Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

A Review on Diabetes Classification Based on Machine Learning Algorithms

Jihan Askandar Mosa[1], Adnan Mohsin Abdulazeez[2]
[1]jihan.musa@dpu.edu.krd,[2]adnan.mohsin@dpu.edu.krd
[1]Technical College of Duhok, Duhok Polytechnic University, Kurdistan Region, Iraq
[2]Technical College of Engineering, Duhok Polytechnic University, Kurdistan Region, Iraq

Article Information

Submitted : 31 Mar 2024
Reviewed: 7 Apr 2024
Accepted : 24 Apr 2024

Keywords

Diabetes, Machine Learning, Classification, Algorithms, Healthcare.

Abstract

Diabetes, a chronic metabolic disorder, is a significant global health concern affecting millions of individuals worldwide. Early and accurate diagnosis of diabetes is crucial for effective management and prevention of complications. Machine learning (ML) techniques have emerged as powerful tools for analyzing diabetes-related data, aiding in the classification and prediction of diabetes types. This review provides a comprehensive overview of recent advancements in diabetes classification using ML algorithms, highlighting their strengths, limitations, and future directions. Various ML algorithms, including but not limited to support vector machines, decision trees, random forests, artificial neural networks, and ensemble methods, are discussed in details. Furthermore, data preprocessing techniques, feature selection methods, and evaluation metrics employed in diabetes classification studies are examined. Additionally, challenges such as data imbalance, interpretability, and generalization across diverse populations are addressed. Finally, potential avenues for future research to enhance the accuracy and applicability of ML-based diabetes classification systems are proposed.

A. Introduction

The number of diabetic patients has increased in recent years, necessitating the deployment of additional devices for patient monitoring [1]. According to WHO estimates, during the next 25 years, the number of people with diabetes will increase from 130 million to 350 million, although only 50% of patients would be aware of their condition [2].Diabetes is a significant health issue in both industrialized and developing nations [3].is a dangerous chronic disease that can lead to serious complications like heart attack, blindness, and kidney diseases, making it one of the deadliest diseases [4][5]. An abnormal blood glucose level, which is a symptom of diabetes, is a chronic metabolic disease brought on by either insufficient or ineffective usage of insulin [6]. Diabetes can be managed early to avoid complications and lower the chance of developing serious health problems. Both automated and manual diagnosis are possible; only manual diagnosis doesn't need the aid of a machine. Frequently, symptoms are too mild for skilled medical professionals to recognize [7]. Therefore, in order to prevent more harm to the body, it is advised to get a check-up as soon as any of these symptoms appear. This is because, unlike other diseases, diabetes can go undetected even in people who lead healthy lifestyles [8]. Building effective healthcare systems is crucial to addressing global health issues in light of the expanding population. These systems, which satisfy patients' concerns about quality and treatment options, are made to promote health and precisely detect illnesses as scientific research develops [9].

Various machine learning methods can be employed on diverse data structures. The study looks at predictive analysis in the medical field. Healthcare data sets are subjected to machine learning algorithms for analysis [10]. Machine learning techniques, such as SVM, NB, ANN, and other algorithms to identify patterns in data, can be extremely helpful in the early diagnosis, prediction, and preventative measures of diabetes in diabetic patients in order to improve the quality of care [11]. Skilled diagnosis aims to reduce erratic admissions by considering unique patient features, clinical and demographic factors. Diabetes is an inherited and ethnic illness, with poor dietary and lifestyle choices being the root cause. It affects more people in cities [2].

The subsequent sections of this paper are structured as follows: Section 2 delves into the Research Methodology, wherein a detailed explanation is provided. Section 3 encompasses the Literature Review focusing on Diabetes, followed by the Results and Discussion derived from the literature review in Section 4. Section 5 encapsulates the Conclusion and Future Directions, presenting the conclusions drawn from the study's findings and outlining potential avenues for future research.

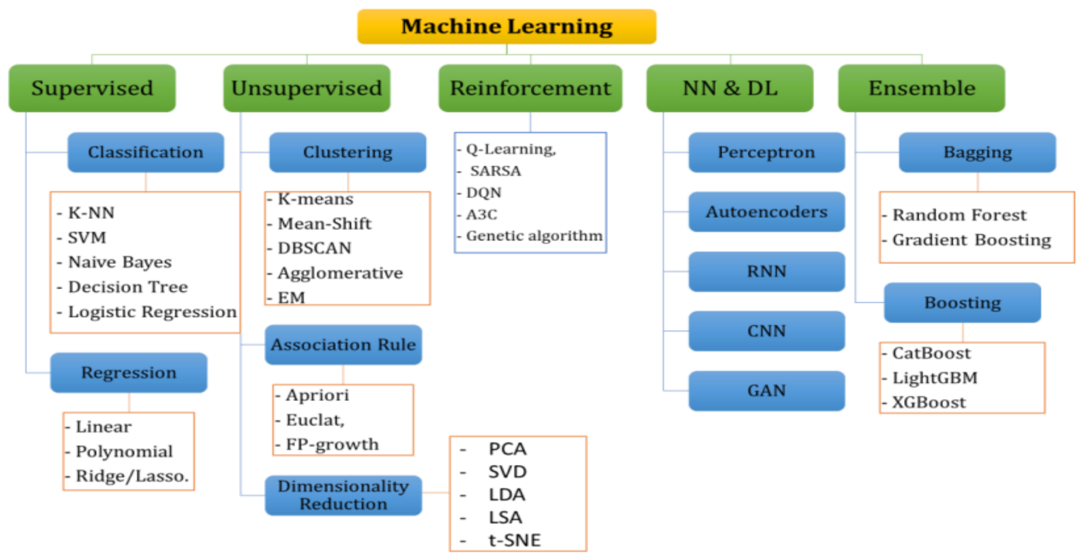

Figure 1 : Classification of ML Techniques [12]

B. Research Method

Effective data preparation and preprocessing techniques are essential for achieving the best classification results. The literature review's identification of a research gap is filled by the suggested algorithm. A meaningful strategy to handling missing values of attributes can significantly enhance the performance of a machine learning model [13].

- **Support Vector Machines**

Vapnik and Alexery Ya developed the supervised learning technique known as Support Vector Machine (SVM) in 1963 [14]. Analyze the provided data and create a function that may be used to the display of further data [15]. Guided learning techniques called Support Vector Machines (SVMs) look for patterns in data. Plotting the disease-predicting qualities in a "multidimensional hyperplane" allows the SVM algorithm to anticipate the occurrence of diabetes. It optimally classifies the classes by calculating the margin between two data groups [16].

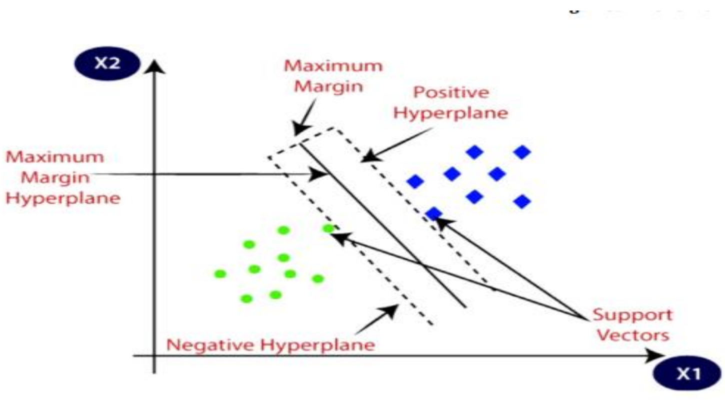

Figure 2 : Support Vector Machines [17]

- **Decision Trees**

The Decision Tree (DT) approach divides data repeatedly based on a specific variable to solve regression and classification problems in supervised machine learning [18]. A node-based, leaf-based, and branch-based hierarchical architecture. Each node in this model represents a feature test. The branches represent collections of features that point to the class labels, and the leaf represents a class label. Classification policies are symbolized by the path that leads from the root to the leaf [19]. A tree is defined in graph theory as a linked graph that is acyclic, undirected, and edge-free [20].

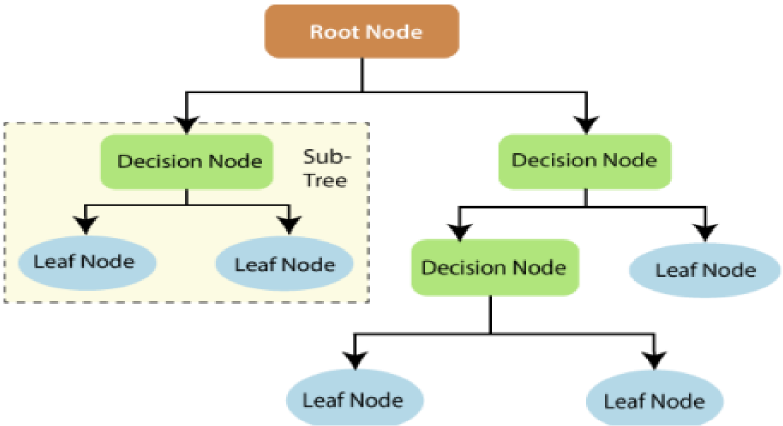

Figure 3 : Decision Tree [21]

- **Random Forests**

The different independent decision trees that make up Random Forest function as a group. The class that receives the most votes becomes the model's prediction, and these individual trees in the random forest divide the class prediction. The integration of multiple uncorrelated trees working together for the prediction process is the main idea behind random forests. To ensure that the behavior of every single tree in the model is not overly associated with the behavior of any other tree [22]. Each decision tree is built using a sample of data taken from the training dataset. The decision tree error will be estimated using the remaining data [23].

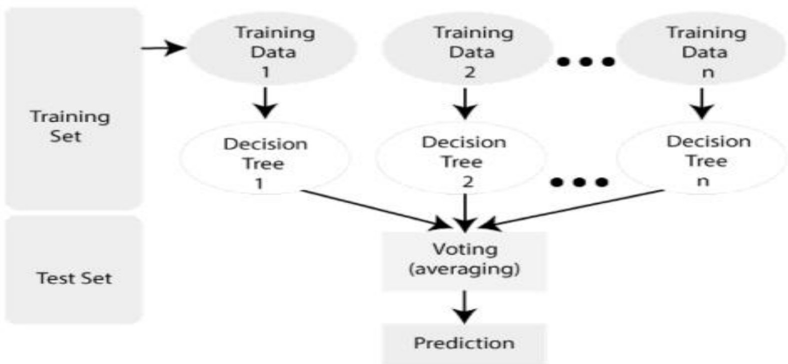

Figure 4 : Random Forests [24]

- **Artificial Neural Network**

The Artificial Neural Network is a systematic system for processing information. It functions similarly to how the human brain does [15].The feedforward neural network used in this paper is trained using Multilayer Perceptron's (MLP) based on the neural architecture of the human brain. A sigmoid activation function is used to facilitate non-linear relationship growth between the diabetes and non-diabetes classes, risk variables, and hidden layers that make up the network [25].

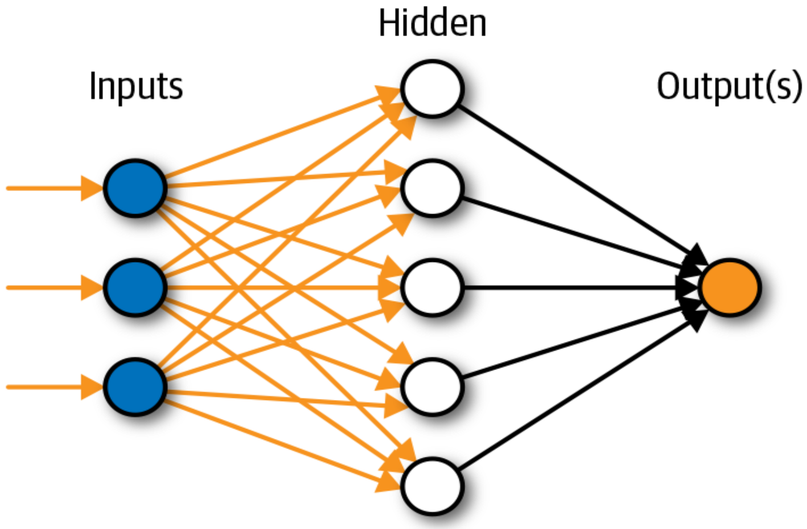

Figure 5: Artificial Neural Network

- **Adaptive Boosting**

AdaBoost, sometimes known as Adaptive Boosting, is a well-liked iterative boosting EML algorithm that works well with decision trees. It was first presented by Freund and Schapire in 1996. By overcoming their shortcomings and using an iterative process to correct the mistakes made by weak learners, the AdaBoost technique turns weak learners into strong ones [26]. This algorithm can be used in conjunction with many categorization algorithms to increase their efficiency [27].

- **Gradient Boosting**

Gradient Boosting is a popular ensemble technique first presented by for classification and regression tasks. This method improves the overall performance of the model by gradually adding weak learners to create an additive model [28]. Regression trees are an iterative decision tree for estimating continuous real-valued functions, and the gradient boosting model begins with a single leaf. With the goal of minimizing residual error, all potential splits on the available predictors are used to divide the data into two groups [29].

- **K-Nearest Neighbors**

The K-Nearest Neighbors (K-NN) technique remains one of the earliest and simplest classification algorithms in the field of machine learning [30].The dataset is stored, and each new observation is classified according to its likelihood of falling into the diabetes or non-diabetes class. The algorithm determines the separation between the new observation and every other observation in the dataset. After that, it allocates the new observation to the class that shows up in a set of k (positive integer parameter) neighbors the most times [25].

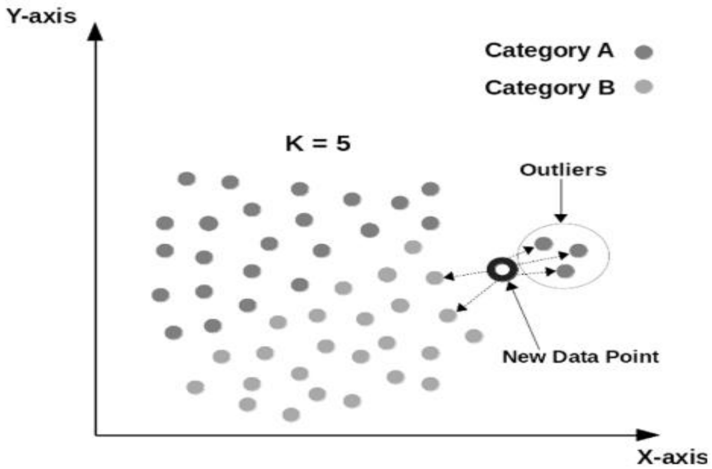

Figure 6 : K-Nearest Neighbors [24]

- **Logistic Regression**

Data can be categorized into discrete groups using regression analysis. Logistic regression often involves a dependent variable that can be true or false. In our case study, a diabetes diagnosis of 1 or 0 indicates a positive diagnosis or negative diagnosis. A linear classification model is what is meant to be understood when one speaks of "logistic regression," not regression [31].The cost function, often known as the sigmoid function, is used by the logistic function. This function converts probabilities into forecasts. Belagavi and associates [32].

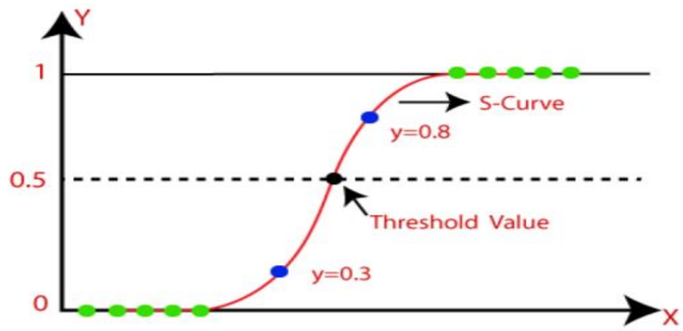

Figure 7: Logistic Regression [33]

- **Naive Bayes Classifier**

The Bayes Theorem of Probability underpins the Naïve Bayes algorithm. Another name for the Bayes Theorem is the "Bayes hypothesis." Because it can anticipate the output without requiring every observation in the training set, Naïve Bayes is an eager learning algorithm. From 1950 onwards, Naïve Bayes became a prominent issue in machine learning. It gained notoriety at the time for its effectiveness in content recovery. The Naïve Bayes approach was a superb content-order methodology in the 1960s. In medical diagnosis, it was applied [34]. A probability-based classifier combined with the Bayes theorem is the foundation of the Naïve Bayes classifier." High dimensional datasets can be well-characterized by the NB method [16].

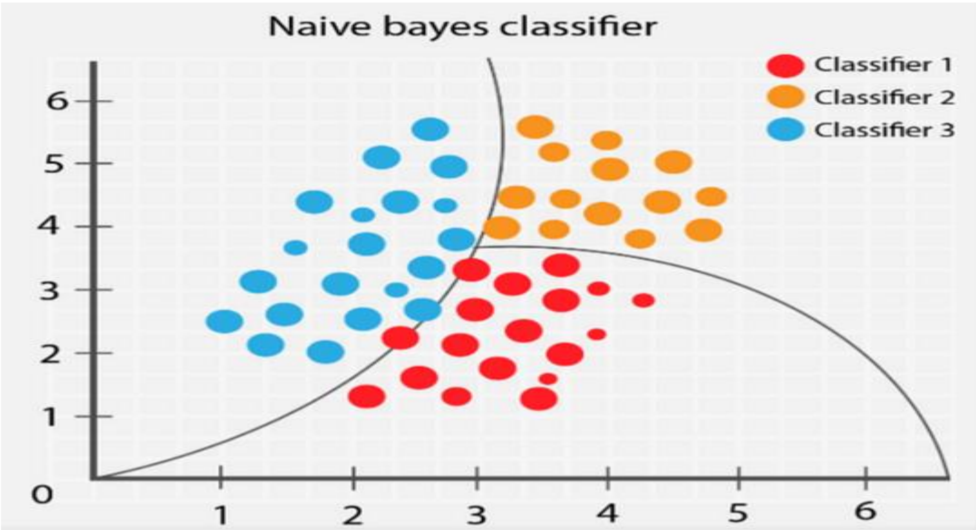

Figure 8 : Naive Bayes Classifier

C. Literature Review in Diabetes

Chang et al. [35] Presented a feature selection and model interpretability while discussing the application of interpretable AI techniques for diabetes prediction. It emphasizes how crucial it is that end users can comprehend and utilize models that are clear to them. The Pima Indian Diabetes dataset is used to test different machine learning algorithms for diabetes prediction, including Random Forest and J48 decision tree. The research highlights the importance of feature selection and the influence of distinct features on the outcomes of model prediction.

ÖZÇELİK and ALTAN. [36] Presented a machine learning techniques based on entropy-based features are used to classify diabetic retinopathy (DR). One of the main causes of blindness is diabetic retinal disease (DR), which is brought on by damage to the retina's blood vessels. Preventing vision loss in persons with diabetes mellitus requires early identification of diabetic retinopathy (DR). utilizing color retinal pictures, the study suggests a classifier model for early diagnosis of DR utilizing a genetic algorithm and the k-nearest neighbor (KNN) technique. 2400 retinal image data, including images from the "Non-Proliferative DR" (NPDR) and "Proliferative DR" (PDR) classes, were used in the model's training. Ten-fold cross-validation was used to assess the model's validity. Machine learning techniques were used to test the model's performance to Gaussian Naive Bayes (GNB).

Phongying and Hiriote. [37] Described how to create decision trees and how to identify the final class of test objects by aggregating votes from various trees. It also includes references to related studies on the use of classification algorithms to predict diabetes and the comparison of the accuracy levels of KNN and SVM algorithms in predicting heart disease. The application of metabolomics in prediabetes and diabetes as well as the usage of big data mining for diabetes risk prediction are also discussed in the document.

AHMED et al.[38] Described a fuzzy logic-based machine learning-based diabetes decision support system that combines two popular machine learning approaches. With an accuracy of 94.87%, the suggested model outperformed current systems and may save lives by facilitating early diagnosis and preventative actions. The document underscores the importance of this field in healthcare by citing a number of studies and research projects on machine learning algorithms for diabetes prediction.

Sadeghi et al. [39] Worked to Evaluate machine learning algorithms for handling class imbalance in medical informatics and decision making. It uses metrics like g-mean, F1-measure, MCC, ROC curve, and AUC, and discusses strategies like threshold moving, cost-sensitive learning, and sampling techniques. Early detection and prediction of type two diabetes mellitus incidence could reduce complications.

Mushtaq et al. [40] Presented the implementation of machine learning techniques to predict diabetes mellitus is discussed in the paper, with a number of studies concentrating on several facets like risk score prediction, deep learning approaches, performance analysis, and prognostic modeling. The publication also offers statistical details regarding the dataset that was utilized in the research, including characteristics like age, result, skin thickness, insulin, BMI, glucose levels, blood pressure, and pregnancies. The outcomes of under sampling and oversampling techniques are also discussed, as well as the data balancing techniques used in the study.

Sivaranjani S et al. [41] Discussed diabetes prediction with SVM and RF machine learning algorithms. To increase prediction accuracy, dimensionality reduction strategies and feature selection methodologies are used. RF shows itself to be more effective, with an accuracy rate of 75%. To accurately predict the beginning of diabetes, three key processes are required: feature selection, dimensionality reduction, and data pre-processing.

Aamir et al. [42] Discussed the use of machine learning and fuzzy logic techniques for diabetes detection using the Pima Indians Diabetes (PID) dataset. It outlines the data pre-processing steps, including data normalization and division into training and testing sets. The classification process involves the construction of fuzzy classifiers and the determination of the degree of belongingness for each instance of the dataset. Additionally, it presents a summary of various machine learning techniques for diabetes detection, along with their respective accuracies. Furthermore, it highlights the application of fuzzy logic techniques, such as fuzzy rules generation and optimization, for diabetes prediction, achieving accuracies of 81% and 85.33% respectively.

Aftab et al.[43] Presented an early diabetes detection utilizing a cloud-based intelligent framework enabled by supervised machine-learning methods and fuzzy systems is described in the publication. The two layers that make up the framework are training and testing, each having several steps. In the training layer, three supervised classification techniques—ANN, DT, and naïve Bayes—are used for classification along with dataset selection, pre-processing (data cleaning, normalization, and splitting). To address the uncertainty in the base classifier findings, the paper also presents an ensemble classification model with a fuzzy rule inference engine. By enabling early type-2 diabetes diagnosis, the study hopes to help patients take better care of their diets, lifestyles, and medications before the condition gets worse.

Ahmed et al. [44] Discussed Diabetes Mellitus is a global disease-causing significant death. Machine learning (ML) approaches are being used to detect the disease early. This study explores supervised ML models like Decision tree, Naive Bayes, k-nearest neighbor, Random Forest, Gradient Boosting, Logistic Regression, and Support Vector Machine for diagnosing diabetes. The results show improved accuracy, with the highest accuracy model integrated into a web application.

Azad et al. [45] Presented a prediction and detection of diabetes is discussed in relation to the use of machine learning techniques in healthcare. In order to overcome class imbalance, it emphasizes the need for oversampling approaches like SMOTE. It also underlines the difficulties associated with outliers, missing values, and class imbalance in medical datasets. Decision trees (DT) are used for prediction in the suggested model, while genetic algorithms (GA) are used for feature selection. Additionally, the publication includes experimental results that demonstrate how several approaches, including GA and SMOTE, affect the prediction accuracy of diabetes diagnosis.

Bansal and Singhrova. [46] Discussed the application of various machine learning algorithms for predicting and diagnosing medical conditions such as diabetes and breast cancer. It includes references to studies that have utilized algorithms like Support Vector Machine (SVM), J48 classifier, adaboost with Choice Stump, and others to achieve high precision in predicting diabetes and breast cancer. The document also highlights the use of AI techniques for classifying medical datasets and emphasizes the importance of machine learning in medical data analysis and prediction.

Gayathri et al. [47] Discussed the assessment of classifiers for diabetic retinopathy (DR) grading through the application of cross-validation techniques and M-CNN features. It highlights how crucial it is to use stratified random sampling to address imbalanced databases and how confusion matrices may be used to calculate a variety of assessment measures. The examination made use of three databases with a large number of images: IDRiD, Kaggle, and MESSIDOR. The paper also presents important assessment metrics and emphasizes their importance in evaluating classifier performance, including accuracy, precision, recall, F1-score, specificity, and Kappa-score. It also discusses the challenge of applying precise class efficiency measures for analysis and the computation of weighted average values for a simple system evaluation.

Ghosh et al. [48] Worked a study to explores the application of machine learning algorithms for detecting diabetes. Various techniques, including Random Forest, Support Vector Machine, AdaBoost, and Gradient Boosting, are compared using the Pima Indians diabetes dataset. Results indicate that Random Forest outperforms other algorithms in terms of accuracy. The research emphasizes the potential of machine learning in enhancing disease detection and management, particularly in the context of diabetes.

Jian et al. [49] Focused to use supervised classification algorithms on a dataset from the Rashid Center for Diabetes and Research in the United Arab Emirates to predict and categorize eight diabetes complications. Feature selection and data normalization were used in conjunction with preprocessing procedures to handle missing values and unbalanced data. Model performance was enhanced by feature selection and the application of balancing strategies such as SMOTE and cluster centroids. In addition to defining different diabetes problems such as obesity, dyslipidemia, metabolic syndrome, neuropathy, nephropathy, diabetic foot, hypertension, and retinopathy, the study also indicated differences in training times between the models.

Khaleel and Al-Bakry [50] Discussed how the Pima Indian Diabetes dataset can be used to train machine learning algorithms to predict the onset of diabetes. It emphasizes the value of early prediction in reducing diabetes severity and risk factors while showcasing machine learning's potential in the medical industry. When comparing the accuracy of several machine learning algorithms' predictions, the study finds that Logistic Regression is more effective in predicting diabetes than both Naïve Bayes and K-nearest Neighbor algorithms. The article also describes how to rescale features using Min Max Scaler and gives a general review of the K-nearest Neighbor algorithm, highlighting its versatility and ease of use in pattern recognition.

Nishat et al. [51] Presented the significance of early detection and precise prediction for successful treatment as it examines the prevalence and effects of diabetes mellitus. It emphasizes how machine learning algorithms can be used to predict diabetes by referencing several studies that have used diverse methods, including neural networks, logistic regression, support vector machines, and others. The report also highlights how difficult it is to forecast diabetes with high accuracy using machine learning models and how current technology can both enhance predictions and reduce healthcare costs. It also describes the performance metrics (accuracy, sensitivity, precision, F1-score, specificity, and ROC_AUC) that were used to assess the various methods.

Khanam and Foo. [52] Focused the study highlights the significance of finding hidden patterns in data for precise decision-making as it explores the application of these techniques to preprocess healthcare data and automate diabetes prediction. Several researchers have used the Pima Indian Diabetes dataset and machine learning approaches to predict diabetes, with accuracy ranging from 75.7% to 77.21%. The study uses a variety of classification algorithms to predict diabetes in individuals and assesses their effectiveness through a range of testing techniques.

Nadeem et al. [53] Discussed the fusion-based prediction method for diabetes detection that combines Artificial Neural Networks (ANN) and Support Vector Machines (SVM). Outperforming solo SVM and ANN models, this method obtained an accuracy of 94.67%. True-positive rate, misclassification rate, and system performance indicators were all improved by the fusion method. With the potential to forecast the beginning of diabetes, it compiles a cohesive dataset from several sources for improved alignment with machine learning algorithms.

Nahzat and Yağanoğlu. [54] Focused the study highlights the value of early diagnosis and treatment to enhance patient outcomes as it explores the application of machine learning approaches for diabetes prediction. It gives a general review of diabetes, its different forms, and the health hazards that go along with it, emphasizing the disease's worldwide effects. In order to predict diabetes, the study makes use of the Pima Indian Diabetes Dataset and a number of machines learning methods, including K-Nearest Neighbors, Random Forest, Support Vector Machine, Artificial Neural Network, and Decision Tree.

Alpan and Ilgi. [55] Presented the study different classification algorithms—including WEKA as a data mining engine—are used to categorize a diabetic dataset. There are 520 cases in the dataset with 17 attributes. Different classification methods, including k-NN and SVM, performed differently in accurately classifying the examples. For each method, the outcomes of evaluation metrics for accuracy, sensitivity, specificity, positive and negative precision, correctness, and error rate are also given. At 98.07%, the k-NN algorithm demonstrated the highest accuracy, while Bayes Net demonstrated the lowest accuracy, at 86.92%.

Assegie and Nair [56] Presented the role of early diabetes diagnosis in medicine is discussed in this work, along with the use of machine learning models for diabetes disease categorization, including Random Forest, Gaussian Naive Bayes, and Linear Support Vector Machine (LSVM). Because many machine learning models vary in their accuracy and complexity, it draws attention to the difficulties in creating an ideal model for disease classification. LSVM, Gaussian Naive Bayes, and Random Forest algorithms for diabetes prediction are developed and performed, and these research questions are addressed in this study. It also highlights the usefulness of LSVM in diabetes dataset categorization and addresses previous studies on diabetes diagnosis using machine learning models.

Daanouni et al. [57] Discussed the use of machine learning algorithms like KNN, Decision Tree, ANN, and DNN for predicting diabetes. It highlights the importance of classification accuracy, sensitivity, and specificity in evaluating these models. Feature selection is used to enhance classifier efficiency. The study shows that after pre-processing the dataset, more accurate results were achieved. DNN demonstrates high capability in classifying diabetic disease with a ROC of 92.36%. Comparison with related work indicates promising results in accuracy and sensitivity.

HASAN et al. [58] Discussed how the PID dataset might be used to train machine learning models for diabetes prediction. It draws attention to how preprocessing methods like outlier rejection and missing value imputation affect the quality of the dataset. The study contrasts the effectiveness of several machine learning models and highlights the superiority of feature selection techniques like correlation-based selection over PCA and ICA. It also discusses how important model assembling is to getting good performance for diabetes prediction, specifically with the XB model. The paper also presents a comparison between the suggested method and previous research, demonstrating the improved performance of the former in terms of balanced accuracy and AUC.

Katarya and jain. [59] Presented the study emphasizes the importance of early detection and prediction of diabetes and suggests the potential for further improvement using other ensemble machine learning methods. the application of six different machine learning algorithms (KNN, Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest, and Logistic Regression) for the detection and prediction of diabetes. It compares the performance of these algorithms based on five metrics: accuracy, recall, precision, f1-score, and ROC-AUC curve. The results indicate that Random Forest outperforms the other algorithms with an accuracy of 84%, precision of 83, recall of 76, f1-score of 86, and ROC-AUC score of 83.

Pethunachiyar. [60] Discussed the significance of early detection of diabetes mellitus (DM) using machine learning algorithms, particularly Support Vector Machines (SVM). It emphasizes the global impact of diabetes, especially in India, and the potential risks associated with untreated diabetes. The paper highlights the role of machine learning in healthcare and medical fields, emphasizing the need for early diagnosis for improved quality of life. It also provides an overview of related work in the field, including the use of different classification and regression techniques for diabetes prediction and treatment.

Soni and Varma. [61] Discussed various research studies on predicting diabetes onset using machine learning techniques. Different supervised machine learning methods such as SVM, Logistic regression, ANN, Bayesian, KNN, and other algorithms are employed to predict diabetes disease. The studies compare the performance and accuracy of these algorithms and propose effective techniques for earlier detection of diabetes. Additionally, the document explains the concept of finding the better hyperplane by calculating the distance between the planes and the data, known as Margin, and discusses the K-Nearest Neighbor (KNN) algorithm as a lazy prediction technique for solving classification and regression problems based on similarity measures.

Tigga and Grag. [62] Discussed the prevalence of diabetes and prediabetes in urban and rural India, presenting the results of the Indian Council of Medical Research–India Diabetes (ICMR–INDIAB) study. It also explores various machine learning and data mining methods used in diabetes research, comparing their effectiveness in predicting diabetes. The study evaluates the performance of different classification methods such as logistic regression, K-nearest neighbor, support vector machine, naive Bayes, decision tree, and random forest on a specific dataset and the PIMA database. The results indicate that the random forest classifier demonstrates the highest accuracy, sensitivity, specificity, precision, and F-measure, making it the most effective method for the dataset.

tripathi and Kumar. [63] Discussed the use of machine learning algorithms for the early prediction of diabetes mellitus. It highlights the significance of personalized healthcare and the role of machine learning in identifying diseases and their symptoms at an early stage. The study compares the performance of four classification algorithms - Linear Discriminant Analysis (LDA), K-nearest neighbor (KNN), Support Vector Machine (SVM), and Random Forest (RF) - using the Pima Indian Diabetes Database (PIDD) for experimental analysis. The document emphasizes the impact of diabetes on various organs and the importance of early diagnosis, citing statistics on the prevalence of diabetes globally.

Xue et al. [64] Discussed the application of machine learning algorithms for diabetes prediction, comparing the performance of three classification algorithms: naive Bayes, SVM, and LightGBM. It presents the confusion matrix evaluation test results, indicating that SVM has the highest accuracy for diabetes prediction. The document emphasizes the importance of early detection of diabetes and the role of machine learning in revolutionizing diabetes risk prediction.

Table 1. Literature Review Summary Table

Authors and year of pub.	Dataset	Algorithms	Pros	Cons	Result
Chang et al.2023 [35]	PID	J48 DT, RF, NB,	diabetes efficiently processes and analyze large data sets, aiding decision-making and patient management, leading to accurate predictions and personalized feedback.	a lack of diversity in the dataset used for training and testing machine learning models, potential data imbalance, and insufficient discussion of potential biases.	80%
ÖZÇELİK and ALTAN.2023[36]	Asia Pacific Tele-Ophthalmology Society (APTOS 2019)	KNN, GNB	The proposed model outperforms the GNB model in enabling early diagnosis of DR disease with high accuracy and low computational	The GNB model's precision, recall, and F1-score values are around 85% lower than the proposed model's performance	93.83%

			cost.		
Phongying and Hiriote.2023 [37]	Department of Medical Services, Bangkok	KNN, DT, SVM, RF	terms enhance efficiency by incorporating risk factors like body mass index and family history of diabetes, enhancing their effectiveness in classifying potential patients.	diabetes does not include certain risk factors like exercise, lifestyle, and dietary management, as well as certain metabolites associated with prediabetes and diabetes	97.5%
AHMED et al.2022 [38]	UCI	SVM, ANN	disease detection by using real-time patient data, achieving a prediction accuracy of 94.87%.	the dataset used in the research contains only 520 instances and 17 attributes based on diabetic symptoms.	94.87%
Sadeghi et al.2022 [39]	Tehran Lipid and Glucose Study (TLGS)	DNN, XGBoost, RF	improving diabetes prediction accuracy through imbalance solving strategies such as sampling methods, cost-sensitive learning, and threshold moving.	minority class performance due to noisy rare samples, small sample size, and biased evaluation metrics towards the majority class.	54.8%

Mushtaq et al.2022 [40]	PIMA	SVM, NB, KNN, RF,	The logistic regression model employs L1 and L2 regularization structures to minimize overfitting, reducing coefficient val ues and preventing zero-valued coefficients.	logistic regression model is the potential for overfitting, which can be mitigated	Range of 80.7%, to 82.0%
S et al.2021 [41]	PIMA	SVM, RF	analyze large healthcare data volumes to predict diabetes onset, potentially improving patient outcomes and reducing healthcare costs.	The limitation is the lack of information about the specific machine learning techniques used in the studies mentioned.	83% RF, 81.4% SVM
Aamir et al.2021 [42]	PID	RF, NN, KNN	The proposed fuzzy classifiers outperform existing techniques in accuracy, demonstrating their effectiveness in detecting diabetes.	the proposed fuzzy classifiers may not perform as effectively in detecting diabetes.	96.47%
Aftab et al.2021 [43]	PID	ANN, DT, NB	Early detection of type-2 diabetes allows patients to improve	the potential for a 4% miss rate in achieving	95.2%

			lifestyles, dietary habits, and start taking medication before the disease worsens.	96% accuracy with the (ANN) during the training process.	
Ahmed et al.2021 [44]	PID, UCI	DT, NB, KNN, RF, GB, LR, SVM.	improved through various pre-processing techniques such as outliers' removal, handling missing values, data standardization, and label encoding.	the model's accuracy may vary depending on the dataset and the machine learning method used, with a range of accuracy from 2.71% to 13.13%.	Range of 2.71% to 13.13%
Azad et al.2021 [45]	PID	GA, DT, PMSGD	The PMSGD model effectively addresses data imbalance and dimensionality issues in diabetes datasets	conversations between providers and patients due to unstructured data and incomplete, redundant, irrelevant, and noisy information	82.1256%.
Bansal and Singhrova.2021 [46]	Kaggle site	LR	Bagging enhances model accuracy and stability by combining multiple bootstrapped samples, reducing variance and improving the	bagging algorithm is that it requires a larger number of base learners, which can increase computational complexity and training time.	75.32%

			stability of classifiers through diversification of models.		
Gayathri et al.2021 [47]	IDRiD, Kaggle, MESSIDOR	SVM, RF, J48	M-CNN extracts features from a small database, training machine learning classifiers for diabetic retinopathy grading, evaluating performance and aiding in selecting the most effective classifier.	The disadvantage is that training a CNN with a small database doesn't produce a good classification model.	99.62%
Ghosh et at .2021 [48]	PID	RF, SVM, AB, GB	The Random Forest approach achieves 99.35% accuracy for early diabetes diagnosis, with pipeline structures and an Android application promising further improvement in prediction accuracy.	The disadvantage of the SVM and AB methods is that they exhibited the lowest performance when compared to the Random Forest approach.	99.35%
Jian et al.2021[49]	PIDD, RCDR	SVM, LR, DT, RF, AdaBoost,	used in data mining and machine learning	may be due to the small dataset size	97.8%

		XGBoost	algorithms to classify and predict diabetes complications, including metabolic syndrome, dyslipidemia, hypertension, obesity, diabetic foot, neuropathy, retinopathy, and nationality.	and the use of specific machine learning algorithms and techniques, which may affect the generalizability of the findings.	
Khaleel and Al-Bakry .2021 [50]	PID	LR, NB, KNN	Logistic Regression was found to be more effective in predicting diabetes than other classifiers	The limitation of the study is that it focuses on a specific dataset and may not be generalizable to other populations or datasets.	LR 94%, NB 79%, KNN 69%
Nishat et al.2021 [51]	Kaggle Diabetes Dataset, Frankfurt hospital, Germany	LR, NB, SVM, GB, ADB, RF, GP, SGD, ANN, KNN	analyzing large datasets to identify patterns and make predictions, which can be particularly useful in the field of healthcare for predicting and diagnosing diseases such as diabetes.	the potential for overfitting, which can occur when a machine learning algorithm performs well on the training data but fails to generalize to new, unseen data.	Range of 78% to 98.25%
Khanam and Foo.2021 [52]	PID	DT, KNN, RF, NB, AB, LR, SVM	The dataset has undergone preprocessing, including outlier removal, feature selection, and	the dataset contains missing values for certain attributes, which can	Range of 75.7% to 77.21%

			normalization, to enhance machine learning model performance by ensuring clean, relevant, and standardized data.	affect the accuracy of the machine learning models trained on this data.	
Nadeem et al.2021 [53]	NHANES, PID	SVM, ANN	The advantage is an improvement in performance owing to the fusion of both Support Vector Machine and Artificial Neural Network approaches.	Low-quality contextual data for training and validation of algorithms in data-driven applications and services poses a disadvantage, compromising the accuracy of the resulting models.	94.67%.
Nahzat and Yağanoğlu.2021 [54]	PID	KNN, RF, SVM, ANN, DT	The Random Forest algorithm is renowned for its simplicity and usability, making it a popular choice for classification and regression due to its ability to handle large datasets.	the K-Nearest Neighbors (KNN) algorithm is its slow learning manner, which delays data generalization until classification.	88.31%
Alpan and Ilgi,2020 [55]	UCI	BN, NB, DT(J48), RT, RF, KNN,	The Gain Ratio adjusts the information gain for each attribute	the information gain measure is its bias toward tests with	98.07%

		SVM	to account for the breadth and uniformity of the attribute values, allowing for a more balanced selection of attributes.	many outcomes, which can lead to a preference for selecting attributes with a large number of values.	
Assegie and Nair.2020 [56]	kaggle, MNIST, UCI	GNB, LSVM, RF	LSVM has better performance compared to the other algorithms on diabetes prediction, on the random tests conducted on the models in the classification of diabetes.	the Random Forest model has the lowest accuracy compared to the other models in the classification of diabetes.	SVM 78.39%,GNB 74.15%,RF 72.72%
Daanouni et al.2020[57]	UCI	DT, KNN, ANN, DNN	Feature selection in classification improves performance, reduces complexity, and enhances efficiency, making it effective in tasks like machine learning and computer vision.	using feature selection for classification is aimed at reducing the dimensionality and noise in datasets in order to improve performance and reduce complexity and efficiency of classifier methods.	92.36%.
HASAN et al.2020 [58]	PID	KNN, DT, RF, AdaBoost, NB, XGBoost	lies in the ability of the proposed framework to be applied to various medical contexts, demonstrating	is that data standardization cannot guarantee improved performance in the case of	0.950

			its generalizability and versatility in predicting disease classes.	tree-based classifiers.	
Katarya and jain.2020 [59]	PIMA	KNN, NB, SVM, DT, RF, LR	Support Vector Machine (SVM) enhances data classification by finding the maximum margin hyperplane, creating clear separation between data points, and providing superior generalization and performance.	using support vector machine (SVM) is that it may not perform well with large datasets and can be sensitive to noise in the data.	84%
Pethunachiyar.2020 [60]	UCI	SVM	Machine learning algorithms are crucial in early disease detection, especially in medical fields like diabetes, providing confidence in diagnosis and improving quality of life.	Machine learning algorithms face challenges in processing and mining knowledge from large volumes of medical data with different formats, which may hinder early diabetes detection.	Range 90% to 100%

Soni and Varma.2020[61]	UCI	KNN, LR, DT, SVM, GB, RF	The method achieves 77% classification accuracy, aiding healthcare in early diabetes diagnosis and decision-making, potentially saving lives.	the experimental results achieved a classification, which may not be considered high enough for some applications.	77%
Tigga and Grag.2020 [62]	PID	LR, KNN, SVM, NB, DT, RF	Naïve Bayes, Decision Tree, and Random Forest classification methods are all excellent for their simplicity, accuracy, stability, and visualization. Naïve Bayes outperforms others, Decision Tree offers high accuracy, and Random Forest aggregates votes.	do not account for the potential imbalance in the dataset, where non-diabetic cases outnumber diabetic ones. This can lead to biased results and affect the accuracy of the methods.	94.10%
tripathi and Kumar.2020[63]	PID	LDA, KNN, SVM, RF	K-fold cross-validation effectively evaluates model performance by dividing the dataset into multiple sections, providing statistically reliable results	the Support Vector Machine algorithm is the challenging task of selecting the optimal hyperplane in the dimensional space, which	87.66 %

			and a more accurate assessment of the model's performance.	can be difficult.	
Xue et al.2020 [64]	UCI	SVM, NB, LightGBM	LightGBM is known for its faster training efficiency and is considered to be distributed and efficient.	The disadvantage of the Naïve Bayes classifier is its assumption of strong (naive) independence between features, which may not hold true in real-world data.	Range of 88.46% to 96.54%

D. Result and discussion

Between 2020 and 2023, the field of machine learning witnessed a surge in research and practical applications focused on predicting and diagnosing diabetes. Researchers and practitioners delved into a variety of datasets, including the Pima Indian Diabetes dataset (PID), UCI datasets, and others, to explore novel approaches. Commonly utilized ML models during this period included Gaussian Naive Bayes (GNB), K-Nearest Neighbors (KNN), Decision Trees (DT), Random Forest (RF), Support Vector Machines (SVM), among others.

Accuracies reported in studies varied widely, with some achieving impressive rates above 90%, while others fell within the 70% to 80% range. These reported outcomes were heavily influenced by factors such as evaluation metrics, preprocessing techniques, dataset characteristics, and the choice of ML algorithms. Notably, different ML models demonstrated distinct performance levels across experiments. For instance Random Forest and Support Vector Machine methods often stood out for their superior accuracy in diabetes prediction tasks. K-Nearest Neighbors and Logistic Regression, though commonly employed, each possessed its own set of advantages and disadvantages. Each study meticulously highlighted both the strengths and limitations of its proposed methodologies. While some approaches showcased high accuracy in diabetes prediction, enabling early detection and personalized interventions, they also acknowledged challenges such as dataset imbalance, data quality issues, computational complexity, and model interpretability

E. Conclusion and Future work

The comprehensive review of literature on the application of machine learning (ML) algorithms for diabetes prediction highlights a promising avenue for improving early diagnosis and patient outcomes. The studies reviewed underscore the importance of leveraging ML techniques to address the global challenge of diabetes mellitus effectively. Through the utilization of diverse ML algorithms such as decision trees, support vector machines, random forests, and neural networks, researchers have demonstrated significant advancements in predicting diabetes onset and complications. These techniques offer valuable insights into disease progression, risk factors, and personalized treatment strategies.
Future research should focus on refining ML algorithms, enhancing model interpretability, and integrating predictive analytics into clinical practice seamlessly. By harnessing the full potential of ML in diabetes management, we can improve patient outcomes, reduce healthcare costs, and ultimately mitigate the burden of this prevalent chronic disease on individuals and healthcare systems worldwide.

F. References

[1] A. Rghioui, J. Lloret, S. Sendra, and A. Oumnad, “A smart architecture for diabetic patient monitoring using machine learning algorithms,” *Healthcare (Switzerland)*, vol. 8, no. 3, 2020, doi: 10.3390/healthcare8030348.

[2] J. S. K. Dalam No, S. Salih, and A. M. Abdulazeez, “Classification of Diabetic Retinopathy Images through Deep Learning Models-Color Channel Approach: A Review,” *Indonesian Journal of Computer Science Attribution*, vol. 13, no. 1, p. 450, 2024.

[3] V. Jaiswal, A. Negi, and T. Pal, “A review on current advances in machine learning based diabetes prediction,” *Primary Care Diabetes*, vol. 15, no. 3. Elsevier Ltd, pp. 435–443, Jun. 01, 2021. doi: 10.1016/j.pcd.2021.02.005.

[4] T. Chauhan, S. Rawat, S. Malik, and P. Singh, “Supervised and Unsupervised Machine Learning based Review on Diabetes Care,” in *2021 7th International Conference on Advanced Computing and Communication Systems, ICACCS 2021*, Institute of Electrical and Electronics Engineers Inc., Mar. 2021, pp. 581–585. doi: 10.1109/ICACCS51430.2021.9442021.

[5] N. Sethi, ^ Shiva, S. Reddy, and R. Rajender, “A COMPREHENSIVE ANALYSIS OF MACHINE LEARNING TECHNIQUES FOR INCESSANT PREDICTION OF DIABETES MELLITUS,” *International Journal of Grid and Distributed Computing*, vol. 13, no. 1, pp. 1–22, 2020, doi: 10.33832/ijgdc.2020.13.1.01.

[6] H. M. Deberneh and I. Kim, “Prediction of type 2 diabetes based on machine learning algorithm,” *Int J Environ Res Public Health*, vol. 18, no. 6, Mar. 2021, doi: 10.3390/ijerph18063317.

[7] J. Chaki, S. Thillai Ganesh, S. K. Cidham, and S. Ananda Theertan, “Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6. King Saud bin Abdulaziz University, pp. 3204–3225, Jun. 01, 2022. doi: 10.1016/j.jksuci.2020.06.013.

[8] M. Rout and A. Kaur, *Prediction of Diabetes Risk based on Machine Learning Techniques*. 2020 International Conference on Intelligent Engineering and Management (ICIEM) 246 978-1-7281-4097-1/20/$31.00 ©2020 IEEE Prediction of Diabetes Risk based on Machine Learning Technique, 2020.

[9] T. Sharma and M. Shah, "A comprehensive review of machine learning techniques on diabetes detection," *Visual Computing for Industry, Biomedicine, and Art*, vol. 4, no. 1. Springer, Dec. 01, 2021. doi: 10.1186/s42492-021-00097-7.

[10] R. Krishnamoorthi *et al.*, "Retracted: A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques," *Journal of healthcare engineering*, vol. 2023. NLM (Medline), p. 9872970, 2023. doi: 10.1155/2023/9872970.

[11] S. Shafi and G. Ahmad Ansari, "Early Prediction of Diabetes Disease & Classification of Algorithms Using Machine Learning Approach." [Online]. Available: https://ssrn.com/abstract=3852590

[12] Z. Arif Ali, Z. H. Abduljabbar, H. A. Tahir, A. Bibo Sallow, and S. M. Almufti, "Exploring the Power of eXtreme Gradient Boosting Algorithm in Machine Learning: a Review," *Academic Journal of Nawroz University*, vol. 12, no. 2, pp. 320–334, May 2023, doi: 10.25007/ajnu.v12n2a1612.

[13] K. Roy *et al.*, "An Enhanced Machine Learning Framework for Type 2 Diabetes Classification Using Imbalanced Data with Missing Values," *Complexity*, vol. 2021, 2021, doi: 10.1155/2021/9953314.

[14] A. M. Abdulazeez, D. Zeebaree, D. M. Abdulqader, and D. Q. Zeebaree, "Machine Learning Supervised Algorithms of Gene Selection: A Review," 2020. [Online]. Available: https://www.researchgate.net/publication/341119469

[15] J. T. C, *A Study on Various Machine Learning Classification Algorithms for Diabetes Prediction*.

[16] S. ,Oluwafemi Abe, O. O. Obe, O. K. Boyinbode, and O. N. Biodun, "Classifier Algorithms and Ensemble Models for Diabetes Mellitus Prediction: A Review," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 10, no. 1, pp. 430–439, Feb. 2021, doi: 10.30534/ijatcse/2021/641012021.

[17] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning," *Decision Analytics Journal*, vol. 3, p. 100071, Jun. 2022, doi: 10.1016/j.dajour.2022.100071.

[18] I. Ibrahim and A. M. Abdulazeez, "The Role of Machine Learning Algorithms for Diagnosing Diseases," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 10–19, Mar. 2021, doi: 10.38094/jastt20179.

[19] G. T. Reddy *et al.*, "An Ensemble based Machine Learning model for Diabetic Retinopathy Classification," in *International Conference on Emerging Trends in Information Technology and Engineering, ic-ETITE 2020*, Institute of Electrical and Electronics Engineers Inc., Feb. 2020. doi: 10.1109/ic-ETITE47903.2020.235.

[20] M. O. Edeh *et al.*, "A Classification Algorithm-Based Hybrid Diabetes Prediction Model," *Front Public Health*, vol. 10, Mar. 2022, doi: 10.3389/fpubh.2022.829519.
[21] B. Charbuty and A. M. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.
[22] G. T. Reddy *et al.*, "An Ensemble based Machine Learning model for Diabetic Retinopathy Classification," in *International Conference on Emerging Trends in Information Technology and Engineering, ic-ETITE 2020*, Institute of Electrical and Electronics Engineers Inc., Feb. 2020. doi: 10.1109/ic-ETITE47903.2020.235.
[23] M. Méndez, M. G. Merayo, and M. Núñez, "Machine learning algorithms to forecast air quality: a survey," *Artif Intell Rev*, vol. 56, no. 9, pp. 10031–10066, Sep. 2023, doi: 10.1007/s10462-023-10424-4.
[24] A. Shrivastava, M. Chakkaravarthy, and M. A. Shah, "A new machine learning method for predicting systolic and diastolic blood pressure using clinical characteristics," *Healthcare Analytics*, vol. 4, Dec. 2023, doi: 10.1016/j.health.2023.100219.
[25] L. Ismail, H. Materwala, M. Tayefi, P. Ngo, and A. P. Karduck, "Type 2 Diabetes with Artificial Intelligence Machine Learning: Methods and Evaluation," *Archives of Computational Methods in Engineering*, vol. 29, no. 1, pp. 313–333, Jan. 2022, doi: 10.1007/s11831-021-09582-x.
[26] M. Zounemat-Kermani, O. Batelaan, M. Fadaee, and R. Hinkelmann, "Ensemble machine learning paradigms in hydrology: A review," *Journal of Hydrology*, vol. 598. Elsevier B.V., Jul. 01, 2021. doi: 10.1016/j.jhydrol.2021.126266.
[27] A. M. Alajlan, "A Model-Based Approach for an Early Diabetes Prediction Using Machine Learning Algorithms," 2021.
[28] M. M. Chowdhury, R. S. Ayon, and M. S. Hossain, "An investigation of machine learning algorithms and data augmentation techniques for diabetes diagnosis using class imbalanced BRFSS dataset," *Healthcare Analytics*, vol. 5, Jun. 2024, doi: 10.1016/j.health.2023.100297.
[29] J. Yoon, "Forecasting of Real GDP Growth Using Machine Learning Models: Gradient Boosting and Random Forest Approach," *Comput Econ*, vol. 57, no. 1, pp. 247–265, Jan. 2021, doi: 10.1007/s10614-020-10054-w.
[30] S. F. Khorshid and A. M. Abdulazeez, "BREAST CANCER DIAGNOSIS BASED ON K-NEAREST NEIGHBORS: A REVIEW."
[31] Z. Mushtaq, M. F. Ramzan, S. Ali, S. Baseer, A. Samad, and M. Husnain, "Voting Classification-Based Diabetes Mellitus Prediction Using Hypertuned Machine-Learning Techniques," *Mobile Information Systems*, vol. 2022, 2022, doi: 10.1155/2022/6521532.
[32] T. Saranya, S. Sridevi, C. Deisy, T. D. Chung, and M. K. A. A. Khan, "Performance Analysis of Machine Learning Algorithms in Intrusion Detection System: A Review," in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 1251–1260. doi: 10.1016/j.procs.2020.04.133.

[33] A. Arista, “Comparison Decision Tree and Logistic Regression Machine Learning Classification Algorithms to determine Covid-19,” *Sinkron*, vol. 7, no. 1, pp. 59–65, Jan. 2022, doi: 10.33395/sinkron.v7i1.11243.
[34] R. Pradhan, M. Aggarwal, D. Maheshwari, A. Chaturvedi, and D. K. Sharma, “Diabetes Mellitus Prediction and Classifier Comparitive Study,” in *2020 International Conference on Power Electronics and IoT Applications in Renewable Energy and its Control, PARC 2020*, Institute of Electrical and Electronics Engineers Inc., Feb. 2020, pp. 133–139. doi: 10.1109/PARC49193.2020.236572.
[35] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, “Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms,” *Neural Comput Appl*, vol. 35, no. 22, pp. 16157–16173, Aug. 2023, doi: 10.1007/s00521-022-07049-z.
[36] Y. Bahri Özçelik Bülent Ecevit Üniversitesi and A. Altan Bülent Ecevit Üniversitesi, “Classification of diabetic retinopathy by machine learning algorithm using entropy-based features,” 2023. [Online]. Available: https://www.researchgate.net/publication/370254891
[37] M. Phongying and S. Hiriote, “Diabetes Classification Using Machine Learning Techniques,” *Computation*, vol. 11, no. 5, May 2023, doi: 10.3390/computation11050096.
[38] U. Ahmed *et al.*, “Prediction of Diabetes Empowered With Fused Machine Learning,” *IEEE Access*, vol. 10, pp. 8529–8538, 2022, doi: 10.1109/ACCESS.2022.3142097.
[39] S. Sadeghi, D. Khalili, A. Ramezankhani, M. A. Mansournia, and M. Parsaeian, “Diabetes mellitus risk prediction in the presence of class imbalance using flexible machine learning methods,” *BMC Med Inform Decis Mak*, vol. 22, no. 1, Dec. 2022, doi: 10.1186/s12911-022-01775-z.
[40] Z. Mushtaq, M. F. Ramzan, S. Ali, S. Baseer, A. Samad, and M. Husnain, “Voting Classification-Based Diabetes Mellitus Prediction Using Hypertuned Machine-Learning Techniques,” *Mobile Information Systems*, vol. 2022, 2022, doi: 10.1155/2022/6521532.
[41] S. Sivaranjani, S. Ananya, J. Aravinth, and R. Karthika, “Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction,” in *2021 7th International Conference on Advanced Computing and Communication Systems, ICACCS 2021*, Institute of Electrical and Electronics Engineers Inc., Mar. 2021, pp. 141–146. doi: 10.1109/ICACCS51430.2021.9441935.
[42] K. M. Aamir, L. Sarfraz, M. Ramzan, M. Bilal, J. Shafi, and M. Attique, “A fuzzy rule-based system for classification of diabetes,” *Sensors*, vol. 21, no. 23, Dec. 2021, doi: 10.3390/s21238095.
[43] S. Aftab, S. Alanazi, M. Ahmad, M. A. Khan, A. Fatima, and N. S. Elmitwally, “Cloud-Based Diabetes Decision Support System Using Machine Learning Fusion,” *Computers, Materials and Continua*, vol. 68, no. 1, pp. 1341–1357, Mar. 2021, doi: 10.32604/cmc.2021.016814.
[44] N. Ahmed *et al.*, “Machine learning based diabetes prediction and development of smart web application,” *International Journal of Cognitive*

Computing in Engineering, vol. 2, pp. 229–241, Jun. 2021, doi: 10.1016/j.ijcce.2021.12.001.

[45] C. Azad, B. Bhushan, R. Sharma, A. Shankar, K. K. Singh, and A. Khamparia, "Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus," in *Multimedia Systems*, Springer Science and Business Media Deutschland GmbH, Aug. 2022, pp. 1289–1307. doi: 10.1007/s00530-021-00817-2.

[46] A. Bansal and A. Singhrova, "Performance Analysis of Supervised Machine Learning Algorithms for Diabetes and Breast Cancer Dataset," in *Proceedings - International Conference on Artificial Intelligence and Smart Systems, ICAIS 2021*, Institute of Electrical and Electronics Engineers Inc., Mar. 2021, pp. 137–143. doi: 10.1109/ICAIS50930.2021.9396043.

[47] S. Gayathri, V. P. Gopi, and P. Palanisamy, "Diabetic retinopathy classification based on multipath CNN and machine learning classifiers," *Phys Eng Sci Med*, vol. 44, no. 3, pp. 639–653, Sep. 2021, doi: 10.1007/s13246-021-01012-3.

[48] P. Ghosh, S. Azam, A. Karim, M. Hassan, K. Roy, and M. Jonkman, "A comparative study of different machine learning tools in detecting diabetes," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 467–477. doi: 10.1016/j.procs.2021.08.048.

[49] Y. Jian, M. Pasquier, A. Sagahyroon, and F. Aloul, "A machine learning approach to predicting diabetes complications," *Healthcare (Switzerland)*, vol. 9, no. 12, Dec. 2021, doi: 10.3390/healthcare9121712.

[50] F. Alaa Khaleel and A. M. Al-Bakry, "Diagnosis of diabetes using machine learning algorithms," *Mater Today Proc*, vol. 80, pp. 3200–3203, Jan. 2023, doi: 10.1016/j.matpr.2021.07.196.

[51] M. M. Nishat, "Performance Assessment of Different Machine Learning Algorithms in Predicting Diabetes Mellitus," *Biosci Biotechnol Res Commun*, vol. 14, no. 1, pp. 74–82, Mar. 2021, doi: 10.21786/bbrc/14.1/10.

[52] J. J. Khanam and S. Y. Foo, "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.

[53] M. W. Nadeem, H. G. Goh, V. Ponnusamy, I. Andonovic, M. A. Khan, and M. Hussain, "A fusion-based machine learning approach for the prediction of the onset of diabetes," *Healthcare (Switzerland)*, vol. 9, no. 10, Oct. 2021, doi: 10.3390/healthcare9101393.

[54] S. NAHZAT and M. YAĞANOĞLU, "Makine Öğrenimi Sınıflandırma Algoritmalarını Kullanarak Diyabet Tahmini," *European Journal of Science and Technology*, Apr. 2021, doi: 10.31590/ejosat.899716.

[55] K. Alpan and G. S. Ilgi, "Classification of Diabetes Dataset with Data Mining Techniques by Using WEKA Approach," in *4th International Symposium on Multidisciplinary Studies and Innovative Technologies, ISMSIT 2020 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., Oct. 2020. doi: 10.1109/ISMSIT50672.2020.9254720.

[56] T. A. Assegie and S. Nair, "The Performance Of Different Machine Learning Models On Diabetes Prediction," *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, vol. 9, p. 1, 2020, [Online]. Available: www.ijstr.org

[57] O. Daanouni, B. Cherradi, and A. Tmiri, “Diabetes Diseases Prediction Using Supervised Machine Learning and Neighbourhood Components Analysis,” in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Mar. 2020. doi: 10.1145/3386723.3387887.
[58] 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.
[59] R. Katarya and S. Jain, “Comparison of different machine learning models for diabetes detection,” in *Proceedings of 2020 IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering, ICADEE 2020*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020. doi: 10.1109/ICADEE51157.2020.9368899.
[60] G. A. Pethunachiyar, “Classification Of Diabetes Patients Using Kernel Based Support Vector Machines,” in *2020 IEEE Recent Advances in Intelligent Computational Systems, RAICS*, Institute of Electrical and Electronics Engineers Inc., Jun. 2020, pp. 122–127. doi: 10.1109/RAICS.2015.7488400.
[61] M. Soni and S. Varma, “Diabetes Prediction using Machine Learning Techniques,” 2020. [Online]. Available: www.ijert.org
[62] N. P. Tigga and S. Garg, “Prediction of Type 2 Diabetes using Machine Learning Classification Methods,” in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 706–716. doi: 10.1016/j.procs.2020.03.336.
[63] G. Tripathi and R. Kumar, *Early Prediction of Diabetes Mellitus Using Machine Learning*. 2020, pp. 1–6.
[64] J. Xue, F. Min, and F. Ma, “Research on diabetes prediction method based on machine learning,” in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Nov. 2020. doi: 10.1088/1742-6596/1684/1/012062.