

Review of Computer Engineering Research

2024 Vol. 11, No. 1, pp. 16-29

ISSN(e): 2410-9142

ISSN(p): 2412-4281

DOI: 10.18488/76.v11i1.3598


© 2024 Conscientia Beam. All Rights Reserved.



Machine learning algorithms-based decision support model for diabetes

 **Karthick Kanagarathinam¹⁺**

 **R. Manikandan²**

 **T. Sathish Kumar³**

¹Department of Electrical and Electronics Engineering, GMR Institute of Technology, Rajam, Andhra Pradesh, India.

Email: karthick.k@gmr.it.edu.in

²Department of Electronics and Communications Engineering, Panimalar Engineering College, Chennai, India.

Email: money_kandan2004@yahoo.co.in

³Department of Electrical and Electronics Engineering, S.A. Engineering College, Chennai, India.

Email: sathishkumart@saec.ac.in



(+ Corresponding author)

ABSTRACT

Article History

Received: 27 September 2023

Revised: 30 November 2023

Accepted: 13 December 2023

Published: 11 January 2024

Keywords

Boruta feature selection
Classification
Data mining
Diabetes
Machine learning
Prediction model.

This research explores the application of machine learning (ML)-based risk prediction models in early diabetes disease detection for healthcare professionals. Diabetes affects millions of people worldwide. In light of significant advancements in biomedical sciences, vast volumes of data have been generated, including high-throughput genetic and diagnostic data sourced from extensive health records. Leveraging an initial diabetes risk prediction dataset from the University of California Irvine (UCI) ML repository, our research focused on supervised learning techniques, constituting 85% of the employed methods. The remaining 15% comprised unsupervised learning approaches, specifically association rules. A key contribution of this study lies in the development of an optimal prediction model utilizing supervised ML algorithms. The Boruta feature selection algorithm was employed to identify pertinent features, and the subsequent models were validated using a preprocessed dataset containing 10 attributes. Notably, the risk prediction models generated through random forest, extreme gradient boosting (XGBoost), and light gradient boosting machine (LightGBM) exhibited impressive average accuracies of 98.13%, 97.37%, and 97.22%, respectively, as determined via 10-fold cross-validation with 15 repetitions. Furthermore, these models achieved exceptional area under the ROC curve (AUC) values of 1, 0.99, and 0.99, respectively, showcasing their robustness and efficacy in diabetes risk prediction.

Contribution/Originality: This study introduces a novel approach to diabetes risk prediction by employing a combination of Boruta feature selection algorithm and advanced machine learning classifiers (XGBoost, LightGBM, and Random Forest). This research is unique because it uses these techniques along with strict validation methods and in-depth insights into feature selection. This makes early-stage diabetes detection models more accurate and useful.

1. INTRODUCTION

Diabetes, a chronic disease, manifests in humans when blood glucose levels, commonly referred to as blood sugar levels, become abnormally elevated. Predicting diabetes at an early stage holds the potential for improved treatment outcomes. The primary source of energy for humans is glucose in the blood, derived from the consumption of food. Insulin enables the transfer of glucose from food into cells, providing the body with vital energy. However, inadequate insulin production leads to a situation where glucose accumulates in the bloodstream

instead of reaching the cells. This prolonged elevation of blood glucose can result in various health complications [1]. Alarming, it is projected that diabetes could cause the death of approximately 592 million people by 2035. The economic burden of diabetes has also shown a global increase [2]. Significantly, diabetes is a significant factor in the occurrence of visual impairment, kidney issues, heart attacks, strokes, and amputations of lower limbs. In the year 2019, diabetes mellitus (DM) was the ninth most prevalent reason for mortality, contributing to around 1.5 million reported deaths [3].

The process of extracting insights and uncovering patterns within datasets through the utilization of ML, statistical techniques, and database systems is termed data mining [4]. Even with the constrained exploration within the realm of bioinformatics, progress in the domains of computing and statistics has cleared the path for creating investigative models grounded in computation and statistical methods. The diagnosis of DM presents itself as a complex quantitative research challenge [5-7]. To make the best diabetes prediction model, you need to use a dataset with important features like frequent urination, weight loss, frequent eating, genital thrush, blurred vision, slow healing, and paresis.

Polyuria, characterized by abnormally frequent urination, is a medical condition [8]. Numerous urinary conditions characterized by excessive urination are associated with polydipsia, which refers to an intense feeling of thirst. As a consequence, individuals may experience an ongoing need to compensate for fluid and electrolyte loss through frequent urination [9].

Weight loss, an unexplained reduction in body weight, signifies a loss of 10 pounds (approximately 4.5 kg) or 5% of normal body weight over a period of 6 months to one year without a discernible cause. Diabetic patients typically encounter a decline in their average weight [10]. Weakness, also known as diabetes-related asthenia, encompasses feelings of fatigue or tiredness in the body. This weakness may lead to an inability to properly move specific body parts due to a lack of energy [11].

Polyphagia, or excessive hunger or increased appetite, is one of the three primary diabetes symptoms [12]. Thrush is more prevalent among diabetics due to elevated sugar levels that create favorable conditions for yeast growth. Diabetes exacerbates yeast infections by promoting the growth of candida. Increased sugar levels in the blood trigger heightened production of sweat, saliva, and urine within the body, fostering yeast growth in areas like the mouth and genitals, leading to thrush. Individuals with DM are more prone to developing vulvovaginal candidiasis compared to those with normal blood glucose levels [13].

Despite advancements in understanding ocular diseases and identifying effective treatments, DM and its associated retinal complications persist as major causes of blindness. Timely diagnosis and intervention can prevent all ocular issues related to DM [14]. Diabetes foot ulcers (DFUs), the primary cause of amputations, affect around 15% of diabetic individuals. Reduced reactions of cells and growth factors lead to decreased blood circulation in the surrounding areas and limited local formation of new blood vessels. These factors collectively contribute to compromised healing in individuals afflicted with DFUs [15].

Paresis denotes weakened muscle movement, where individuals retain some degree of control over affected muscles, in contrast to paralysis. Diabetic gastroparesis, a complex ailment, demands a multifaceted approach [16]. Diabetes and obesity have been classified as epidemics by the World Health Organization (WHO) due to their escalating prevalence. Obesity not only underpins the aetiology of the most prevalent type of diabetes globally, type 2 DM, but also contributes to its progression [17].

In the realm of health informatics, ML [18-20] plays a pivotal role in the early prediction of diseases. Datasets derived from the healthcare sector serve as essential resources for crafting optimal prediction models. These models aid medical practitioners in making informed decisions.

This research endeavor is dedicated to constructing the finest early-stage diabetes risk prediction model using ML classifier algorithms. A dataset containing essential features such as polyuria, polydipsia, polyphagia, genital thrush, delayed healing, and obesity, among others, was identified for the development of this predictive model. The

diabetes prediction model was evaluated using 19 ML classification algorithms. To ensure the model's robustness, its stability will be verified through repeated k-fold cross-validation.

The major contributions and findings are outlined below:

- The diabetes prediction dataset encompasses attributes such as age, gender, and various symptoms related to diabetes. These attributes are described in terms of their qualitative or quantitative nature, and data visualization reveals the relationships between these attributes and the likelihood of diabetes.
- In this research, the Boruta feature selection algorithm has been employed for the purpose of selecting pertinent features for the model's development. The 10 features have been selected for model development.
- The diabetes prediction model has been developed by employing 19 ML classification algorithms. The best machine learning algorithms GBoost, LightGBM, and random forest were tested over and over with k-fold cross-validation to make sure the model was stable.
- Diverse metrics were used to assess the model's efficiency, including accuracy, precision, recall, and F1-score. The evaluation of the model's effectiveness involved utilizing both the confusion matrix and ROC curves.
- The findings of this research highlight the model's accuracy and effectiveness in predicting diabetes risk, contributing valuable insights for medical practitioners and researchers alike.

2. RELATED WORKS

ML contributes to enhancing business decisions [21] boosting productivity, diagnosing diseases [22], forecasting weather [23], text recognition [24], identifying power quality issues [25], and much more.

Shetty, et al. [26] constructed an intelligent system for predicting DM diseases, utilizing a Bayesian and k-Nearest Neighbour (kNN) algorithm that evaluates DM conditions based on a diabetes diagnosis database. Similarly, Georga, et al. [27] formulated a model for predicting subcutaneous (SC) glucose concentrations using support vector regression. The model includes factors like the subcutaneous glucose profile, density of plasma insulin, presence of glucose from meals in the overall bloodstream, and energy demands during physical exercises.

Fitriyani, et al. [28] created a forecasting model for type 2 DM, hypertension, prehypertension, and chronic kidney disease using an ensemble learning technique. They attained accuracies of 96.74%, 85.73%, 75.78%, and 100% across distinct datasets. In predicting type 2 DM, they categorized the outcome as positive or negative based on the glycosylated hemoglobin (glyhb) value.

Barakat, et al. [29] formulated a support vector machine (SVM)-based framework to diagnose diabetes, achieving accuracy of 94%, sensitivity of 93%, and specificity of 94%. Similarly, Le, et al. [30] developed a machine learning model aimed at forecasting the occurrence of early-onset diabetes in patients. This plan used a wrapper-based method to pick features, improving the Multilayer Perception (MLP) with Grey Wolf Optimization (GWO) and Adaptive Particle Swarm Optimization (APSO) to lower the number of input features that are needed. The outcomes they obtained demonstrated enhanced predictive precision, reaching 96% using GWO-MLP and 97% employing Adaptive Particle-Grey Wolf Optimization (APGWO)-MLP.

The collection of data containing important attributes will assist in creating an optimal predictive model and offer enhanced support to healthcare professionals. In their research, Singh, et al. [31] introduced eDiaPredict, an ensemble-based approach, for predicting diabetes using the PIMA Indian diabetes dataset (PIDD). This technique yielded an accuracy level of 95%. This prediction approach involves an ensemble of various ML algorithms, including XGBoost, Random Forest, SVM, Neural Network, and Decision Tree (DT), to determine the diabetes status of individuals.

Hasan, et al. [32] used the PIDD to develop an ensembling classifier. Classifiers from the ML toolbox including kNN, DTs, Random Forest, AdaBoost, Naive Bayes, and XGBoost, as well as the MLP, were used. To further enhance diabetes prediction, this research proposes weighted ensembling of several ML models, with individual model weights computed using the ML's AUC. Calculations reveal an AUC of 0.95.

Rajendra and Latifi [33] conducted their analysis with the Python Integrated Development Environment (IDE) and utilised logistic regression (LR) as the primary algorithm. The study employs information from both the PIMA Indian Diabetes dataset and the Vanderbilt dataset. There are two primary techniques used in the process of feature selection. Furthermore, they employ ensemble approaches, which, in comparison to a single model, provide more accurate predictions and boost overall performance. For dataset 1, the Max Voting Ensemble method gave the best accuracy, which was about 78%. For dataset 2, the Max Voting and Stacking Ensemble methods gave the best accuracy, which was about 93%.

In order to determine the chance that a patient would suffer from type 2 diabetes, Raghavendran, et al. [34] analysed a dataset containing information from actual patients. With the use of classification algorithms, they examined the patient dataset to make diabetes forecasts. SVM, LR, kNN, DT, Random Forest, AdaBoost (AdaBoost), and Nave Bayes Classification are tested on the PIDD to determine which produces better results. As shown in their study, AdaBoost is highly effective, with a success rate of 95%.

In this article, we utilized the "early-stage diabetes risk prediction dataset" from UCI's ML repository, which comprises significant attributes [35, 36]. The Boruta feature selection algorithm was employed for feature selection, leading to the construction of the diabetes risk prediction model using the XGBoost, LightGBM, and random forest ML algorithms.

3. MATERIALS AND METHODS

Figure 1 depicts the proposed ML-based diabetes risk prediction model. The dataset has been preprocessed so that the categorical data has been replaced with integers. The data visualisation was accomplished in order to ascertain the relationship between the attributes and diabetes. The feature selection is accomplished through Boruta feature selection algorithm. Furthermore, up to the feature selection stage, the prediction accuracy has been evaluated using 19 ML classification algorithms. The dataset was split into two segments: the training set and the testing set, distributed in an 80:20 ratio. ML classifiers were used, and the model accuracy was calculated using the confusion matrix. The stability of the top three ML models, which exhibited the highest accuracy, was evaluated through repeated k-fold cross-validation.

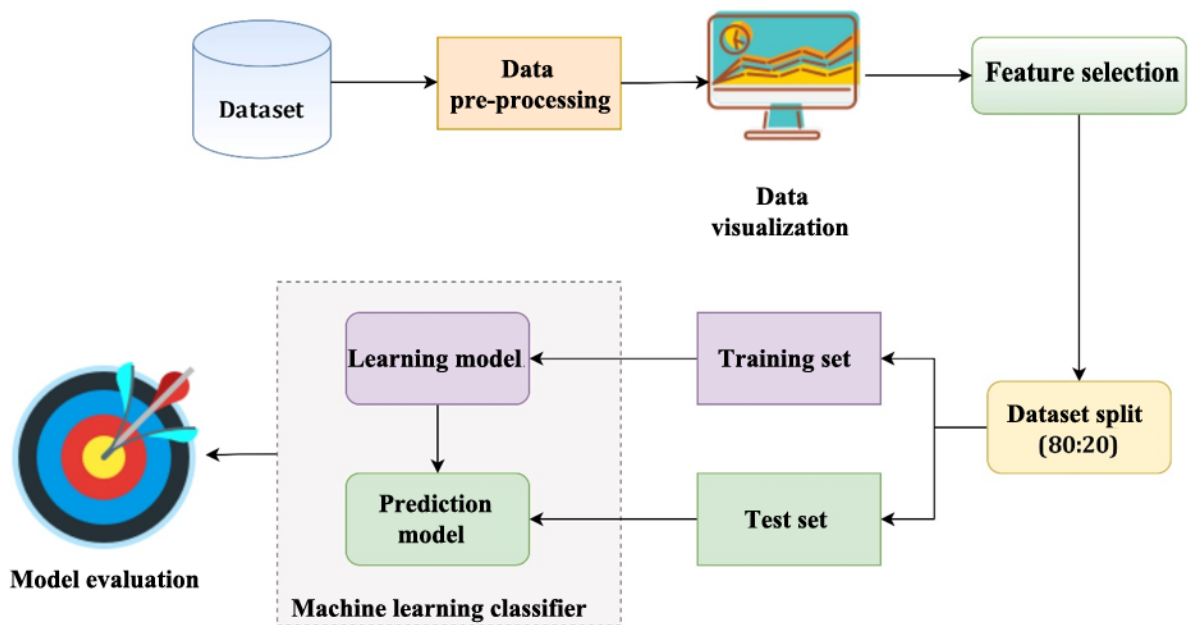


Figure 1. Machine learning based diabetes risk prediction model.

3.1. Data

The publicly available diabetes dataset is collected from UCI's ML repository and is available at [35]. The dataset contains 520 instances with 17 attributes. A medical professional approved the data collection process, which involved conducting direct surveys with patients at Sylhet Diabetes Hospital located in Sylhet, Bangladesh [36]. The description of the features is shown in Table 1. Except for the 'age' attribute, the dataset features are of the qualitative type.

Table 1. Features description.

Attribute name	Data type	Attribute description
Age	Quantitative	Patients age in years
Gender	Qualitative	Male or female
Polyuria	Qualitative	Yes or no
Polydipsia	Qualitative	Yes or no
Sudden weight loss	Qualitative	Yes or no
Weakness	Qualitative	Yes or no
Polyphagia	Qualitative	Yes or no
Genital thrush	Qualitative	Yes or no
Visual blurring	Qualitative	Yes or no
Itching	Qualitative	Yes or no
Irritability	Qualitative	Yes or no
Delayed healing	Qualitative	Yes or no
Partial paresis	Qualitative	Yes or no
Muscle stiffness	Qualitative	Yes or no
Obesity	Qualitative	Yes or no
Class	Qualitative	Positive (1) or negative (0)

3.2. Data Pre-Processing

Preprocessing and feature selection are significant steps in achieving a higher-precision model. The dataset's missing data has been checked, and categorical data has been replaced with integers. The 'Age' feature's quantitative value has not been changed. The qualitative value of the 'Gender' feature has been replaced with 'Male' as '0' and 'Female' as '1'. "Positive" and "negative" qualitative values in the "Class" feature are replaced with 1 and 0, respectively. The other 14 features with 'Yes' and 'No' information have been replaced with '1' and '0', respectively. According to the dataset, there are 320 instances of the diabetes positive class and 200 instances of the negative class.

3.3. Data Visualization

Numerous genetic factors play a role in diabetes. Symptoms of diabetes include polyuria, polydipsia, and significant weight loss. Figure 2 displays the distribution plots of the 16 attributes, highlighting their symmetry. The findings reveal that the features follow a normal distribution, eliminating the need for data normalization techniques on this dataset.

Data insights further indicate that males have a higher infection rate compared to females. Patients presenting symptoms like polyuria, polydipsia, sudden weight loss, weakness, polyphagia, visual blurring, or partial paresis is more prone to diabetes.

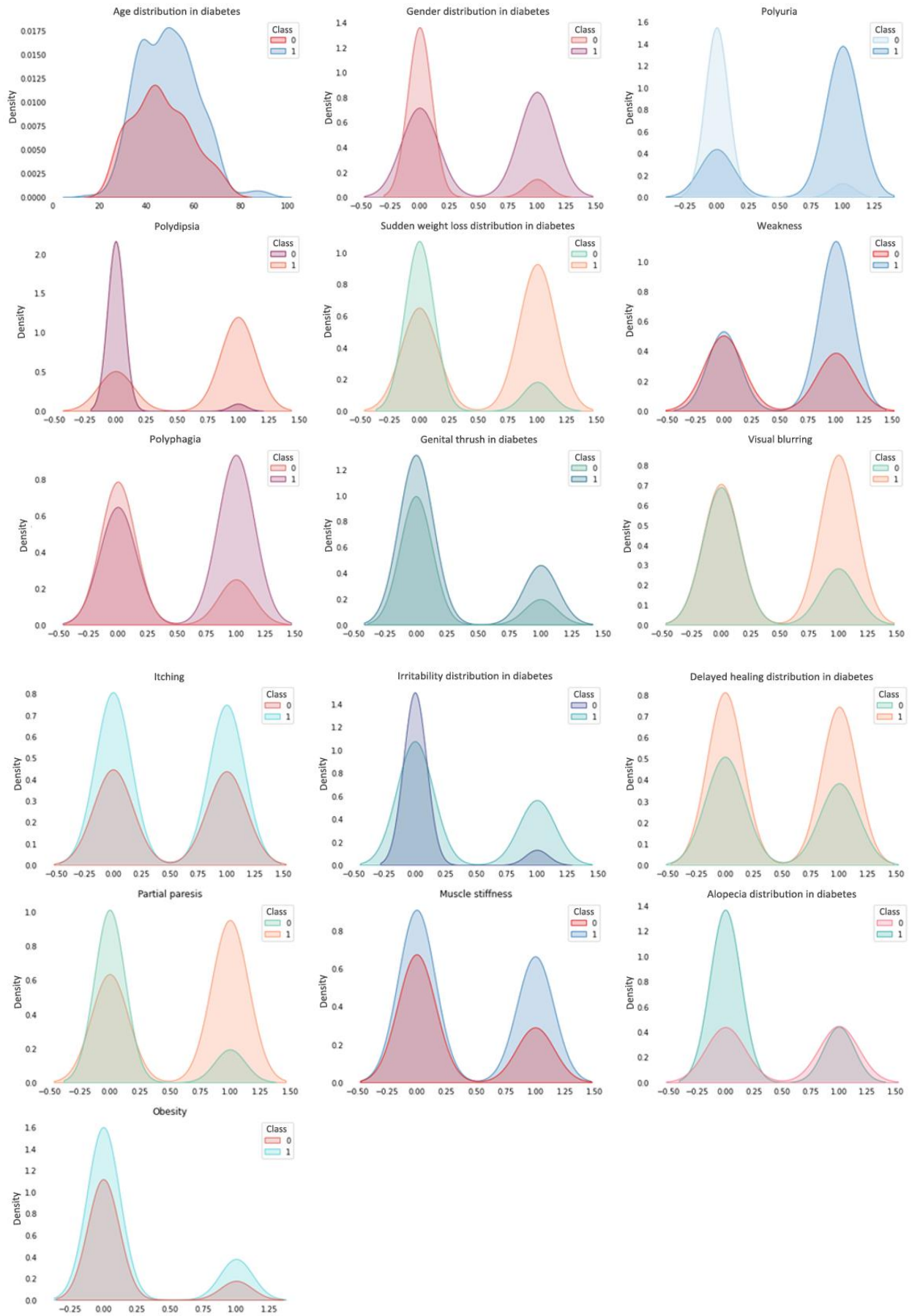


Figure 2. Distribution plots.

3.4. Boruta Feature Selection Algorithm

One of the most essential aspects of knowledge discovery from databases is feature selection. It is described as the method of identifying a subset of the feature set that is more appropriate and insightful for constructing the model. The Boruta feature selector is a wrapper method for feature selection designed to work with ensemble machine learning models like Random Forests. Its primary purpose is to identify and select important features for a predictive modeling task.

Figure 3 displays the selected features using the Boruta feature selector algorithm. Boruta creates a duplicate copy of your feature set, forming a shadow set of features. This shadow set is used for comparison to determine feature importance. To ensure that Boruta is not biased by the original features, it randomly shuffles the values in the shadow features, breaking any existing relationships between the original features and the shadow features.

Boruta performs multiple iterations of feature selection. It trains a Random Forest classifier on both the original features and the shadow features to evaluate the importance of each feature by measuring how effectively it separates the target variable. It then compares the performance of the original features with that of the shadow features. Features that significantly outperform their shadow counterparts are considered important, while those that do not are marked as unimportant and become candidates for removal.

Features are selected based on their performance in separating the target variable, and Boruta keeps track of which features are considered important in each iteration. Boruta continues the iterations until no more features are marked as important.

The ten features listed in the figure are considered important features and are used for model development. The features 'weakness,' 'genital thrush,' 'itching,' 'delayed healing,' 'muscle stiffness,' and 'obesity' have been excluded from the model's development.

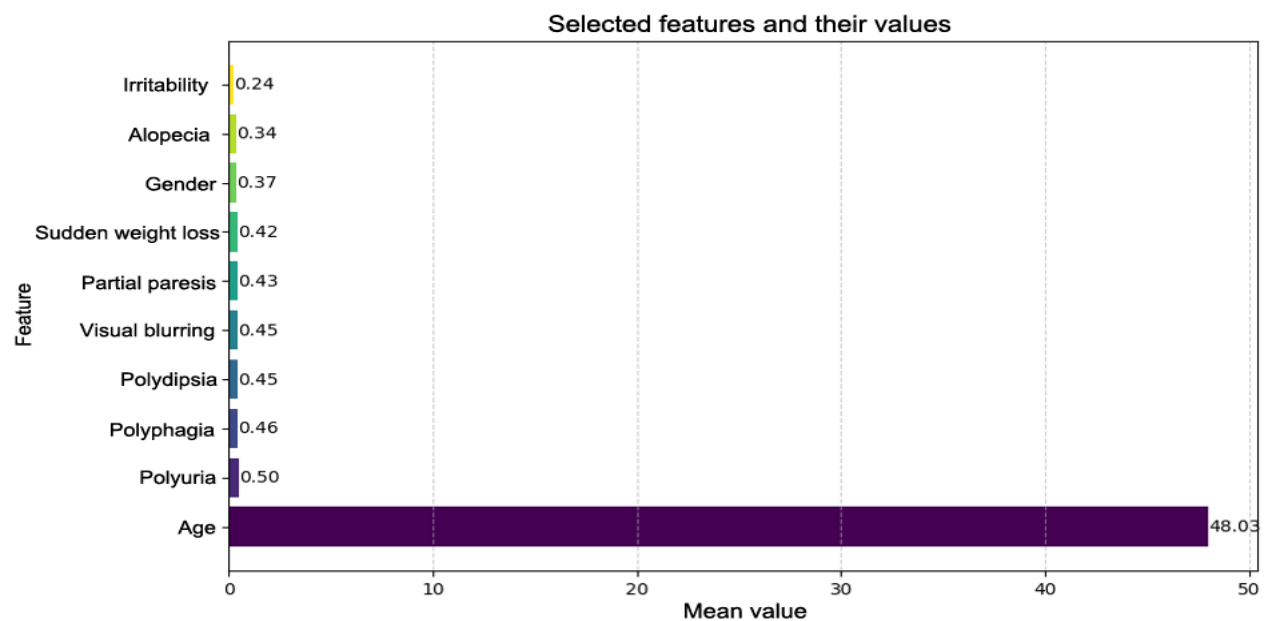


Figure 3. Feature selection using Boruta algorithm.

4. MACHINE LEARNING CLASSIFIERS

In an inductive approach, the framework is designed to "learn" a function referred to as the "target function" in the context of supervised learning. This function represents a model that reflects the data. For this implementation, the dataset has been split into training and test sets with an 80:20 ratio. The training set comprises 416 instances, while the test set contains 104 instances. To predict diabetes, a set of 19 ML algorithms was utilized. The accuracy, AUC-ROC, and F1 score of 16 of these algorithms are presented in Table 2. Notably, the DT and extra tree

classifiers achieved an accuracy of 97% among these 16 algorithms. The remaining three ML algorithms—XGBoost, LightGBM, and Random Forest—will be discussed in more detail.

Table 2. ML classifier accuracy.

ML algorithm	Accuracy	ROC AUC	F1 score
Decision tree classifier	0.97	0.97	0.97
Extra tree classifier	0.97	0.97	0.97
Stochastic gradient descent (SGD) classifier	0.96	0.96	0.96
Quadratic discriminant analysis	0.96	0.95	0.96
Logistic regression	0.95	0.95	0.95
Linear SVC	0.95	0.95	0.95
Calibrated classifier cross-validation (CV)	0.95	0.95	0.95
Nu support vector classification (SVC)	0.94	0.93	0.94
K Neighbors classifier	0.92	0.93	0.92
Linear discriminant analysis	0.93	0.93	0.93
Ridge classifier	0.93	0.93	0.93
AdaBoost classifier	0.94	0.93	0.94
Ridge classifier CV	0.92	0.92	0.92
Passive aggressive classifier	0.91	0.91	0.91
Gaussian Naive Bayes (NB)	0.91	0.9	0.91
Perceptron	0.89	0.88	0.89

4.1. XGBoost

XGBoost algorithm is adopted from Chen and Guestrin [37] and the objective function is defined as:

$$o = \sum_{i=1}^n L(y_i, F(x_i)) + \sum_{k=1}^i R(f_k) + C \tag{1}$$

Where $R(f_k)$ represents the regularization term at the k th iteration, and C is a constant, which can be excluded selectively and $R(f_k)$ is denoted as:

$$R(f_k) = \alpha H + \frac{1}{2} \eta \sum_{j=1}^n w_j^2 \tag{2}$$

Where α denotes the leaves complexity, H represents the number of leaf, η represents the penalty variable, and w_j is each leaf node output result.

4.2. LightGBM

LightGBM [38] operates in the direction of the gradient space G , originating from the input space X . A training set is assumed with instances such as x_1, x_2 , and up to x_n , where every attribute is a vector in the space X with s dimensions. All loss function negative gradients corresponding to the output model represented as g_1, g_2, \dots, g_n in each restatement of a gradient boosting.

Let 'O' represents a set of data for training of a DT, the mean squared error of dividing measure 'j' at a point 'd' is stated as,

$$V_{j|o}(d) = \frac{1}{n_o} \left(\frac{(\sum_{X_i \in O: X_{ij} \leq d} g_i)^2}{n_{l|o}^j(d)} + \frac{(\sum_{X_i \in O: X_{ij} > d} g_i)^2}{n_{r|o}^j(d)} \right) \tag{3}$$

Where $n_o = \sum I[X_i \in o]$, $n_{l|o}^j(d) = \sum I[X_i \in o: X_{ij} \leq d]$ and $n_{r|o}^j(d) = \sum I[X_i \in O: X_{ij} > d]$

4.3. Random Forest

The random forest [39] creates numerous decision trees and combines them to generate predictions that are both more accurate and consistent. The number of estimators is assigned as 50 for the random forest classifier. The node's importance is calculated as,

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \tag{4}$$

The equation represents the impurity value (C_j) of node j , where w_j corresponds to the weighted sample size entering node j . Additionally, $right(j)$ and $left(j)$ denote the child nodes resulting from the right and left splits on node j , respectively.

The relevance of individual attribute is,

$$fi_i = \frac{\sum j: node\ j\ split\ on\ feature\ i\ ni_j}{\sum k \in all\ nodes\ ni_k} \tag{5}$$

4.4. Performance Metrics

The evaluation of the classification model's accuracy is conducted through performance metrics extracted from the confusion matrix. Accuracy, precision, recall, and F1 score can be formally expressed as shown in Equations 6, 7, 8, and 9 correspondingly.

$$\% \text{ Accuracy} = \frac{\text{TruePositive} + \text{TrueNegative}}{\text{TruePositive} + \text{TrueNegative} + \text{FalsePositive} + \text{FalseNegative}} \times 100 \quad (6)$$

$$\text{Precision (p)} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}} \quad (7)$$

$$\text{Recall (r)} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} \times 100 \quad (8)$$

$$\text{f1 score} = \frac{(2 \times p \times r)}{(p + r)} \quad (9)$$

5. RESULTS AND DISCUSSION

Significant conclusions about ML and data mining are drawn from the existing comprehensive review. It's interesting to note that most of the articles that have been published have improved the accuracy of DM prediction classification by more than 80%.

In this work, for initial validation, the number of training set and test set instances is 416 and 104, respectively, with 10 attributes. The number of estimators is assigned as 50 for the random forest classifier. Table 3 displays the performance metrics using Equations 6 to 9 for the three ML classifiers that are proposed. Figure 4 illustrates the confusion matrix for the three models. The ROC curves of the ML classifiers are shown in Figure 5. The ROC plot serves as an assessment tool for gauging the performance of each classification model. Enhanced testing involves reference points clustered towards the upper-left corner of the ROC chart. The accuracy of the model has been computed using Equation 6. The training accuracy of the XGBoost, LightGBM, and Random Forest ML algorithms was 98.79%, 99.52%, and 100%, respectively, with an 80:20 data split. The testing accuracy is 99.33% for all of the models. The AUC values for the XGBoost, LightGBM, and Random Forest ML algorithms were 0.99, 0.99, and 1, respectively. Based on the findings, all three models exhibited strong performance, with the Random Forest model outperforming the other two.

Table 3. ML classifier accuracy.

ML classifier	Accuracy (%)		Precision		Recall		f1-score	
	Training set	Test set	Negative (0)	Positive (1)	Negative (0)	Positive (1)	Negative (0)	Positive (1)
XGBoost	98.79	99.03	1	0.98	0.98	1	0.99	0.99
LightGBM	99.52	99.03	1	0.98	0.98	1	0.99	0.99
Random forest	100	99.03	0.98	1	1	0.98	0.99	0.99

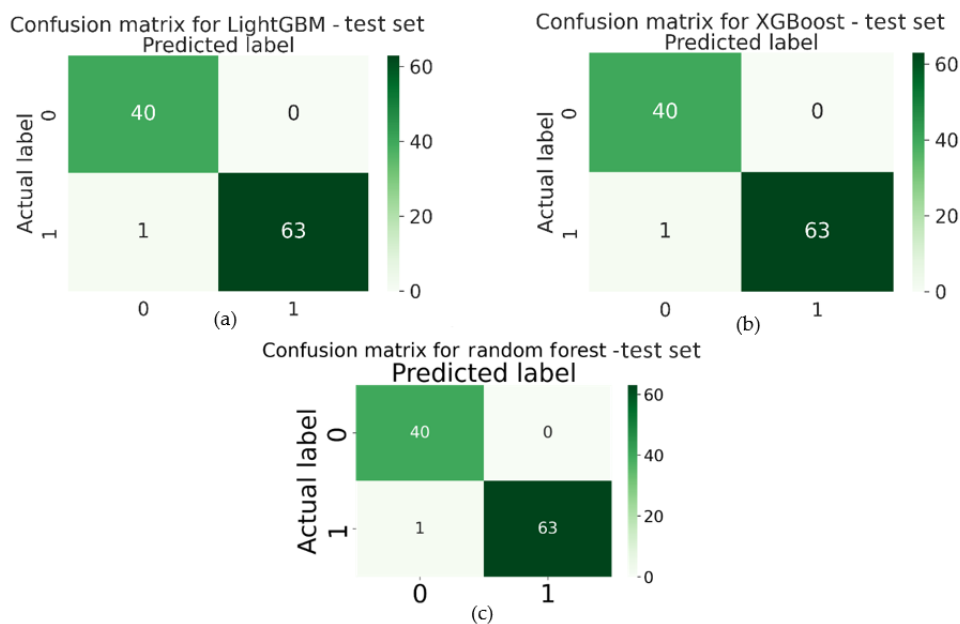


Figure 4. Confusion matrix of ML Classifiers (a) LightGBM (b) XGBoost (c) Random forest.

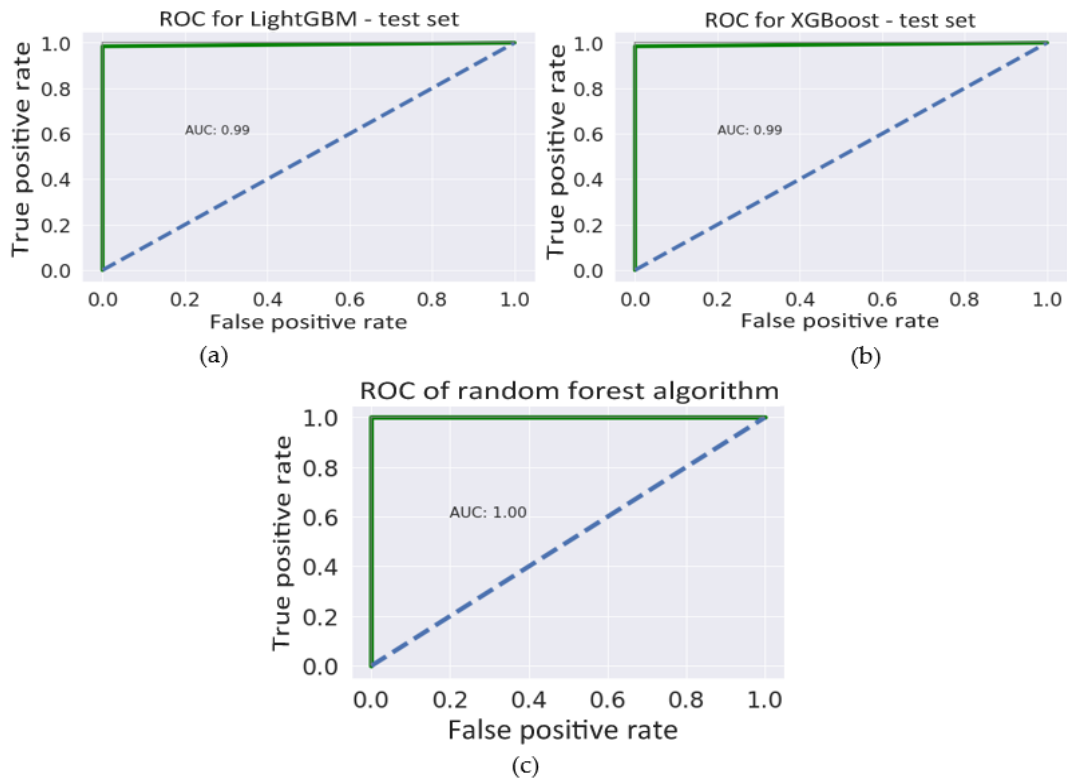


Figure 5. Receiver operating characteristic (ROC) curves of ML Classifiers (a) LightGBM (b) XGBoost (c) Random forest.

5.1. Repeated K-Fold Cross Validation

The standard method for assessing the performance of a ML algorithm on a dataset is by using k-fold cross-validation. Yet, a single run of this method might yield a less precise performance evaluation due to potential variations caused by data splitting. To enhance the reliability of evaluating a ML model, a technique called repeated k-fold cross-validation is utilized. This method uses cross-validation more than once. The average result from all the iterations and folds give a more accurate picture of how well the algorithm actually worked on the dataset, and this evaluation lessens the effect of standard errors. The results of repeated k-fold cross-validation for the LightGBM, XGBoost, and random forest algorithms using the early-stage diabetes risk prediction dataset are shown in Figure 6. The repeated k-fold cross-validation was conducted with 10-fold and 15 repetitions. Table 4 provides the numerical figures of the three ML algorithms. The LightGBM, XGBoost, and random forest classification algorithms provide the mean accuracy of 97.22%, 97.37%, and 98.13%, respectively, with 10-fold and 15 repeats on the dataset.

Table 4. 10-Fold cross-validation with 15 repeats.

No of repeat	Random forest		XGBoost		LightGBM	
	Mean accuracy	Standard error	Mean accuracy	Standard error	Mean accuracy	Standard error
1	0.9827	0.008	0.9769	0.007	0.9731	0.009
2	0.9827	0.004	0.9740	0.005	0.9712	0.006
3	0.9821	0.004	0.9744	0.005	0.9718	0.005
4	0.9793	0.003	0.9736	0.004	0.9712	0.004
5	0.9827	0.003	0.9731	0.003	0.9712	0.003
6	0.9804	0.002	0.9734	0.003	0.9718	0.003
7	0.9802	0.002	0.9734	0.003	0.9720	0.003
8	0.9822	0.002	0.9733	0.002	0.9721	0.003
9	0.9812	0.002	0.9731	0.002	0.9722	0.002
10	0.9813	0.002	0.9727	0.002	0.9721	0.002
11	0.9811	0.002	0.9731	0.002	0.9726	0.002
12	0.9809	0.002	0.9736	0.002	0.9729	0.002
13	0.9812	0.001	0.9734	0.002	0.9728	0.002
14	0.9812	0.002	0.9732	0.002	0.9729	0.002
15	0.9810	0.002	0.9736	0.002	0.9732	0.002
Mean accuracy	0.9813		0.9737		0.9722	

Table 5. Comparison of ML classifier accuracy with existing models.

	ML classifier	Accuracy
Existing model	Type 2 DM – ensemble learning approach [28]	96.74%
	Support vector machine-based model [29]	94%
	GWO - MLP [30]	96%
	APGWO – MLP [30]	97%
Proposed model	XGBoost	97.37%
	LightGBM	97.22%
	Random forest	98.13%

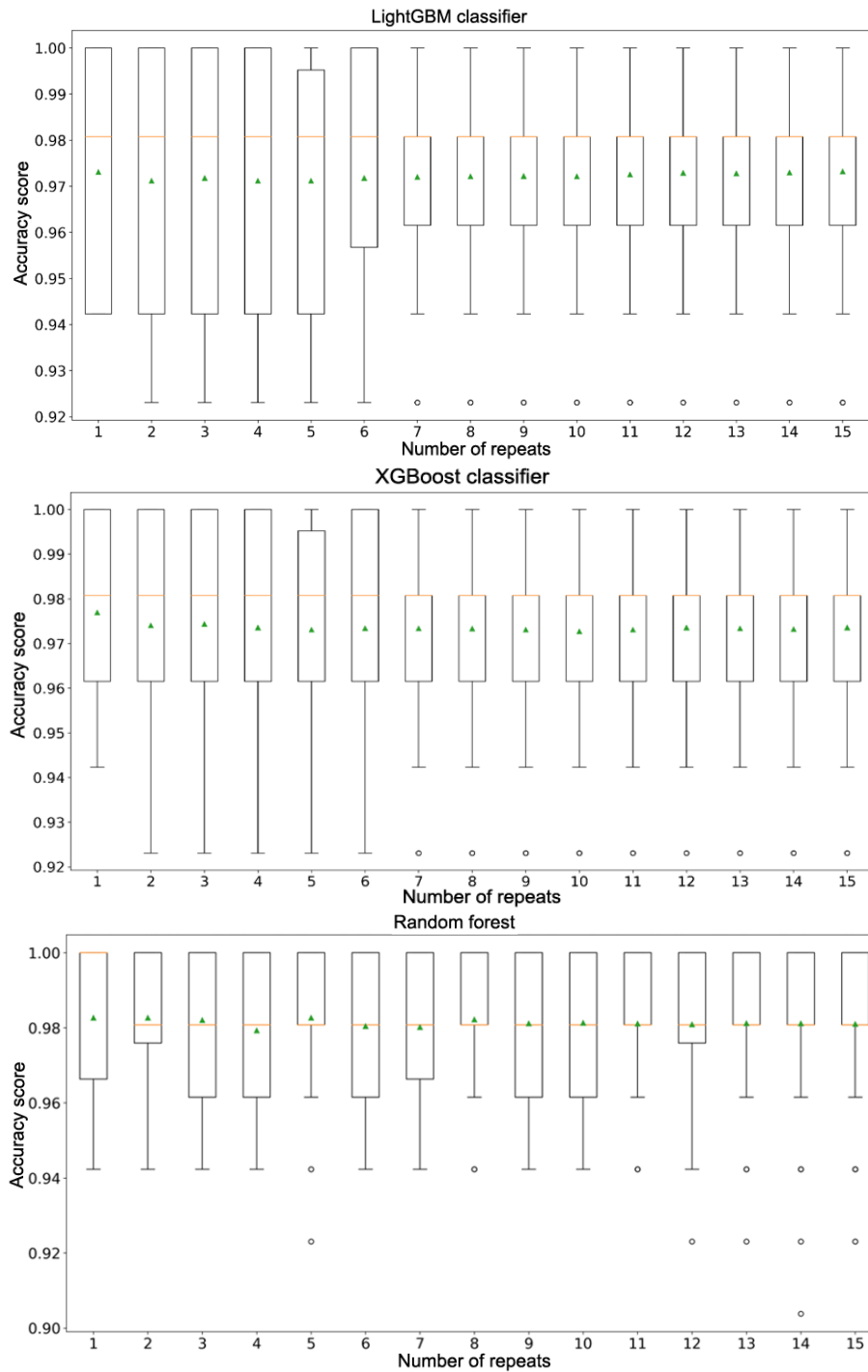


Figure 6. Repeated k-fold cross-validation (a) LightGBM (b) XGBoost (c) Random forest.

Table 5 provides a comparison of accuracy between the newly developed model and existing ones, highlighting the superior performance of the developed model. Specifically, the XGBoost, LightGBM, and Random Forest ML

algorithms demonstrate enhanced accuracy compared to other methods, with the random forest classification algorithm achieving the highest accuracy rate of 98.13%.

6. CONCLUSION

The identification of datasets with significant features aids in the development of the best diabetic risk prediction model. The data insights will be useful in analyzing the features' relationship with the target. The preprocessing and attribute selection methods based on the Boruta feature selection algorithm aided in achieving high-accuracy results. For training and evaluation, 10 features and 520 instances were taken into account. The training accuracy of the XGBoost, LightGBM, and Random Forest ML algorithms was found to be 98.79%, 99.52%, and 100%, respectively, with an 80:20 data split. For all of the models, the test set accuracy is obtained at 99.03 % with an 80:20 data split. The LightGBM, XGBoost, and random forest classification algorithms provide the mean accuracy of 97.22%, 97.37%, and 98.13%, respectively, with 10-fold and 15 repeats on the dataset. According to the results, all three models performed well, with the random forest algorithm gaining better accuracy as well as ROC. This work has the potential to be applied in real-world medical care and be a useful tool for practitioners. In the future, a large dataset can be made, and then a deep learning-based prediction model can be put into place.

Funding: This study received no specific financial support.

Institutional Review Board Statement: Not applicable.

Transparency: The authors state that the manuscript is honest, truthful, and transparent, that no key aspects of the investigation have been omitted, and that any differences from the study as planned have been clarified. This study followed all writing ethics.

Competing Interests: The authors declare that they have no competing interests.

Authors' Contributions: Conceptualization, methodology validation, roles/writing - original draft, K.K.; investigation; methodology supervision; formal analysis, R.M.; data curation; writing - review & editing, T.S.K. All authors have read and agreed to the published version of the manuscript.

REFERENCES

- [1] NIDDK, "What is diabetes?," Retrieved: <https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes>. 2023.
- [2] Z. Tao, A. Shi, and J. Zhao, "Epidemiological perspectives of diabetes," *Cell Biochemistry and Biophysics*, vol. 73, pp. 181-185, 2015. <https://doi.org/10.1007/s12013-015-0598-4>
- [3] WHO, "Diabetes," Retrieved: <https://www.who.int/news-room/fact-sheets/detail/diabetes>. 2023.
- [4] A. Hidouri, S. Jabbour, B. Raddaoui, and B. B. Yaghlane, "Mining closed high utility itemsets based on propositional satisfiability," *Data & Knowledge Engineering*, vol. 136, p. 101927, 2021. <https://doi.org/10.1016/j.datak.2021.101927>
- [5] A. D. Association, "Diagnosis and classification of diabetes mellitus," *Diabetes Care*, vol. 33, no. Supplement_1, pp. S62-S69, 2010.
- [6] G. Kaur *et al.*, "Diagnostic accuracy of tests for type 2 diabetes and prediabetes: A systematic review and meta-analysis," *PLoS One*, vol. 15, no. 11, p. e0242415, 2020. <https://doi.org/10.1371/journal.pone.0242415>
- [7] J. A. D. Silva, E. C. F. D. Souza, A. G. Echazú Böschemeier, C. C. M. D. Costa, H. S. Bezerra, and E. E. L. C. Feitosa, "Diagnosis of diabetes mellitus and living with a chronic condition: Participatory study," *BMC Public Health*, vol. 18, pp. 1-8, 2018. <https://doi.org/10.1186/s12889-018-5637-9>
- [8] A. Fathollahi, F. Daneshgari, and A. T. Hanna-Mitchell, "Effect of polyuria on bladder function in diabetics versus non-diabetics: An article review," *Current Urology*, vol. 8, no. 3, pp. 119-125, 2015. <https://doi.org/10.1159/000365702>
- [9] A. Fournier, "Diagnosing diabetes: A practitioner's Plea: Keep it simple," *Journal of General Internal Medicine*, vol. 15, no. 8, pp. 603-604, 2000. <https://doi.org/10.1046/j.1525-1497.2000.00535.x>
- [10] N. De Fine Olivarius, V. D. Siersma, R. Køster-Rasmussen, B. L. Heitmann, and F. B. Waldorff, "Weight changes following the diagnosis of type 2 diabetes: The impact of recent and past weight history before diagnosis. Results from the Danish Diabetes care in general practice (DCGP) study," *PLoS One*, vol. 10, no. 4, p. e0122219, 2015. <https://doi.org/10.1371/journal.pone.0122219>
- [11] A. Saguil, "Evaluation of the patient with muscle weakness," *American Family Physician*, vol. 71, no. 7, pp. 1327-1336, 2005.

- [12] A. T. Kharroubi and H. M. Darwish, "Diabetes mellitus: The epidemic of the century," *World Journal of Diabetes*, vol. 6, no. 6, pp. 850-867, 2015. <https://doi.org/10.4239/wjd.v6.i6.850>
- [13] L. Mohammed, G. Jha, I. Malasevskaia, H. K. Goud, and A. Hassan, "The interplay between sugar and yeast infections: Do diabetics have a greater predisposition to develop oral and vulvovaginal candidiasis?," *Cureus*, vol. 13, no. 2, p. e13407, 2021. <https://doi.org/10.7759/cureus.13407>
- [14] N. Sayin, N. Kara, and G. Pekel, "Ocular complications of diabetes mellitus," *World Journal of Diabetes*, vol. 6, no. 1, pp. 92-108, 2015.
- [15] H. Brem and M. Tomic-Canic, "Cellular and molecular basis of wound healing in diabetes," *The Journal of Clinical Investigation*, vol. 117, no. 5, pp. 1219-1222, 2007. <https://doi.org/10.1172/jci32169>
- [16] S. Krishnasamy and T. L. Abell, "Diabetic gastroparesis: Principles and current trends in management," *Diabetes Therapy*, vol. 9, pp. 1-42, 2018. <https://doi.org/10.1007/s13300-018-0454-9>
- [17] A. Chobot, K. Górowska-Kowolik, M. Sokołowska, and P. Jarosz-Chobot, "Obesity and diabetes—Not only a simple link between two epidemics," *Diabetes/Metabolism Research and Reviews*, vol. 34, no. 7, p. e3042, 2018. <https://doi.org/10.1002/dmrr.3042>
- [18] M. N. Tajerian *et al.*, "Artemisia: Validation of a deep learning model for automatic breast density categorization," *Journal of Medical Artificial Intelligence*, vol. 4, pp. 1-8, 2021. <https://doi.org/10.21037/jmai-20-43>
- [19] J. SMDAC and G. Ganegoda, "Involvement of machine learning tools in healthcare decision making," *Journal of Healthcare Engineering*, vol. 2021, pp. 6679512-6679512, 2021. <https://doi.org/10.1155/2021/6679512>
- [20] J. N. Kather *et al.*, "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study," *PLoS Medicine*, vol. 16, no. 1, p. e1002730, 2019. <https://doi.org/10.1371/journal.pmed.1002730>
- [21] Y. R. Shrestha, V. Krishna, and G. von Krogh, "Augmenting organizational decision-making with deep learning algorithms: Principles, promises, and challenges," *Journal of Business Research*, vol. 123, pp. 588-603, 2021. <https://doi.org/10.2139/ssrn.3701592>
- [22] N. Kumar, D. Gupta, K. Gupta, and J. Bindra, "Efficient automated disease diagnosis using machine learning models," *Journal of Healthcare Engineering*, vol. 2021, pp. 9983652-9983652, 2021. <https://doi.org/10.1155/2021/9983652>
- [23] F. Q. Kareem, A. M. Abdulazeez, and D. A. Hasan, "Predicting weather forecasting state based on data mining classification algorithms," *Asian Journal of Research in Computer Science*, vol. 9, no. 3, pp. 13-24, 2021. <https://doi.org/10.9734/ajrcos/2021/v9i330222>
- [24] K. Kanagarathinam and K. Sekar, "Text detection and recognition in raw image dataset of seven segment digital energy meter display," *Energy Reports*, vol. 5, pp. 842-852, 2019. <https://doi.org/10.1016/j.egy.2019.07.004>
- [25] K. Sekar, "Power quality disturbance detection using machine learning algorithm," presented at the 2020 IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering (ICADEE), 2020.
- [26] D. Shetty, K. Rit, S. Shaikh, and N. Patil, "Diabetes disease prediction using data mining," presented at the 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017.
- [27] E. I. Georga *et al.*, "Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 1, pp. 71-81, 2012. <https://doi.org/10.1109/titb.2012.2219876>
- [28] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "Development of disease prediction model based on ensemble learning approach for diabetes and hypertension," *IEEE Access*, vol. 7, pp. 144777-144789, 2019. <https://doi.org/10.1109/access.2019.2945129>
- [29] N. Barakat, A. P. Bradley, and M. N. H. Barakat, "Intelligible support vector machines for diagnosis of diabetes mellitus," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 4, pp. 1114-1120, 2010. <https://doi.org/10.1109/titb.2009.2039485>

- [30] T. M. Le, T. M. Vo, T. N. Pham, and S. V. T. Dao, "A novel wrapper-based feature selection for early diabetes prediction enhanced with a metaheuristic," *IEEE Access*, vol. 9, pp. 7869-7884, 2020. <https://doi.org/10.1109/access.2020.3047942>
- [31] A. Singh, A. Dhillon, N. Kumar, M. S. Hossain, G. Muhammad, and M. Kumar, "eDiaPredict: An ensemble-based framework for diabetes prediction," *ACM Transactions on Multimedia Computing Communications and Applications*, vol. 17, no. 2s, pp. 1-26, 2021. <https://doi.org/10.1145/3415155>
- [32] 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. <https://doi.org/10.1109/access.2020.2989857>
- [33] P. Rajendra and S. Latifi, "Prediction of diabetes using logistic regression and ensemble techniques," *Computer Methods and Programs in Biomedicine Update*, vol. 1, p. 100032, 2021. <https://doi.org/10.1016/j.cmpbup.2021.100032>
- [34] C. V. Raghavendran, G. Naga Satish, N. Kumar Kurumeti, and S. M. Basha, "An analysis on classification models to predict possibility for type 2 diabetes of a patient," in *Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2021, 2022*: Springer, pp. 181-196.
- [35] K. Kanagarathinam, "Early stage diabetes risk prediction dataset," *IEEE Dataport*, 2021. <https://dx.doi.org/10.21227/k01r-x481>
- [36] M. M. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, *Likelihood prediction of diabetes at early stage using data mining techniques. In Computer Vision and Machine Intelligence in Medical Image Analysis*. Singapore: Springer, 2020.
- [37] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," presented at the Proceedings of the 22nd acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016.
- [38] G. Ke *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, vol. 30, pp. 3149-3157, 2017.
- [39] A. Sarica, A. Cerasa, and A. Quattrone, "Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: A systematic review," *Frontiers in Aging Neuroscience*, vol. 9, p. 329, 2017. <https://doi.org/10.3389/fnagi.2017.00329>

Views and opinions expressed in this article are the views and opinions of the author(s), Review of Computer Engineering Research shall not be responsible or answerable for any loss, damage or liability etc. caused in relation to/arising out of the use of the content.