Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 19 No. 15 (2023) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v19i15.42417

PAPER

Early Diagnosis of Diabetes: A Comparison of Machine Learning Methods

Mowafaq Salem Alzboon[1](✉), Mohammad Subhi Al-Batah[1], Muhyeeddin Alqaraleh[1], Ahmad Abuashour[2], Ahmad Fuad Hamadah Bader[3]

[1]Faculty of Science and Information Technology, Jadara University, Irbid, Jordan

[2]Faculty of Computer Studies, Arab Open University, Al-Ardiya, Kuwait

[3]Faculty of Engineering, Jadara University, Irbid, Jordan

malzboon@jadara.edu.jo

ABSTRACT

Detection and management of diabetes at an early stage is essential since it is rapidly becoming a global health crisis in many countries. Predictions of diabetes using machine learning algorithms have been promising. In this work, we use data collected from the Pima Indians to assess the performance of multiple machine-learning approaches to diabetes prediction. Ages, body mass indexes, and glucose levels for 768 patients are included in the data set. The methods evaluated are Logistic Regression, Decision Tree, Random Forest, k-Nearest Neighbors, Naive Bayes, Support Vector Machine, Gradient Boosting, and Neural Network. The findings indicate that the Logistic Regression and Neural Network models perform the best on most criteria when considering all classes together. The SVM, Random Forest, and Naive Bayes models also receive moderate to high scores, suggesting their strength as classification models. However, the kNN and Tree models show poorer scores on most criteria across all classes, making them less favorable choices for this dataset. The SGD, AdaBoost, and CN2 rule inducer models perform the poorest when comparing all models using a weighted average of class scores. The results of the study suggest that machine learning algorithms may help predict the onset of diabetes and for detecting the disease at an early stage.

KEYWORDS

diabetes, decision trees, machine learning, diagnosis, support vector

1 INTRODUCTION

Millions of people worldwide have diabetes, a chronic metabolic disease that is a leading cause of illness and mortality [1]. Complications of diabetes include high blood sugar, which can lead to heart disease, stroke, blindness, and even amputations. It is crucial to diagnose and treat diabetes as soon as possible to reduce the risk of complications, but this can be difficult because the disease often presents with vague or nonexistent symptoms [2]. To produce inferences or predictions from data without being explicitly programmed is the goal of machine learning (ML), a branch of artificial intelligence. Algorithms based on machine learning have several

Alzboon, M.S., Al-Batah, M.S., Alqaraleh, M., Abuashour, A., Hamadah Bader, A.F. (2023). Early Diagnosis of Diabetes: A Comparison of Machine Learning Methods. *International Journal of Online and Biomedical Engineering (iJOE)*, 19(15), pp. 144–165. https://doi.org/10.3991/ijoe.v19i15.42417

Article submitted 2023-06-19. Revision uploaded 2023-08-05. Final acceptance 2023-08-13.

applications, one being healthcare. The likelihood of a patient acquiring diabetes may be predicted from patient data using ML algorithms [3]. Motivation: Finding the best machine learning algorithm for diabetes prediction is the goal of this research. Diagnosis and treatment of diabetes at an early stage can considerably improve symptoms and minimize the severity of the disease. Prediction algorithms for early diabetes detection are an important field of research [4]. The need for more research: This research aims to answer the question, "Which ML algorithm is best for early diabetes prediction?" This research aims to determine the best method for early diabetes prediction by comparing and contrasting many machine learning (ML) algorithms. Objectives: The primary goal of this work is to compare the performance of several machine learning algorithms for early diabetes prediction using the National Health and Nutrition Examination Survey (NHANES) dataset. The following are the precise goals.

Modelling the onset and progression of diabetes using machine learning techniques such as logistic regression, decision trees, random forests, support vector machines, and neural networks [5]. To compare the efficacy of various ML techniques by measuring their F1-score, accuracy, precision, and recall. This study aims to determine the most effective machine learning approach for early diagnosis of diabetes. Those interested in demographic, clinical, and laboratory data from a sample of adults aged 20 and above who participated in the NHANES between 1999 and 2016 can download the NHANES dataset for free. Clinical measures (body mass index, blood pressure), laboratory tests (glucose, cholesterol levels), and demographic information (age, gender, race) are all included in the dataset [6]. Early intervention and better patient outcomes are possible thanks to the study's findings, which can aid in the development of more accurate and efficient prediction models for diabetes diagnosis. The results may also add to the expanding body of knowledge about the usefulness of ML algorithms in medical settings [7].

2 LITERATURE REVIEW

Using publically available physiological data, including age, gender, weight, height, and short (2.1s) Photoplethysmography (PPG) signals from intelligent devices, the previous study [7] attempts to predict the development of Type 2 Diabetes. It was shown that Type 2 Diabetes might be expected from relatively short PPG signals by analyzing morphological features of the PPG waveform and its derivatives. The area under the ROC curve (AUC) is most remarkable for linear discriminant analysis (LDA) (79 percent). The actual implementation of the proposed approach would enable individuals to quickly screen themselves using their smart devices to identify the risk of Type 2 Diabetes and avoid the challenges of late detection [7]. Diabetes is a long-term metabolic disorder characterized by insulin resistance and elevated blood sugar levels. Hyperglycemia describes Types 1 and 2, whereas Alzheimer's disease defines Type 3. Due to the progressive nature of diabetes, its prognosis is crucial. Machine learning classification algorithms may be used to forecast cases of diabetes. Classification models with a conclusion target vector may be fitted to insulin, blood pressure, skin thickness, and glucose levels data. For neural networks, the same is true [8]. To predict who would develop Type 2 diabetes in its early stages, the previous study [9] uses direct questionnaires (DM). Utilizing the information gain method, they constructed models using logistic regression, support vector machine, K-nearest neighbor, Nave Bayes, random forest, and neural networks to detect diabetes in its earliest stages. When compared to other machine learning methods, RF's

accuracy of 100% was unmatched. Based on these results, it is plausible that an easy-to-use questionnaire coupled with a machine-learning algorithm might be used to identify people with undiagnosed DM [9] accurately.

Diabetes mellitus is one of the world's leading causes of disability and premature mortality. The authors in [10] utilized a logistic regression model and a decision tree, a machine learning technique, to predict type 2 diabetes in Pima Indian women and better understand risk variables. Their study found that glucose, pregnancy, BMI, diabetes pedigree function, and age were the five most important predictors of type 2 diabetes. The preferred specification yields a cross-validation error rate of just 21.74 percent and a prediction accuracy of 78.26 percent. Their strategy may be used to reduce diabetes rates and costs in conjunction with other preventative measures [10]. Diabetes is a fatal illness, although early diagnosis can decrease its effects. The previous study presents a methodology for improving diabetes forecasting using data called Diabetes Expert System Using Machine Learning Analytics (DESMLA). The model uses SMOTE, Borderline SMOTE, ADASYN, KMeans SMOTE, Gaussian SMOTE, Decision Tree (DT), and Random Forest as classifiers, all oversampling approaches (RF). The trials validate the superior performance of the DESMLA model combined with KMeans SMOTE and Gaussian SMOTE [11]. Artificial intelligence has the potential to revolutionize diabetes diagnosis and treatment. The diagnosis of diabetes mellitus was predicted using a total of six supervised machine-learning methods [12].

The Random Forest classification algorithm outperformed previous state-of-the-art techniques with a 92% accuracy rate, achieved by combining many methods for handling missing information. Using the Pima diabetes data, this strategy beats prior studies [12]. There are currently 537 million individuals affected by diabetes across the world, making it the most lethal and widespread non-communicable disease. Several machine learning techniques were used on a private dataset of female patients in Bangladesh to develop an autonomous diabetes prediction system. Researchers examined data on diabetes among Pima Indians and collected samples from 203 workers at a nearby Bangladesh textile factory. The insulin features of the private dataset were predicted using a semi-supervised model with heavy gradient boosting. The SMOTE and ADASYN methods [13] were used to the problem of class disproportion. With an accuracy of 81%, an F1 coefficient of 0.81, and an area under the curve (AUC) of 0.84, the suggested system outperformed the XGBoost classifier using the ADASYN method. The flexibility of the proposed system was further shown by including the domain adaption technique. At long last, a web-based framework and an Android mobile app have been created to take in some factors and generate an immediate diabetes prediction [14].

Diabetes is a devastating metabolic disease that manifests itself in several ways. Toxic or chemical substances, obesity, the work culture, poor nutrition, an atypical diet, unusual eating habits, and environmental factors all contribute to its rapid development. Using Machine Learning Methods, the researchers can build a better healthcare system that can foresee complications with diabetes. The previous study article [15] uses Machine Learning Techniques in a Diet Recommendation System to detect diabetes and provide dietary guidance for those with the disease (DRS). Patients with diabetes can benefit from data analysis when deciding on the best diet [15]. The method used in that study might be used to diagnose diabetes. The dataset utilized in that study [15] was obtained from the UCI machine learning repository, and it contained data on 768 patients, each characterized by eight numbers. The best features were selected using a genetic algorithm (GA), and k-fold cross-validation was used to partition the dataset. Both the chosen datasets and the baseline dataset were analyzed by utilizing GA with several classifiers, including K-nearest neighbor

(KNN), Multilayer Perceptron (MLP), Deep Neural Network (DNN), and Naive Bayes (NB) (8 features). The relative accuracy of these classifiers was compared. KNN was about 93.33 percent accurate, DNN was approximately 77.27 percent accurate, MLP was roughly 74.92 percent accurate, and NB was 74.89 percent correct [16].

Millions of people worldwide are struggling with diabetes, a devastating disease. Scientists are urged to work on a Machine Learning method for diabetes forecasting. Researchers in the previous study [17] compared many MLAs for early diabetes risk assessment. The experimental research successfully implemented six MLAs, with RF as the most trustworthy classifier (with a 98 percent success rate). The study's findings provide a solid foundation for estimating diabetes's prevalence and foreseeing its development [17]. Untreated diabetes can cause damage to several body systems over time. Predicting the onset of sickness early can help save lives and give medical professionals more time to treat patients. Ensemble learning is a method of data analysis that combines many ways into one superior prediction model. The UCI repository was mined for diabetes data, and prediction models, including AdaBoost, Bagging, and Random Forest, were used to make predictions. The Random Forest Ensemble Method (97%) has better accuracy, precision, recall, and F1 scores than AdaBoost and Bagging [18]. The authors in [19] aimed to calculate an estimated duration of stay for diabetes patients admitted to the intensive care unit by applying machine learning algorithms to their clinical data from the first 8 hours of admission. The time spent in the intensive care unit (ICU) and whether or not an ICU stay is considered lengthy or short based on a 10-day threshold were studied as prediction tasks. The number of days spent in the intensive care unit could be predicted most accurately by the neural network model, with an R2 of 0.3969 and a mean absolute error of 1.94 days. The gradient boosting model successfully differentiated between long and short ICU stays [19] with an accuracy of 0.8214.

Using information from the free medical examination service program for those over 65, they built machine-learning models to predict the chance of incident diabetes. The average annual rate of increase to diabetes in prediabetic older individuals was 14.21%. Each model was trained with data from 9607 prediabetic adults on eight attributes and one outcome variable and then tested on 2402 prediabetes patients. XGBoost was a successful model (ROC: 0.6742 for 2019 and 0.6707 for 2020). Although the four models yielded comparable results, the XGBoost model had a high ROC value and showed promise for further study [20]. It's no secret that Diabetes Mellitus (DM) has far-reaching consequences for individuals, societies, and states. Due to its high diabetes incidence, Saudi Arabia is among the world's top ten countries. If it were possible to predict a patient's diabetic state using only a few indicators, widespread, rapid, and inexpensive diabetes screening would be possible. The authors in [21] investigate using HbA1c and FPG as input features for diabetes patient prediction. Using five separate machine learning classifiers, feature removal through feature permutation, and hierarchical clustering, they achieved good accuracy, precision, recall, and F1-score of the models on the dataset. Risk factors and their indirect effects on diabetes classification were identified through the data analysis. Their results jived with those from the American Diabetes Association (ADA) and other international health agencies listing risk factors for diabetes and prediabetes. They conclude that critical elements particular to the Saudi population may be identified by analysis of the illness, and their management can result in disease control [21].

The researcher in [22] aimed to improve the accuracy of diabetes prediction using machine learning and preprocessing techniques. The Pima Indian Diabetes dataset was classified using J48, Nave Bayes, Support Vector Machine, Logistic Regression,

Multilayer Perceptron, K Nearest Neighbor, Logistic Model Tree, and Random Forest. The preprocessing techniques featured selection, missing value imputing, normalization, and standardization. With an accuracy score 80.869 [1], Random Forest algorithm emerged as the clear victor. High blood sugar levels are a hallmark of diabetes mellitus, a chronic illness. Automating disease forecasting could be possible with the use of data mining tools. Two different data mining classification methods are used in the hybrid classifier model used by the proposed data analytics system. Patients who tested positive were separated into groups—1 and t—using a Multilayer Perceptron Neural Network designed and trained using the Back Propagation Algorithm. The trained neural network's identification rate obtained in tests was 80%, and the mean square error was 0.1213 [22]. The previous study [23–24] aimed to use machine learning and natural language processing (NLP) techniques and resources from the Unified Medical Language System (UMLS) to estimate the probability of mortality in patients with diabetes in the critical care scenario. Several machine learning modelling and natural language processing (NLP) approaches were employed in a secondary analysis of Medical Information Mart for Intensive Care III (MIMIC-III) data. The definitions of clinical terms defined by domain experts form the basis of healthcare domain knowledge. From conceptual entities and their connections, knowledge-guided models may automatically extract knowledge from clinical notes or biological literature. Mortality was classified using a matrix of characteristics and indicators informed by existing information. They trained a convolutional neural network on word embeddings using the UMLS entity embedding (CNN). An AUC of 0.97 was achieved by strategically placing the machine learning models. Predicting death in diabetic patients in a critical care setting using UMLS resources and clinical notes is a practical and valuable approach [23–24].

The previous article [25] compares traditional categorization methods with neural network-based machine learning on the diabetes dataset. The techniques evaluated are naive Bayes, K-nearest neighbor, extra trees, decision trees, radial basis function, and multilayer perceptron. Results show that the multilayer perceptron approach has the best area under the curve (86%) and the lowest false positive and false negative rates (MSE = 0.19) for making predictions. Clinical researchers use predictive modelling methods to establish baseline health status and describe change patterns. The Hidden Markov Model (HMM) and its variants are a class of forecasting models. When using unequally sampled longitudinal Electronic Medical Records (EMR) data, Newton's Divide Difference Method (NDDM) experiences Runge Phenomenon. An innovative approximation approach based on NDDM as a component with HMM was provided to estimate a person's 8-year risk of developing Type 2 Diabetes Mellitus (T2DM). The results showed that the proposed method has the potential to successfully approximate and improve prediction accuracy utilizing the already gathered EMR data on an ad hoc basis [25]. To create a more precise prediction model, that research utilized machine learning techniques on the Pima Indians diabetes dataset (PIDD). The results showed that glucose, insulin, and body mass index were more strongly associated with diabetes. Data standardization allowed the support vector machine (SVM) to achieve 85.06 percent accuracy, the highest of all tested methods. This SVM achieves the highest accuracy (87.01%) after adjustment for predicting a diabetes diagnosis [26].

Dyslipidemia, neuropathy, nephropathy, diabetic foot, hypertension, obesity, and retinopathy are only a few severe complications that can arise from diabetes mellitus (DM). Algorithms for predicting and diagnosing eight complications of diabetes were developed using data from the Rashid Centre for Diabetic and Research (RCDR). Several strategies for handling missing values and skewed data were evaluated via

preprocessing steps [27]. Artificial neural networks, decision trees, random forests, naive Bayes, K-nearest neighbors, support vector machines, and logistic regression are some methods for detecting diabetes investigated in that study. Artificial neural networks, decision trees, random forests, naive Bayes, K-nearest neighbors, support vector machines, and logistic regression are some machine learning methods that may be trained with the Pima Indian diabetic dataset. The consequences, benefits, and drawbacks are discussed in detail [28]. The previous study [29] examined how a Hidden Markov Model (HMM) would affect the accuracy of the well-known Framingham Diabetes Risk Scoring Model (FDRSM). To determine an individual's 8-year risk of developing diabetes, researchers analyzed electronic medical record data from 172,168 primary care patients using HMM. Their study found an AROC of 86.9% in a sample size of 911 persons, which is on par with the AROCs of 78.6% and 85.2% found in previous validation studies of FDRSM in the same Canadian and Framingham populations, respectively. Their suggested HMM outperforms the FDRSM validation study in differentiating between the Canadian and Framingham populations [29]. There were benefits and drawbacks to the abundance of data created by healthcare organizations. AI systems can help with this using medical records, genome-omics data, imaging scans, and wearable devices [4].

Doctors, patients, and their families may all reap the rewards of early diabetes diagnosis enabled by machine learning algorithms. Shankar used neural networks to forecast the onset of diabetes mellitus using the Pima Indian diabetes dataset, proving the efficacy of his method. To detect the onset of diabetes in diabetic patients, a study [30] evaluates and contrasts several machine-learning classification methods. High blood sugar levels are the root cause of the devastating disease known as diabetes. It has been the subject of computer-based detection systems for identification and evaluation, but with the development of machine learning, they can finally build a remedy. We've built an infrastructure that can tell if a person has diabetes. To create an Interactive Web Application for Diabetes Prediction [31], they used the Pima Indian dataset, which is 82.35 percent accurate. Annually, between 2 and 5 million people are diagnosed with Type 2 Diabetes, also known as Non-Insulin Dependent Diabetes Mellitus. Predicting the onset of diabetes and associated complications like cardiovascular and kidney disease can help keep people healthier. The Pima Indian Diabetes data collection (PID) is a widely utilized resource retrieved from the UCI repository. Predictions of Type II diabetes mellitus were made using KNN, Logistic Regression, SVM, Random Forest, LightGBM, and XGBoost ML models. For an 80–20 train test split, the lightGBM model's highest accuracy was 91.47 percent [32]. Predictions of T2DM are made using some different machine learning methods in that investigation. The classifiers used a range from logistic regression and XGBoost to gradient boosting and decision trees to ExtraTrees and random forests, with even lighter gradient boosting machines being used (LGBM). The LGBM classifier outperforms the competition with an accuracy of 95.20 percent [5].

Increased sodium-glucose cotransporter two inhibitor canagliflozin usage in diabetic patients is driven by the drug's beneficial effects on cardiovascular and renal outcomes. According to clinical trials, canagliflozin has been linked to increased amputations performed on the lower extremities (LEA). To calculate the potential for LEA in canagliflozin-treated diabetics, they turned to machine learning methods. The results showed that over a median follow-up period of 1.5 years, the incidence rate of LEA was 0.57 percent. Among the 16 factors examined, a previous diagnosis of LEA and the use of loop diuretics were shown to have the strongest associations with the development of LEA. The risk of LEA in canagliflozin-treated diabetics was predicted correctly by their machine learning method. Patients with diabetes

may benefit from the risk score in making treatment decisions [33]. That investigation utilizes machine learning strategies to provide a preliminary diagnosis of diabetes in women. That can potentially halt the spread of illness and reduce the likelihood of disastrous outcomes. According to the random forest classifier, the model's accuracy was 82% [34]. The study [35] aims to develop a machine learning (ML) method for accurate diabetes risk assessments. Methods logistic regression (LR) uses the p-value and the odds ratio to identify risk factors for developing diabetes (OR). Nave Bayes, Decision Tree, Adaboost, and Random Forest are the four classifiers for predicting diabetes (RF). These classifiers' efficacy is measured by their accuracy (ACC) and area under the curve (AUC) (AUC). Age, education, body mass index, systolic blood pressure, diastolic blood pressure, direct cholesterol, and total cholesterol are seven risk variables for diabetes identified by the LR model. The overall ACC for a system built on ML is 90.62 percent. LR-based feature selection and RF-based classifier for the K10 protocol achieve 94.25 percent ACC and 0.95 AUC [35]. Type 2 diabetes significantly contributes to several serious health problems, including vision loss, kidney failure, heart attack, stroke, and limb loss. Traditional risk classification techniques overlook socio-demographic factors, self-management skills, and healthcare accessibility. The research [36] sought to develop and validate a machine learning-based method for identifying T2DM patients at high risk of clinical deterioration using a comprehensive collection of patient-level indicators collected from a population health-linked dataset. Retinopathy, chronic renal disease, myocardial infarction, stroke, peripheral artery disease, and mortality are all possible results of clinical deterioration. The ability of patients to self-manage their condition, clinical and metabolic signs, and the use of healthcare services will all serve as predictors. Predictive models will be defined using multi-dependence Bayesian networks. The results may be accessed by people with T2DM, their caregivers, funders, diabetic care organizations, and other researchers [36].

The previous study [37] was a retrospective observational study of 22,242 singleton pregnancies at a tertiary maternity hospital in China from 2013.1.1 to 2017.12.31, and it aimed to determine how well machine learning algorithms performed in predicting gestational diabetes mellitus (GDM). Eight popular machine learning algorithms (GDBT, AdaBoost, LGB, Logistic, Vote, XGB, Decision Tree, and Random Forest) and two popular regressions (stepwise logistic regression and logistic regression with RCS) were used to make predictions about the prevalence of GDM. The GBDT model performed better than the other machine learning methods (AUC 0.74, 95% CI 0.71–0.76). Fasting glucose, HbA1c, lipids, and body mass index significantly impacted GDM. The GBDT model's negative predictive value was 74.1% (95 percent CI: 69.5%–78.5%), and its sensitivity was 90% (95 percent CI: 88.0%–91.7%) when the cutoff value was 0.3. Insulin resistance and diabetes are two of the most common metabolic disorders. In that work, they applied machine learning techniques to predict diabetes by extracting information from medical diagnostic datasets. Predictions of diabetes mellitus were commonly made using SVM, Naive Bayes, K-nearest neighbor, and C4.5 Decision Tree, all of which are examples of machine learning algorithms. Experimental results showed that C4.5 Decision Tree was more accurate than other machine learning algorithms [37]. Cardiovascular disease, stroke, neuropathy, and renal failure are among the potential outcomes of diabetes mellitus. To diagnose diabetes in its earliest stages, machine learning techniques were used to examine data from several sources and synthesize crucial knowledge. Using R-Studio and the Pima Indian database housed in the UCI repository [38], that group constructed a fantastic model that predicts and identifies diabetic sickness earlier. With 246 million sufferers and 3.8 million fatalities attributable to complications each year, diabetes

mellitus is a severe threat to public health worldwide. The work [39] aimed to use massive data platforms like Spark and distributed machine learning to develop a system that can predict diabetes. The results of the experiments showed that out of the five machine learning classification techniques used, LR achieved the highest percentages of accuracy (82%), recall (92%), and precision (82%), respectively [39]. The previous research [40] aims to develop a machine learning (ML) model that extrapolates data from the current year to forecast the prevalence of type 2 diabetes (T2D) in year Y+1. The data was collected from a medical organization's EHRs between 2013 and 2018, and 80,692 individuals' longitudinal data was used to train the ML model. The RF classifier achieved an accuracy of 73.3%, while the XGBoost classifier achieved an accuracy of 73.82%. The prediction of type 2 diabetes incidence in the year Y+1 was enhanced by factors such r-GTP, uric acid, triglycerides, and lifestyle variables.

Potential GDM predictors are included in a 19-week risk prediction model [41].

Classifying and predicting data is an essential part of data mining, and it is used in many fields to give context to readily available data and valid prediction results. In that paper [42], they tested and used a modified version of the extreme learning machine to differentiate between people with and without diabetes. It also discusses and compares the use of two standard machine learning algorithms as binary classifiers to deal with the diabetes prediction problem: the backpropagation neural network and the modified extreme learning machine. UCI's learning repository [42] provided the datasets used in that study. The suggested model combines two machine learning techniques, Support Vector Machine, and Random Forest, to make diabetes predictions. Random Forest outperformed Support Vector Machine [43] with 98% accuracy and 99% ROC. Managing diabetes requires an individualized, patient-specific approach to lowering risk factors. A risk engine is an analytical tool that gathers large amounts of population data to model the onset and progression of diabetes. Recently created data cohorts enable the development of adaptive and generalizable risk engines. The Building, Relating, Assessing, and Validating Outcomes in (BRAVO) diabetes model can accurately predict diabetes comorbidities globally since it was calibrated using data from international clinical trials. It may find use in areas such as risk assessment, intervention evaluation, and cost-effectiveness modelling in the long run [44]. E-health, m-Health, and other "smart healthcare" forms also aid in preventative medicine. In the research [45], they used supervised machine learning to train a model and unsupervised machine learning to classify a dataset for predicting diabetes risk at an early stage. When evaluating a new patient, they use the supervised machine learning algorithm with the highest accuracy rate for diagnosing diabetes. Using machine learning and a patient questionnaire, a web app is created to provide a rough assessment of the patient's risk for type 2 diabetes in the early stages of the disease. A deep learning approach further evaluates the prediction to improve accuracy.

3 METHODOLOGY

The information was initially collected in the United States by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Due to the high prevalence of diabetes among Pima Indians, the dataset has been widely used in diabetes studies [46]. The data was collected from Pima Indians in the Gila River region of Arizona. There are nine columns in the dataset, the first 8 of which detail patient demographics and clinical details, and the ninth is the outcome variable indicating

whether the patient has diabetes. Patient age, gender, and number of pregnancies are examples of demographic characteristics. Body mass index (BMI), blood pressure, skin thickness, insulin level, and glucose concentration are all examples of clinical features. A result of 1 implies diabetes, whereas a value of 0 suggests the patient does not have it. In machine learning models, the outcome variable is the dependent variable, while demographic and clinical factors are the independent variables. The diabetes among Pima Indians dataset is a one-of-a-kind resource for ML study. Several prediction models for diagnosing and treating diabetes early on have been built using this information. Machine learning models that may accurately predict whether a patient has diabetes are designed and evaluated using this dataset, and researchers examine the connection between demographic and clinical factors and the outcome variable. A thorough analysis of how the models were created and the parameters used in their construction for the machine-learning inquiry utilizing the diabetes dataset.

Acquiring Databases: This study's diabetes dataset was collected from a publicly available source. It comprises 768 rows representing individual patients and contains numerous diabetes-related data, such as glucose level, blood pressure, and BMI. Before the development of the models, the dataset was preprocessed to guarantee its quality. Outliers, which are severe or abnormal results, were found and eliminated so that they would not negatively affect the model's performance. In addition, missing values, if any, were handled with the appropriate methods, such as imputation or deletion. The data were standardized to promote fair comparisons and precise model training. Normalization guarantees that all variables are measured on the same scale, preventing one characteristic from dominating the learning process. Techniques such as min-max scaling and standardization were used to standardize the numeric properties of the dataset.

The dataset was divided into two subsets: a training set and a testing set, to examine the models' performance on unobserved data and evaluate their generalization capacity. The training set, which usually consists of 70 percent of the data, was used to train machine learning models, while the testing set, consisting of the remaining 30 percent of the data, was used to evaluate the performance of the models. Several machine learning algorithms, including logistic regression, k-nearest neighbors, decision trees, random forests, and support vector machines, were examined for the inquiry. These algorithms were selected due to their aptitude for classification tasks and their past effectiveness in diabetes prediction research.

Feature Selection: A subset of relevant features from the dataset were chosen to train the models effectively. The selection of these properties as model input variables was based on their potential predictive power and clinical significance. Domain knowledge and approaches such as correlation analysis or feature importance from ensemble models were utilized for feature selection. The chosen machine learning algorithms were trained on the training set using the selected features. The models learned the underlying patterns and correlations between the input characteristics and the diabetic outcome variable during the training procedure. Using approaches such as grid search and cross-validation, the algorithms' parameters were modified to maximize their performance.

Using a variety of evaluation indicators, the effectiveness of the trained models was evaluated. Standard classification task metrics include accuracy, sensitivity, specificity, precision, and the F1 score. Accuracy assesses the overall accuracy of the predictions, whilst sensitivity and specificity assess the model's ability to distinguish positive and negative examples correctly. Precision measures the fraction of accurately anticipated positive instances, whereas the F1 score combines precision and

recall into a single metric. Additional measures, such as the area under the curve (AUC) and the receiver operating characteristic (ROC) curve, were utilized to evaluate the performance of the models. The area under the receiver operating characteristic (ROC) curve depicts the trade-off between accurate positive and false favorable rates at various categorization criteria. Generalization and Statistical Analysis: They were evaluated on an unseen testing set to ensure the models' ability to generalize to new data. Statistical analysis tools, such as the t-test or analysis of variance (ANOVA), were used to assess the efficacy of various machine learning approaches and to establish whether there were statistically significant variations between their performance measures. Tables and graphs were used to illustrate the data to assist in interpreting and presenting the findings. These visualizations gave a clear and concise assessment of the performance of the models, making the results easier to comprehend and analyze.

In conclusion, the models in this machine-learning investigation utilizing the diabetes dataset were developed using a systematic procedure that included dataset acquisition, data preprocessing, normalization, data splitting, model selection, feature selection, model training, evaluation using various metrics, generalization testing, statistical analysis, and data visualization. These processes confirm the validity of the models and provide insight into their performance and diabetes prediction capabilities.

3.1 Collect and preprocess the dataset

Collect and preprocess the diabetes dataset from a publicly available source. The data will next be preprocessed to remove any outliers or missing values. Next, the information was normalized to measure all variables on the same scale. After that, we'll split the data set in two, using 70% for training and 30% for testing. This separation ensures the model is trained on a large enough dataset to reveal hidden trends while providing adequate testing grounds.

Finding the most important qualities for predicting the result variable is a primary goal of machine learning, and feature extraction and selection play a crucial role in this process. This investigation uses principal component analysis (PCA) and correlation analysis to extract features, whereas RFE and SelectKBest are employed for feature selection. These methods will help determine which diabetes prediction features are the most important to feed into machine learning models.

This analysis will compare popular machine learning methods, such as logistic regression, k-nearest neighbors, decision trees, random forests, and support vector machines. Previous studies using these algorithms for diabetes prediction have shown encouraging results. Grid search will be used to fine-tune the hyperparameters of the trained algorithms once they have been taught to use the training set's given features. This will help locate the sweet spot for each algorithm's hyperparameters and maximize their efficiency.

3.2 Evaluation and performance metrics for models

Accuracy, sensitivity, specificity, precision, and the F1 score are only a few metrics that will be used to evaluate the efficacy of the machine-learning approaches. Using these metrics, we can determine which diabetes prediction algorithms perform the best and refine our models accordingly. The area under the curve (AUC) and the

receiver operating characteristic (ROC) curve will be used to evaluate the model's efficacy. The AUC measures the model's overall performance, whereas the ROC curve compares the actual positive rate to the false positive rate at different cutoffs of the predicted probability.

This technique offers a thorough structure for evaluating several machine learning algorithms for diabetes diagnosis forecasting at an early stage. Early detection and management of diabetes can be improved by using the project's findings to inform decisions about which machine learning algorithms and methods are most effective for making such predictions.

4 EXPERIMENTAL RESULTS

4.1 Procedures for doing a machine learning study using the diabetes dataset instances

The collection comprises a total of 768 records. Each data instance represents a patient, complete with the demographic and clinical information linked with that patient. The dataset includes a total of nine distinct aspects or qualities. In total, the dataset contains nine different attributes. The details are as follows: The number of successful pregnancies carried to term by a patient is indicated by the first feature. The second feature reflects the amount of glucose in the blood. Thirdly, diastolic blood pressure will be discussed in this section. The depth of the skin folds in the triceps is the fourth distinguishing feature. The model's fifth feature portrays insulin concentration in the serum. The individual's body mass index is represented by characteristic 6 (BMI). The feature 7 designation represents the diabetic pedigree analysis. The patient's chronological age is displayed in feature 8, as may be deduced from the number assigned to it. The first eight traits are quantitative, while the ninth is a categorical outcome variable that can take on one of two possible values (0 indicating no diabetes and 1 showing diabetes). This dataset is frequently utilized by researchers in the field of machine learning to construct diagnostic and therapeutic prediction models for type 2 diabetes. Researchers can use this dataset to train a model, which allows them to investigate the link between the many factors and the final variable. This allows them to more correctly predict whether or not a patient has diabetes based on demographic and clinical data.

4.2 Research method

An experimental study design will compare machine learning techniques on the diabetes dataset. This research aims to identify the most effective algorithm for making diabetes predictions based on demographic and clinical patient variables.

4.3 Obtaining and cleaning up raw data

The diabetes dataset will be gathered from a source accessible to the general public. There are eight numeric attributes and a two-class categorical outcome variable across the dataset's 768 rows and nine columns. During preprocessing, outliers and missing values will be eliminated from the data. The data will be standardized to measure all variables on the same scale. Following this, we will divide the data

set in half, utilizing 70% for training and 30% for testing. Several machine-learning approaches will be studied here, including logistic regression, k-nearest neighbors, decision trees, random forests, and support vector machines. Previous research utilizing these algorithms for diabetes prediction has demonstrated promising results. The features will be utilized to train the algorithms on the training set, while grid search will be utilized to fine-tune the hyperparameters of the algorithms. The models' efficacy will be evaluated based on various factors, including their accuracy, sensitivity, specificity, precision, and F1 score. The model's effectiveness will be evaluated using the area under the curve (AUC) and receiver operating characteristic (ROC) curves. The models will be evaluated using a test set to ensure robust generalization to new data. Statistical analysis will be used to compare the efficacy of various machine learning algorithms. They compared the performance measures using statistical tests such as the t-test and the analysis of variance. The data will be displayed using tables and graphs for ease of comprehension.

The diabetes dataset utilized in this study will be taken from a publicly accessible source containing information about diabetes diagnosis. The dataset consists of 768 rows representing patients, and 9 columns containing 8 numerical parameters such as glucose level, blood pressure, and body mass index (BMI), and a two-class categorical outcome variable indicating the presence or absence of diabetes. The data will be preprocessed before the analysis is conducted to ensure quality. Extreme or aberrant numbers will be found and eliminated from the dataset. In addition, any missing values will be dealt with using appropriate methods, such as imputation or deletion. Normalizing the data to permit accurate comparisons and model training is vital. All variables will be normalized to measure them on the same scale. This stage is crucial for machine learning algorithms that rely on distance-based computations.

Two subsets will be created from the dataset: training and testing sets. Seventy percent of the data will be given to the training set, while the remaining thirty percent will be allotted to the testing set. This division enables us to evaluate the performance of trained models on unseen data and their generalizability. Several algorithms will be studied to assess the predictive ability of various machine-learning techniques for diabetes. Among them are logistic regression, K-nearest neighbors, decision trees, random forests, and support vector machines. These algorithms have shown promise in prior diabetes prediction research. The selected characteristics from the dataset will be used to train machine learning models on the training set. During this training phase, the algorithms will discover patterns and associations between the input features and the diabetic outcome. The grid search technique will be utilized to maximize the models' performance. Grid search includes carefully investigating various combinations of hyperparameters (e.g., learning rate, regularization) to determine the optimal configuration. A set of evaluation metrics will be utilized to determine the efficacy of the trained models. Accuracy measures the overall correctness of the predictions; sensitivity and specificity evaluate the model's ability to identify positive and negative instances correctly; precision quantifies the proportion of correctly predicted positive instances; and the F1 score combines precision and recall into a single metric. In addition to these metrics, the area under the curve (AUC) and receiver operating characteristic (ROC) curves will be used to assess the performance of the models. The area under the receiver operating characteristic (ROC) curve depicts the trade-off between accurate positive and false favorable rates at various classification thresholds. To ensure the robustness of the models, they will be evaluated on the testing set, which consists of data that has never been seen. This evaluation highlights the models' ability to generalize predictions to new and unobserved cases.

Statistical analytic techniques will be utilized to compare the efficacy of various machine learning algorithms. These methods may include the t-test and analysis of variance (ANOVA) to determine whether or not there are statistically significant differences between the performance metrics of the various algorithms.

Tables and graphs will be used to visually depict the data to assist in comprehending and presenting the findings. These visualizations will provide a clear and brief summary of the data, making the performance of the machine-learning models easier to comprehend and interpret. This work uses a publicly available dataset to investigate several machine-learning algorithms for diabetes prediction. Through preprocessing, normalization, model training, and evaluation utilizing a variety of performance indicators, the study aims to determine the predictive accuracy of several algorithms for diabetes presence. The statistical analysis and visualizations will illuminate the comparative performance of the models.

4.4 Reproducibility

The procedure will be designed with the ability to replicate the results. The experiment's code will be released as open-source to the public. The experiment's dataset, data preparation procedures, and feature selection algorithms will be publicly available. A complete plan for conducting a machine-learning investigation with the diabetes dataset is provided by this experimental design. Researchers may use this setup to compare the performance of several machine learning algorithms for diabetes prediction, ultimately identifying the most effective algorithm for early diagnosis and treatment of the disease. Eight features are ranked and scored using multiple feature selection methods [46].

4.5 A condensed description of each method and its associated scores

Gain in information measures how important each aspect is in determining the conclusion. Each feature is given a score that reflects how much information it contributes to the overall prediction of the outcome variable. The amount of possible feature values is included in the gain ratio, which is then used to change the information gain score. Each feature is given a score that reflects its gain ratio; higher scores indicate a feature's increased usefulness in predicting the outcome variable. The Gini index quantifies the level of a feature's impureness. Lower scores indicate that the quality is purer and more helpful in predicting the outcome variable; the scores represent each feature's Gini index. The analysis of variance (ANOVA) method determines whether there is a statistically significant difference between the means of a characteristic for the various groups of the outcome variable. More substantial scores for a feature correspond to a higher significance of that characteristic in predicting the outcome variable, as measured by the F-statistic. Method 2 examines whether a character depends on the result variable.

Each characteristic is given a score corresponding to its two statistics; a higher score indicates a stronger correlation between the feature and the result. Feature quality is quantified using ReliefF by comparing the feature values of neighbors belonging to the same and different classes. Higher scores indicate that the feature is more helpful in predicting the outcome variable, while the values represent each feature's ReliefF score. The Feature Selection by Benefit Function (FCBF) method chooses highly relevant characteristics to the outcome variable and is minimally

redundant. Each feature's score reflects its Functional Conservancy Based Fidelity (FCBF) score, with higher scores suggesting greater relevance and fewer redundancies in the feature's ability to predict the outcome variable. According to Table 1, Feature 2 (plasma glucose concentration) is the most informative feature for predicting the outcome variable, as it receives the highest scores from most feature selection procedures. Most methods also give reasonably high scores to variables 8 and 6 (age and BMI, respectively), indicating they are helpful factors for predicting diabetes. The lower scores for the other indicators suggest that they contribute less to accurately predicting the outcome variable.

Table 1. Rank for the feature from 1–8 where the target is feature 9

Feature	Info. Gain	Gain Ratio	Gini	ANOVA	χ^2	ReliefF	FCBF
Perimeter	0.367	0.184	0.216	57.322	34.142	0.085	0.33
Area	0.349	0.174	0.206	45.347	32.711	0.07	0
Compactness	0.249	0.125	0.151	34.86	25.467	0.053	0.203
Id	0.108	0.054	0.068	10.94	7.244	0.062	0.079
Symmetry	0.048	0.024	0.032	5.627	5.032	0.018	0
Smoothness	0.045	0.022	0.029	3.983	3.862	0.021	0
Radius	0.031	0.015	0.02	3.168	3.227	0.01	0
Texture	0.014	0.007	0.009	0.493	0.633	–0.004	0
Fractal dimension	0.002	0.001	0.001	0.007	0.017	0.017	0

5 SAMPLING TYPE

Seventy percent of the information was gleaned from a randomly selected sample, then stratified and evaluated. This means that 70% of the data was chosen randomly to be as representative as possible of the whole population. Stratification was used where practical to guarantee that each group of interest in the outcome variable was appropriately represented in the sample. Deterministic sampling ensured reproducibility by always drawing from the same pool. Input: Patient demographics and clinical data such as the number of pregnancies, plasma glucose concentration, and body mass index were included in the input dataset's 768 occurrences. Sample: According to the sampling strategy, we randomly selected 538 occurrences or 70% of the data. This selection is a subset of the actual events that can be used in further research and modelling. The remaining cases did not cut for any reason. In this case, 30% of the original dataset, or 230 occurrences, would remain. These data were not utilized in the research and modelling but might be used for verification. Overall, a deterministic random sample was used here, representing 70% of the available data, and it was stratified if possible. This sampling method is commonly used in data analysis and modelling to ensure that the piece is representative of the population and can be used to make accurate conclusions. Machine learning and statistical analysis frequently make use of the algorithms above. Each algorithm is described below: A random forest is an ensemble learning method that uses many decision trees to construct a more accurate and robust model. It creates a set of decision trees using random data subsets to get at an overall prediction. Logistic Regression: A Statistical Method for Evaluating

Datasets Where Multiple Independent Factors determine the Outcome. The values of the independent variables are used as inputs into a model that predicts the probability of a specific outcome. A decision tree is a diagrammatic structure that looks like a flowchart and shows the results of attribute tests as internal nodes, the outcomes of those tests as branches, and the labels assigned to classes as leaf nodes. Support vector machines (SVMs) are supervised learning techniques in regression and classification. SVMs find the hyperplane in the feature space that most accurately separates classes.

AdaBoost is a machine learning algorithm that pools the results of several underperforming classifiers into a single, highly accurate model. Each data point is given a weight, adjusted with each iteration so that the algorithm can zero in on misclassified data. Neural networks are a machine learning system inspired by how human brains work. Among its many applications are pattern recognition, categorization, and regression analysis. K-nearest neighbors (kNN) is a machine learning method used for regression and classification. It works by finding the K-nearest neighbors of a new data point and making a label prediction for that point based on the labels of its neighbors. Naive Bayes is a machine learning classification method based on probabilistic inference. The likelihood of a hypothesis (here, a class label) is proportional to the evidence's probability, as stated by Bayes' theorem (in this case, the features). Decision rules may be learned from data using CN2, a rule induction technique. Repeated iterations of adding new rules improve the model's precision. It is common practice in machine learning to utilize Stochastic Gradient Descent (SGD), an optimization method, while training models. The model is iteratively refined by adjusting its input parameters to reduce the gap between predicted and observed values. Pattern recognition, regression analysis, and classification applications all use these techniques.

6 TESTING

Five metrics (Accuracy, F1, Precision, and Recall) are typically utilized when comparing classification models on the same dataset. The results were averaged across all classes, as shown in Table 2, which shows that the test was done using stratified 5-fold cross-validation. The following may be inferred from the data in Table 2: On average, across all classes, the provided dataset best suits the Logistic Regression and Neural Network models, which obtain the highest scores on most criteria. When considering all classes together, the SVM, Random Forest, and Naive Bayes models all receive moderate to high scores, showing they are also strong classification models.

With poorer scores on most criteria across all classes on average, the kNN and Tree models may not be the best choices for this data set. When comparing all models using a weighted average of class scores, SGD, AdaBoost, and CN2 rule inducer models fare the poorest. Notably, a standard method for evaluating a classification model's efficacy, stratified 5-fold cross-validation, was used in this evaluation. This method prevents unfair evaluation by splitting the dataset into five sections containing the same percentage of each type. Table 2 provides helpful insight into the performance of several classification models on the provided dataset when averaged across all classes using stratified 5-fold cross-validation. However, while evaluating the findings and choosing the best model for the case, it is crucial to account for the problem's context and the application's unique requirements.

Table 2. Sampling type: Stratified 5-fold cross-validation, target class: None, show the average over classes

Model	AUC	CA	F1	Prec	Recall
AdaBoost	0.636	0.667	0.668	0.669	0.667
SGD	0.663	0.686	0.688	0.691	0.686
Tree	0.667	0.708	0.706	0.704	0.708
CN2 rule inducer	0.692	0.638	0.637	0.636	0.638
kNN	0.774	0.714	0.708	0.706	0.714
Random Forest	0.805	0.729	0.722	0.721	0.729
Naive Bayes	0.808	0.729	0.733	0.741	0.729
SVM	0.815	0.747	0.736	0.74	0.747
Neural Network	0.82	0.76	0.751	0.754	0.76
Logistic Regression	0.822	0.76	0.752	0.754	0.76

Five metrics (Accuracy, CA, F1, Precision, and Recall) are typically utilized when comparing classification models on the same dataset. Performance for target class 0 is provided in the results, and Table 3 also reveals that the evaluation was done using stratified 5-fold cross-validation. The following may be inferred from Table 3: On the provided dataset for target class 0, the Logistic Regression, SVM, and Neural Network models achieve the highest scores on most metrics, indicating that they are the best-performing models when averaged across all folds. Random Forest, kNN, Naive Bayes, and Tree models, on average over all folds, score moderately to highly on most criteria, showing that they are also strong classification models for target class 0. Scores for the AdaBoost and CN2 rule inducer models are often lower across all folds, suggesting that they are not optimal for this target class. The SGD model has the lowest average scores for target class 0 when considering all folds together. The evaluation used stratified 5-fold cross-validation to avoid favoring any group or data set. In addition, the assessment was limited to target class 0, suggesting that it may have had some particular importance in the scenario. Table 3 provides helpful insights into the performance of several classification models on the target class 0 for the provided dataset when averaged across all folds using stratified 5-fold cross-validation. However, while evaluating the findings and choosing the best model for the case, it is crucial to account for the problem's context and the application's unique requirements.

Table 3. Sampling type: Stratified 5-fold cross-validation, target class: 0

Model	AUC	CA	F1	Prec	Recall
Tree	0.665	0.708	0.78	0.766	0.794
Random Forest	0.811	0.729	0.801	0.767	0.837
Logistic Regression	0.828	0.76	0.826	0.781	0.877
SVM	0.822	0.747	0.819	0.767	0.877
AdaBoost	0.636	0.667	0.743	0.746	0.74
Neural Network	0.825	0.76	0.827	0.78	0.88
kNN	0.776	0.714	0.788	0.761	0.817
Naive Bayes	0.808	0.729	0.782	0.819	0.749
CN2 rule inducer	0.691	0.638	0.723	0.72	0.726
SGD	0.663	0.686	0.754	0.769	0.74

The results for target class 1 are shown in Table 4, and the evaluation was carried out using stratified 5-fold cross-validation.

The following may be inferred from the data in Table 4: On average, overall folds, the Naive Bayes and Logistic Regression models perform the best when attempting to predict target class 1 from the provided information. When averaged across all folds, the Neural Network and SVM models perform similarly well as classification tools for the target class 1. Overall, the Random Forest, kNN, Tree, and SGD models perform worse than the average, suggesting they are not the best choices for this target class.

When considering the average performance over all folds, the AdaBoost and CN2 rule inducer models fare the poorest, with the lowest scores on most metrics for target class 1. The evaluation used stratified 5-fold cross-validation to avoid favoring any group or data set. In addition, the assessment was limited to the first target class, suggesting that this subset may be especially relevant or essential. The results of multiple classification models on the target class 1 for the provided dataset are summarized in Table 4 when the results of stratified 5-fold cross-validation are averaged over all folds. However, while evaluating the findings and choosing the best model for the case, it is crucial to account for the problem's context and the application's unique requirements.

Table 4. Sampling type: Stratified 5-fold cross-validation, target class: 1

Model	AUC	CA	F1	Prec	Recall
Tree	0.665	0.708	0.567	0.589	0.548
Random Forest	0.811	0.729	0.576	0.635	0.527
Logistic Regression	0.828	0.76	0.613	0.703	0.543
SVM	0.822	0.747	0.583	0.688	0.505
AdaBoost	0.636	0.667	0.528	0.524	0.532
Neural Network	0.825	0.76	0.61	0.706	0.537
kNN	0.776	0.714	0.56	0.605	0.521
Naive Bayes	0.808	0.729	0.64	0.596	0.691
CN2 rule inducer	0.691	0.638	0.477	0.481	0.473
SGD	0.663	0.686	0.566	0.547	0.585

7 DISCUSSION

In this work, the predicted accuracy of multiple machine learning algorithms for diabetes was evaluated using the Pima Indians diabetes dataset. According to the results, the Random Forest approach achieved 76.30 percent accuracy, while the Neural Network algorithm achieved 78.57 percent accuracy. These findings are consistent with other studies that have used the same dataset and evaluated similar approaches. The study also found that body mass index, glucose levels, and age were the most vital indicators of diabetes risk. The study's small dataset and the methods examined within it may not generalize well to other contexts or populations. Furthermore, potential influences on diabetes, such as lifestyle and genetics, were not considered in the study. More research is needed to determine whether machine learning algorithms effectively predict diabetes in larger datasets and across diverse demographics [47–49].

8 CONCLUSION

The Pima Indians diabetes dataset was utilized in this study to assess the predictive accuracy of several different machine-learning algorithms for diabetes. According to the findings, the Random Forest method attained an accuracy of 76.30 percent, while the Neural Network technique produced an accuracy of 78.57 percent. According to the findings of the study, machine learning algorithms have the potential to serve as an effective early detection tool and contribute to diabetes prediction. However, more datasets and populations must be researched to determine the usefulness of these algorithms, and other diabetes-related characteristics must also be considered.

9 REFERENCES

[1] E. N. Haner Kirgil, B. Erkal, and T. E. Ayyildiz, "Predicting diabetes using machine learning techniques," in *2022 5th International Conference on Theoretical and Applied Computer Science and Engineering, ICTASCE* 2022, 2022, pp. 137–141. https://doi.org/10.1109/ICTACSE50438.2022.10009726

[2] 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, 2021, pp. 141–146. https://doi.org/10.1109/ICACCS51430.2021.9441935

[3] D. Tripathi, S. K. Biswas, S. Reshmi, A. N. Boruah, and B. Purkayastha, "Diabetes prediction using machine learning analytics: Ensemble learning techniques," 2022 *2nd Asian Conf. Innov. Technol. ASIANCON* 2022, 2022. https://doi.org/10.1109/ASIANCON55314.2022.9908975

[4] S. Wadhwa and K. Babber, "Artificial intelligence in health care: Predictive analysis on diabetes using machine learning algorithms," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12250, pp. 354–366, 2020, https://doi.org/10.1007/978-3-030-58802-1_26

[5] B. S. Ahamed, M. S. Arya, and A. O. Nancy V, "Prediction of type-2 diabetes mellitus disease using machine learning classifiers and techniques," *Frontiers in Computer Science,* vol. 4. 2022. https://doi.org/10.3389/fcomp.2022.835242

[6] Y. Ye et al., "Comparison of machine learning methods and conventional logistic regressions for predicting gestational diabetes using routine clinical data: A retrospective cohort study," *J. Diabetes Res.,* vol. 2020, 2020, https://doi.org/10.1155/2020/4168340

[7] C. Hettiarachchi and C. Chitraranjan, "A machine learning approach to predict diabetes using short recorded photoplethysmography and physiological characteristics," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* 2019, pp. 322–327. https://doi.org/10.1007/978-3-030-21642-9_41

[8] U. Anand, A. Sehgal, S. Tripathi, G. Singh, R. Sharma, and Manisha, "An analysis of predicting diabetes using machine learning," *Journal of Control System and Control Instrumentation*, vol. 4, no. 3, pp. 16–26, 2018. https://doi.org/10.5281/zenodo.1462319

[9] T. N. Poly, M. M. Islam, and Y. C. J. Li, "Early diabetes prediction: A comparative study using machine learning techniques," in *Studies in Health Technology and Informatics,* 2022. https://doi.org/10.3233/SHTI220752

[10] R. D. Joshi and C. K. Dhakal, "Predicting type 2 diabetes using logistic regression and machine learning approaches," *Int. J. Environ. Res. Public Health,* vol. 18, no. 14, 2021. https://doi.org/10.3390/ijerph18147346

[11] S. Reshmi, S. K. Biswas, A. N. Boruah, D. M. Thounaojam, and B. Purkayastha, "Diabetes prediction using machine learning analytics," in *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, COM-IT-CON* 2022, 2022, pp. 108–112. https://doi.org/10.1109/COM-IT-CON54601.2022.9850922

[12] S. Samet, M. R. Laouar, I. Bendib, and S. Eom, "Analysis and prediction of diabetes disease using machine learning methods," *Int. J. Decis. Support Syst. Technol.,* vol. 14, no. 1, 2022. https://doi.org/10.4018/IJDSST.303943

[13] S. Yuan, Y. Sun, X. Xiao, Y. Long, and H. He, "Using machine learning algorithms to predict candidaemia in ICU patients with new-onset systemic inflammatory response syndrome," *Front. Med.,* vol. 8, 2021. https://doi.org/10.3389/fmed.2021.720926

[14] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," *Healthc. Technol. Lett.,* 2022. https://doi.org/10.1049/htl2.12039

[15] S. S. Bhat, G. A. Ansari, and G. A. Ansari, "Predictions of diabetes and diet recommendation system for diabetic patients using machine learning techniques," 2021 *2nd Int. Conf. Emerg. Technol.,* 2021, https://doi.org/10.1109/INCET51464.2021.9456365

[16] A. Das, S. K. Das, D. Das, and K. M. R. Alam, "A comparative study to predict diabetes using machine learning techniques," in *2021 International Conference on Science and Contemporary Technologies,* ICSCT 2021, 2021. https://doi.org/10.1109/ICSCT53883.2021.9642493

[17] S. S. Bhat, V. Selvam, G. A. Ansari, M. D. Ansari, and M. H. Rahman, "Prevalence and early prediction of diabetes using machine learning in North Kashmir: A case study of district Bandipora," *Comput. Intell. Neurosci.,* 2022. https://doi.org/10.1155/2022/2789760

[18] U. E. Laila, K. Mahboob, A. W. Khan, F. Khan, and W. Taekeun, "An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study," *Sensors,* vol. 22, no. 14, 2022. https://doi.org/10.3390/s22145247

[19] Y. Hu, L. Zheng, and J. Wang, "Predicting ICU length of stay for patients with diabetes using machine learning techniques," in *Proceedings of the International Conference on Cyber-Physical Social Intelligence, ICCSI* 2022, 2022, pp. 417–422. https://doi.org/10.1109/ICCSI55536.2022.9970666

[20] Q. Liu, Q. Zhou, Y. He, J. Zou, Y. Guo, and Y. Yan, "Predicting the 2-year risk of progression from prediabetes to diabetes using machine learning among chinese elderly adults," *J. Pers. Med.,* vol. 12, no. 7, 2022. https://doi.org/10.3390/jpm12071055

[21] H. F. Ahmad, H. Mukhtar, H. Alaqail, M. Seliaman, and A. Alhumam, "Investigating health-related features and their impact on the prediction of diabetes using machine learning," *Appl. Sci.,* vol. 11, no. 3, pp. 1–18, 2021. https://doi.org/10.3390/app11031173

[22] S. P. Kesavan and R. Rajeswari, "Analysis of power and delay in CMOS universal gates using single virtual rail clamping technique," *Asian J. Res. Soc. Sci. Humanit.,* vol. 6, no. 10, p. 68, 2016. https://doi.org/10.5958/2249-7315.2016.00998.9

[23] J. Ye, L. Yao, J. Shen, R. Janarthanam, and Y. Luo, "Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes," *BMC Med. Inform. Decis. Mak.,* vol. 20, 2020. https://doi.org/10.1186/s12911-020-01318-4

[24] P. Theerthagiri, A. U. Ruby, and J. Vidya, "Diagnosis and classification of the diabetes using machine learning algorithms," *SN Comput. Sci.,* vol. 4, no. 1, 2023. https://doi.org/10.1007/s42979-022-01485-3

[25] S. Perveen, M. Shahbaz, T. Saba, K. Keshavjee, A. Rehman, and A. Guergachi, "Handling irregularly sampled longitudinal data and prognostic modeling of diabetes using machine learning technique," *IEEE Access,* vol. 8, pp. 21875–21885, 2020. https://doi.org/10.1109/ACCESS.2020.2968608

[26] Y. Miao, "Using machine learning algorithms to predict diabetes mellitus based on PIMA indians diabetes dataset," *ACM Int. Conf. Proceeding Ser.*, pp. 47–53, 2021. https://doi.org/10.1145/3463914.3463922

[27] Y. Jian, M. Pasquier, A. Sagahyroon, and F. Aloul, "Using machine learning to predict diabetes complications," in *BioSMART 2021 – Proceedings: 4th International Conference on Bio-Engineering for Smart Technologies,* 2021. https://doi.org/10.1109/BioSMART54244.2021.9677649

[28] A. Choudhury and D. Gupta, "A survey on medical diagnosis of diabetes using machine learning techniques," in *Advances in Intelligent Systems and Computing,* 2019, pp. 67–78. https://doi.org/10.1007/978-981-13-1280-9_6

[29] S. Perveen, M. Shahbaz, K. Keshavjee, and A. Guergachi, "Prognostic modeling and prevention of diabetes using machine learning technique," *Sci. Rep.,* vol. 9, no. 1, 2019. https://doi.org/10.1038/s41598-019-49563-6

[30] M. Aminul and N. Jahan, "Prediction of onset diabetes using machine learning techniques," *Int. J. Comput. Appl.,* vol. 180, no. 5, pp. 7–11, 2017. https://doi.org/10.5120/ijca2017916020

[31] S. K. Dey, A. Hossain, and M. M. Rahman, "Implementation of a web application to predict diabetes disease: An approach using machine learning algorithm," in *2018 21st Int. Conf. Comput. Inf. Technol. ICCIT* 2018, 2019. https://doi.org/10.1109/ICCITECHN.2018.8631968

[32] C. Charitha, A. D. Chaitrasree, P. C. Varma, and C. Lakshmi, "Type-II diabetes prediction using machine learning algorithms," in *2022 International Conference on Computer Communication and Informatics, ICCCI* 2022, 2022. https://doi.org/10.1109/ICCCI54379.2022.9740844

[33] L. Yang, N. Gabriel, I. Hernandez, A. G. Winterstein, and J. Guo, "Using machine learning to identify diabetes patients with canagliflozin prescriptions at high-risk of lower extremity amputation using real-world data," *Pharmacoepidemiol. Drug Saf.,* vol. 30, no. 5, pp. 644–651, 2021. https://doi.org/10.1002/pds.5206

[34] N. Abdulhadi and A. Al-Mousa, "Diabetes detection using machine learning classification methods," *2021 Int. Conf. Inf. Technol. ICIT 2021 – Proc.,* pp. 350–354, 2021. https://doi.org/10.1109/ICIT52682.2021.9491788

[35] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, "Classification and prediction of diabetes disease using machine learning paradigm," *Heal. Inf. Sci. Syst.,* vol. 8, no. 1, 2020. https://doi.org/10.1007/s13755-019-0095-z

[36] A. L. Neves et al., "Using electronic health records to develop and validate a machine-learning tool to predict type 2 diabetes outcomes: A study protocol," *BMJ Open,* vol. 11, no. 7, 2021. https://doi.org/10.1136/bmjopen-2020-046716

[37] M. F. Faruque, Asaduzzaman, and I. H. Sarker, "Performance analysis of machine learning techniques to predict diabetes mellitus," *2nd Int. Conf. Electr. Comput. Commun. Eng. ECCE 2019,* 2019. https://doi.org/10.1109/ECACE.2019.8679365

[38] F. Bano, K. Munidhanalakshmi, and R. MadanaMohana, "Predict diabetes mellitus using machine learning algorithms," in *Journal of Physics: Conference Series,* 2021. https://doi.org/10.1088/1742-6596/2089/1/012002

[39] H. Ahmed, E. M. G. Younis, and A. A. Ali, "Predicting diabetes using distributed machine learning based on Apache Spark," in *Proceedings of 2020 International Conference on Innovative Trends in Communication and Computer Engineering, ITCE* 2020, 2020, pp. 44–49. https://doi.org/10.1109/ITCE48509.2020.9047795

[40] H. M. Deberneh et al., "1233-P: Prediction of type 2 diabetes occurrence using machine learning model," *Diabetes,* vol. 69, no. Supplement_1, 2020. https://doi.org/10.2337/db20-1233-P

[41] Y. Xiong et al., "Prediction of gestational diabetes mellitus in the first 19 weeks of pregnancy using machine learning techniques," *J. Matern. Neonatal Med.,* vol. 35, no. 13, pp. 2457–2463, 2022. https://doi.org/10.1080/14767058.2020.1786517

[42] R. Priyadarshini, N. Dash, and R. Mishra, "A novel approach to predict diabetes mellitus using modified extreme learning machine," *2014 Int. Conf. Electron. Commun. Syst. ICECS* 2014, 2014. https://doi.org/10.1109/ECS.2014.6892740

[43] A. S. Alanazi and M. A. Mezher, "Using machine learning algorithms for prediction of diabetes mellitus," *2020 Int. Conf. Comput. Inf. Technol. ICCIT* 2020, 2020. https://doi.org/10.1109/ICCIT-144147971.2020.9213708

[44] H. Shao, L. Shi, Y. Lin, and V. Fonseca, "Using modern risk engines and machine learning/artificial intelligence to predict diabetes complications: A focus on the BRAVO model," *Journal of Diabetes and its Complications,* vol. 36, no. 11, 2022. https://doi.org/10.1016/j.jdiacomp.2022.108316

[45] C. Pankaj, K. V. Singh, and K. R. Singh, "Artificial intelligence enabled web-based prediction of diabetes using machine learning approach," in *Proceedings of IEEE International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications, CENTCON* 2021, 2021, pp. 60–64. https://doi.org/10.1109/CENTCON52345.2021.9688236

[46] "Pima Indians Diabetes Database," 2023. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[47] M. Al-Batah, B. Zaqaibeh, S. A. Alomari, and M. S. Alzboon, "Gene Microarray Cancer classification using correlation based feature selection algorithm and rules classifiers," *Int. J. Online Biomed. Eng.,* vol. 15, no. 8, pp. 62–73, 2019. https://doi.org/10.3991/ijoe.v15i08.10617

[48] D. Novaliendry et al., "Hemodialysis patient death prediction using logistic regression," *Int. J. Online Biomed. Eng.,* vol. 19, no. 9, pp. 66–80, 2023. https://doi.org/10.3991/ijoe.v19i09.40917

[49] Z. Sabouri, N. Gherabi, M. Nasri, M. Amnai, H. El Massari, and I. Moustati, "Prediction of depression via supervised learning models: Performance comparison and analysis," *Int. J. Online Biomed. Eng.,* vol. 19, no. 9, pp. 93–107, 2023. https://doi.org/10.3991/ijoe.v19i09.39823

10 AUTHORS

Mowafaq Salem Alzboon is an Assistant Professor at the Science and Information Technology Faculty in Jadara University, Jordan. He holds a PhD degree in computer science from University Utara Malaysia. Dr. Alzboon's research interests center around Cloud Computing, Autonomic Computing, Visualization technology, Load balancing, Overlay Network, Mobile Application Development, and Internet of Things. He has expertise in a range of disciplines, including Computer Architecture, Computer Communications (Networks), and Distributed Computing. Dr. Alzboon's skillset encompasses Overlay Network, Computer Networks, Load Balancing, Cloud Computing, Parallel and Distributed Computing, Networking, Virtualization, Virtualization technology, Distributed Computing, Autonomic Computing, and Grid Computing. He may be contacted at the following email address: malzboon@jadara.edu.jo.

Mohammad Subhi Al-Batah holds a PhD in Computer Science with a specialization in Artificial Intelligence, which he received from the University of Science Malaysia in 2009. He currently serves as a lecturer at the Faculty of Sciences and Information Technology in Jadara University, Jordan. In 2018, he also served as the Director of the Academic Development and Quality Assurance Center at Jadara

University. Dr. Al-Batah's research interests span a range of topics, including image processing, Artificial Intelligence, real-time classification, and software engineering. He may be contacted at the following email address: albatah@jadara.edu.jo.

Muhyeeddin Alqaraleh is an Assistant Professor at Jadara University. Dr. Alqaraleh's areas of expertise include Computer Engineering, Information and Communication Technology, Computer Networking, Digital Signal Processing, Electronics and Communication Engineering, Signal, Image and Video Processing, Information Technology, Network Communication, Communication & Signal Processing, Networking, Cloud Computing, Network Security, Network Architecture, Wireless Computing, Network technology, Signal Processing, Signal Processing for Communication, Radio Communication, Information Theory, Discrete-Time Signal Processing, Computer Technology, IT Infrastructure, Hardware Troubleshooting, Computer Networks Security, Security, Network Management, IT Security, Network Administration, Network Configuration, Network Simulation, and Information Security. He is fluent in English, Arabic, and Russian. He may be contacted at the following email address: m.qaralleh@jadara.edu.jo.

Ahmad Abuashour is a Professor Assistant in the Department of Information Technology and Computing at Arab Open University in Kuwait. He received his Bachelor of Engineering degree in Computer Engineering from Jordan University of Science and Technology and his Master of Engineering degree in Electrical Engineering from Concordia University. Currently, he is pursuing his PhD in Electrical Engineering at Ecole de Technologie Superieure (ETS), University of Quebec, Canada. His current research interests are focused on Intelligent Transportation Systems (ITS), Vehicular Ad-Hoc Networks (VANETs), cluster-based routing protocols, network management and monitoring, and quality of service. Specifically, he is concentrating on VANET routing protocols. The Faculty of Computer Studies at Arab Open University is in the Ardiya Industrial Area, Farwanya, with the Information Technology and Computing (ITC) department being assigned the postal code 13033. Postal correspondence may be addressed to P.O. Box 3322 Kuwait. To contact Ahmad Abuashour, please use the following email address: aabuashour@aou.edu.kw.

Ahmad Fuad Hamadah Bader is an Assistant Professor at the Department of Communication and Computer Engineering in the Engineering College of Jadara University, located in Irbid, Jordan. He holds a PhD in Engineering with a specialization in Computer Systems and Networks from Donetsk State Technical University in Donetsk, Ukraine, which he earned in 2007. He also holds a Master of Science in Engineering degree in Computer Systems Complexes and Networks from State Technical University named after Beruni in Tashkent, Uzbekistan, which he earned in 1992, and a Bachelor of Science in Engineering degree in Computer Systems Complexes and Networks from Donetsk State Technical University in Donetsk, Ukraine, which he earned in 1989. He may be contacted at the following email address: abader@jadara.edu.jo.