven, J.W. Shavlik(1997), Using neural networks for data mining, Future Gener. Comput. Syst. 13 (2–3) 211–229, http://dx.doi.org/10.1016/s0167-739x(97)00022-8.
[2] Duplaga M (2004), The impact of information technology on quality of healthcare services. In: Bubak M, van Albada GD, Sloot PMA, Dongarra J(eds) Computational science - ICCS 2004. 4th international conference, Kraków, Poland, June 2004.
[3] Poston RS, Reynolds RB, (2006), Gillenson ML Technology solutions for improving accuracy and availability of healthcare records. Inf Syst Manag 24(1):59–71. https://doi.org/10.1080/10580530601038097
[4] C.L. Huang, M.C. Chen, C.J. Wang(2007), Credit scoring with a data mining approach based on support vector machines, Expert Syst. Appl. 33 (4) 847–856, http://dx.doi.org/10.1016/j.eswa.2006.07.007.
[5] Lassi M, Sonnenwald DH (2010), Identifying factors that may impact the adoption and use of a social science collaboratory: a synthesis of previous research.Inf Res15(3)
[6] Duan L, Street WN, Xu E (2011), Healthcare information systems: data mining methods in the creation of a clinical recommender system. Enterp Inf Syst 5(2):169–181. https://doi.org/10.1080/17517575.2010.541287
[7] Bonfiglio S (2012), The role of ICT in a healthcare moving from “clinical-centric” to “patient-centric”. In: Donnelly M, Paggetti C, Nugent C, Mokhtari M (eds) Impact analysis of solutions for chronic disease prevention and management.10th international conference on smart homes and health telematics, June 2012.

[8] Nakahara T, Hyogo H, Yoneda M, Sumida Y, Sumida Y, Fujii H et al (2013), Type 2 diabetes mellitus is associated with the fibrosis severity in patients with non-alcoholic fatty liver disease in a large retrospective cohort of Japanese patients. J Gastroenterol 49(11):1477–1484. https://doi.org/10.1007/s00535-013-0911-1
[9] S.A. Kaveeshwar, J. Cornwall(2014), The current state of diabetes mellitus in India, Australas. Med. J. 7 (1) -45.
[10] Salim Amour Diwani, Anael Sam(2014), Diabetes forecasting using supervised learning techniques, Adv. Comput. Sci.: Int. J. [S.l.] (ISSN:2322-5157) 10–18, Available at: <http://www.acsij.org/acsij/article/view/156.
[11] https://blogs.nvidia.com/blog/2016/07/29/ whats-difference-artificial intelligence-machine-learning-deep-learning-ai/.
[12] Saiti K, Macaš M, Štechová K, Pithová P, Lhotská L (2017), A review of model prediction in diabetes and of designing glucose regulators based on model predictive control for the artificial pancreas.
[13] Zia UA, Khan N (2017), Predicting diabetes in medical datasets using machine learning techniques. Int J Sci Eng Res 5(2):257–267
[14] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah(2018), "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare," in 2018 24th International Conference on Automation and Computing (ICAC), Sep. 2018, pp. 1–6. doi: 10.23919/IConAC.2018.8748992.
[15] D. Sisodia, D.S. Sisodia(2018), Prediction of diabetes using classification algorithms, Procedia Comput. Sci. 132 (2018) 1578–1585.
[16] G. Swapna, R. Vinayakumar, K.P. Soman, Soman KP(2018), diabetes detection using deep learning algorithms, ICT Express 4 (4) (2018) 243–246, http://dx.doi.org/10.1016/j.icte.2018.10.005, Elsevier B.V.
I. Contreras, J. Vehi(2018), Artificial intelligence for diabetes management and decision support: Literature review, J. Med. Internet Res. 20 (5).
[17] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, H. Tang(2018), Predicting Diabetes Mellitus with Machine Learning Techniques, Vol. 9, Frontiers in genetics, 2018, p. 515, http://dx.doi.org/10.3389/fgene.2018.00515.
[18] Haritha R, Sureshbabu D, Sammulal P (2018), Diabetes detection using principal component analysis and neural networks. In: Santosh KC, Hegadi RS (eds) Recent trends in image processing and pattern recognition.Second international conference, RTIP2R 2018, December 2018.
[19] T.M. Alam, et al.(2019), Informatics in medicine unlocked a model for early prediction of diabetes, Inform. Med. Unlocked 16 (2019) 100204.
[20] T. Mahboob Alam et al.(2019), "A model for early prediction of diabetes," Inform. Med. Unlocked, vol. 16, p. 100204, Jan. 2019, doi: 10.1016/j.imu.2019.100204.
[21] N. Yuvaraj and K. R. SriPreethaa(2019), "Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster," Clust. Comput., vol. 22, no. 1, pp. 1–9, Jan. 2019, doi: 10.1007/s10586-017-1532-x.
[22] T.M. Alam, et al.(2019), Informatics in medicine unlocked a model for early prediction of diabetes, Inform. Med. Unlocked 16 (2019) 100204.
A. Mujumdar and V. Vaidehi(2019), "Diabetes Prediction using Machine Learning Algorithms," Procedia Comput. Sci., vol. 165, pp. 292–299, Jan. 2019, doi: 10.1016/j.procs.2020.01.047.
[23] Chen Q, Alrowais R, Burhan M, Ybyraiymkul D, Shahzad MW, Li Y et al (2020), A self-sustainable solar desalination system using direct spray technology. Energy 205:118037. https://doi.org/10.1016/j.energy.2020.118037
[24] J. Chaki, S. Thillai Ganesh, S.K. Cidham, S. Ananda Theertan(2020), Machine learning and artificial intelligence-based diabetes mellitus detection and self-management: A systematic review, J. King Saud Univ. - Comput. Inf. Sci..
[25] G. Tripathi and R. Kumar(2020), "Early Prediction of Diabetes Mellitus Using Machine Learning," in 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp. 1009–1014. doi: 10.1109/ICRITO48877.2020.9197832.
[26] Cheng D, Ting C, Ho C, Ho C (2020), Performance evaluation of explainable machine learning on non-communicable diseases. Solid State Technol 63:2780–2793
[27] N.P. Tigga, S. Garg(2020), Predicting type 2 Diabetes using Logistic Regression accepted to publish in: Lecture Notes of Electrical Engineering, Springer.

[28] F. Alaa Khaleel and A. M. Al-Bakry(2021), "Diagnosis of diabetes using machine learning algorithms," Mater. Today Proc., Jul. 2021, doi: 10.1016/j.matpr.2021.07.196.
[29] J. J. Khanam and S. Y. Foo(2021), "A comparison of machine learning algorithms for diabetes prediction," ICT Express, vol. 7, no. 4, pp. 432–439, Dec. 2021, doi: 10.1016/j.icte.2021.02.004.
[30] Solanki P, Baldaniya D, Jogani D, Chaudhary B, Shah M, Kshirsagar A (2021), Artificial intelligence: new age of transformation in petroleum upstream. Pet Res (in press). https://doi.org/10.1016/j.ptlrs.2021.07.002
[31] M. Rady, K. Moussa, M. Mostafa, A. Elbasry, Z. Ezzat, and W. Medhat(2021), "Diabetes Prediction Using Machine Learning: A Comparative Study," 3rd Novel Intelligent and Leading Emerging Sciences Conference (NILES), pp. 279–282. doi: 10.1109/NILES53778.2021.9600091.
[32] M. U. Emon, M. S. Keya, Md. S. Kaiser, Md. A. islam, T. Tanha, and Md. S. Zulfiker(2021), "Primary Stage of Diabetes Prediction using Machine Learning Approaches," International Conference on Artificial Intelligence and Smart Systems (ICAIS), Mar. 2021, pp. 364–367. doi: 10.1109/ICAIS50930.2021.9395968.
[33] P. Cıhan and H. Coşkun(2021), "Performance Comparison of Machine Learning Models for Diabetes Prediction," 29th Signal Processing and Communications Applications Conference (SIU), Jun. 2021, pp. 1–4. doi: 10.1109/SIU53274.2021.9477824.
[34] D P Singh, J S Jassi, Sunaina, (2023), Exploring the Significance of Statistics in the Research: A Comprehensive Overview, Eur. Chem. Bull., 12(Special Issue 2),2089-2102
[35] https://www.cdc.gov/diabetes/basics/prediabetes.html.
[36] https://www.geeksforgeeks.org/machine-learning/
[37] Dr D P Singh
[38] drdps97@gmail.com
[39] ID: https://orcid.org/0000-0001-9494-4296