Alexandria Engineering Journal 94 (2024) 311–324

Contents lists available at ScienceDirect

Alexandria Engineering Journal

journal homepage: www.elsevier.com/locate/aej

Original article

An optimized diabetes mellitus detection model for improved prediction of accuracy and clinical decision-making

Turke Althobaiti [a,b], Saad Althobaiti [c], Mahmoud M. Selim [d,*]

[a] Department of computer science, Faculty of Science, Northern Border University, Arar, Saudi Arabia
[b] Remote Sensing Unit, Northern Border University, Arar, Saudi Arabia
[c] Department of Sciences and Technology, Ranyah University Collage, Taif University, PO Box 11099, Taif 21944, Saudi Arabia
[d] Department of Mathematics, College of Science and Humanities, Prince Sattam bin Abdulaziz University, Al-Kharj 16273, Saudi Arabia

ARTICLE INFO

Keywords:
Diabetes mellitus
GBM-DRU network
Feature engineering
Ensemble learning
Clinical decision-making

ABSTRACT

Diabetes Mellitus (DM) is an enduring metabolic illness that disturbs many individuals globally. This study addresses the global impact of Diabetes Mellitus (DM) and emphasizes the critical role of accurate DM detection in early diagnosis, effective treatment, and prevention of complications. The research introduces an optimized DM detection model, the GBM-DRU (Gradient Boosting Machine - Data Reduction Unit) network, which integrates feature engineering and ensemble learning techniques to enhance prediction accuracy and support clinical decision-making. The GBM-DRU network combines the powerful gradient boosting machine algorithm with a data reduction unit (DRU) for efficient feature selection, reducing dimensionality and improving computational efficiency. Feature engineering enhances discriminatory power, while ensemble learning methods, including bagging and boosting, improve overall model performance. Rigorous experiments on a comprehensive dataset of DM patients demonstrate that the proposed approach outperforms existing models in terms of accuracy, sensitivity, specificity, and AUC-ROC. The optimized model provides valuable insights into feature importance, aiding clinical decision-making and deepening the understanding of DM risk factors. Therefore, the GBM-DRU network, utilizing feature engineering and ensemble learning, presents a viable approach to precise diagnosis of diabetes mellitus, with favorable implications for patient outcomes, disease control, and public health campaigns. The improved prediction accuracy, feature interpretability, and clinical decision support capabilities of the model may have a beneficial effect on public health campaigns, disease management, and patient outcomes.

1. Introduction

Today, the most severe chronic illness that can result in other serious aggravating illnesses is diabetes. People with diabetes face serious health problems. When blood sugar is very high, diabetes develops. Essentially, blood sugar serves as a person's main source of energy. However, having too much insulin in the circulatory system over an extended period might be unhealthy. However, mellitus can't be cured entirely because routine treatments well manage it. Consequently, the most prominent study field in the biological sciences is diabetic prognosis [1]. The predictive analysis technique can make meaningful choices and projections about healthcare data. ML can be used for predictive analytics as an alternative to regression. Predictive data analysis strives to improve medical results by increasing care for patients, optimizing the utilization of resources, and making disease diagnoses as precise as feasible. To learn more about the risk causes and detection of diabetes and pre-diabetes, many research investigations have been conducted. However, only a few research were done to determine the risk variables for diabetes complications precisely. Due to this, diabetic-related disorders continue to be underutilized in preventing illness, and frequently they become apparent once the disorder manifests in a harmful condition.

In the realm of healthcare, machine learning algorithms play a crucial role in analysing data collected by connected devices, contributing to efficient and precise assessments of patients' well-being [2]. Data mining techniques, including segmentation approaches, prove to be effective in recognizing anomalies and predicting diseases. The integration of contemporary connected devices and sensors in medicine

* Corresponding author.
E-mail address: m.selim@psau.edu.sa (M.M. Selim).

https://doi.org/10.1016/j.aej.2024.03.044
Received 21 October 2023; Received in revised form 14 February 2024; Accepted 17 March 2024
Available online 28 March 2024

has led to the establishment of numerous health monitoring and intelligent healthcare strategies. The prevalence of gestational diabetes mellitus (GDM), characterized by glucose intolerance during pregnancy, varies globally, ranging from 1% to over 30%. Regions such as the Western Pacific and certain Middle Eastern nations exhibit higher rates of GDM, with Chinese women being particularly susceptible in the Western Pacific. Additionally, Asian women, in general, face a higher vulnerability compared to women from diverse ethnic backgrounds [3].

Despite the large amount of medical data generated daily in healthcare, it is valuable for predicting diseases and especially helpful when customizing treatments based on patients' medical history and prior therapies [4], since health-related data is utilized subsequently to make wise decisions about the well-being of patients, that will conceal certain aspects of the details; additionally, this field has to be developed further using instructive healthcare data. In recent years, classification, tool proof of identity, and forecasting have been the three main applications of machine learning (ML) techniques, including artificial neural networks (ANN) and tree classification [5]. Therefore, these machine learning methods are utilized for the accurate identification and prognosis of illnesses in healthcare, including liver disease, inflammation, persistent illnesses of the heart, kidneys, Parkinson's, and Alzheimer's, and in patients with HIV to predict therapeutic responsiveness. Thus, chronic diseases and their circumstances could become a global health and economic problem.

A total of 415 million individuals worldwide had diabetes in 2015, based on data from the International Diabetes Federation (IDF) [6]. Diabetes has become the leading cause of death in many countries, surpassing heart disease. It is a group of metabolic disorders characterized by high blood glucose levels due to insufficient insulin manufacture by the body or the body's inability to use insulin effectively. Individuals with diabetes are at a higher risk of developing various chronic conditions affecting the heart, nervous system, eyes, feet, and more if their blood sugar levels are not well-managed. There are two main types of diabetes: Type-1 Diabetes Mellitus (T1DM) and Type-2 Diabetes Mellitus (T2DM). Type-1 diabetes, also known as glucose-dependent diabetes, is an autoimmune condition where the body's immune system damages pancreatic beta cells, leading to irregular insulin manufacture [7]. The condition known as type 2 diabetes mellitus (T2DM), also called noninsulin-dependent diabetes, is caused by insulin resistance, a condition in which the body's cells are unable to utilise insulin as intended. Insulin resistance and decreased pancreatic secretion of insulin combine to cause elevated blood glucose levels in this condition [8]. For the sake of people with diabetes who want to regulate the blood glucose (BG) quantity constantly, exogenous insulin delivery is necessary. Moreover, persons with T2DM who have become insulin-resistant need to perform self-blood pressure monitoring (SMBG). The BG standards are not random since the BG history has a workable design. Forecasting prospective BG values using data from the previous BG values gathered by Continuous Glucose Monitoring (CGM) is feasible. Hence, conventional statistical techniques have been used in most earlier investigations on BG models for prediction. However, these approaches fail to consider the nonlinear connection among BG levels using models based on machine learning that has primarily been carried out.

However, research on forecasting blood glucose levels has been published in the last five years [9]. Furthermore, convolutional neural networks (CNN), recursive neural networks (RNN), and variants of each of these system replicas have been seen to be often utilized, and it is possible to get estimates that have a small root mean square errors (RMSE). Even though the disease has long been understood to be a consequence of diabetes mellitus, epidemiological studies have not consistently found a link between diabetes mellitus and infection. Moreover, main education measuring the cost of infection-related problems in patients using diabetes mellitus and examining the particular pathogens responsible for this burden was only conducted in the last ten years [10].

Traditional diabetes prediction algorithms mainly depend on statistical models and fundamental machine learning methods like logistic regression or decision trees. While these techniques have been useful in some cases, they frequently lack the accuracy needed for fast and accurate diagnosis, which puts patients at risk for health problems. Deep learning models and ensemble learning strategies are only two of the more complex medical diagnostics methods developed due to recent developments in artificial intelligence and machine learning. Convolutional neural networks (CNNs) [11] and recurrent neural networks (RNNs) [12], in particular, have shown promise in time-series data processing and medical image analysis, which are relevant to diabetes detection. Additionally, by integrating the advantages of many models, ensemble learning techniques like gradient boosting and random forests have demonstrated better prediction accuracy. However, a thorough examination of these cutting-edge methods in the context of diabetes diagnosis is currently lacking, calling for more study and development to realize their full potential. This study aims to bridge this gap by introducing the GBM-DRU Network, a unique ensemble learning strategy specifically designed for diabetes prediction, and merging it with feature engineering techniques to improve accuracy and clinical decision-making in diabetes diagnosis.

This paper uses an optimized GBM-DRU network with feature engineering and ensemble learning techniques to detect DM disease, enhancing accuracy and clinical decision-making. The Gradient Boosting Machine (GBM) will allow the generation of predictions that are out of the data. In contrast, the Data Reduction Unit (DRU) will increase storage efficiency and performance. Moreover, this will also reduce the storage efficiency cost.

1.1. Key contribution

The key contribution of the established outline is briefed as follows:

- Introduction of a novel approach, the GBM-DRU network, combining Gradient Boosting Machine (GBM) with a Data Reduction Unit (DRU).
- Implementation of sophisticated feature engineering techniques to enhance the model's ability to capture relevant information for diabetes detection.
- Utilization of ensemble learning methods, such as bagging and boosting, to improve overall model performance and robustness.
- Rigorous evaluation of the model's performance on a diverse dataset, using key metrics like accuracy, sensitivity, specificity, and AUC-ROC, with comparative analyses against existing models for validation.

Section 1 provide the introduction continue by related work at Section 2, formulation of the problem is presented in Section 3 which provide the proposed method and then result and discussion is presented in Section 4 and finally the paper is concluded in Section 5.

2. Related works

Rufo et al. [13] discovered a significant increase in diabetes mellitus (DM) worldwide. A serious chronic condition that significantly affects people's well-being, and as reported in [13], 50% of all diabetics worldwide remain undiagnosed. As DM's consequences rise to high computational time, new research possibilities and issues will also arise. This research advocates a preemptive analysis approach to enhance early diabetes mellitus (DM) detection, particularly beneficial in regions with limited healthcare professionals. The proposed method focuses on implementing strategies to identify individuals at risk or affected by DM efficiently. This proactive approach is crucial for addressing healthcare resource challenges and improving health outcomes in areas with lower medical professional densities. The Zewditu Memorial Hospital (ZMHDD) in Addis Ababa, Ethiopia, is the source of the diabetes data.

Moreover, the only fresh and positive investigation result using the gradient boosting framework (GDM) that uses tree-based learning procedures is the Light Gradient Boosting Machine (LightGBM). Because of its squat figuring complication, it is appropriate for use in regions with constrained capability, like Ethiopia. As a result, a precise model for recognizing the presence of diabetes has been developed in this work using the LightGBM approach. According to the results of the experiment, the produced diabetic database can be used to predict the occurrence of diabetes mellitus. The LightGBM model outperforms KNN, SVM, NB, Bagging, RF, and XGBoost in the circumstance of the ZMHDD dataset with accuracy, AUC, sensitivity, and specificity of 98.1%, 98.1%, 99.9%, and 96.3%, correspondingly. Healthcare organizations offer individualized care in various settings to assist patients in integrating into their daily routines, said Ali et al.[14]. This work employs an enhancement team approach with an internal randomized group encoder to foresee the kind of diabetes that patients will have depending on their individual and medical information. Moreover, a genuine data set with 100 records assesses the proposed AdaboostM1 method. Based on Weka studies with a ten-fold cross-validation, the predicted accuracy reached is 81.0%. The key flaw of this paper is that it is challenging to anticipate the existence of diabetes.

Diabetes mellitus, a metabolic illness marked by high blood sugar, is a major global public health issue disclosed by Islam et al.[15]. Most instances can be decreased by detecting diabetes at an extremely young age. Therefore, the primary aim of this investigation is to create a scheme grounded on machine learning that will significantly improve diabetic patient detection. Researchers employed a dataset created by collecting direct questions from Sylhet Diabetic patients at the hospital to design such a system. The dataset includes details on the early symptoms and indicators of people newly diagnosed with diabetes or likely to have it. The Gradient Boosting Machine (GBM) surpassed the other algorithms' best F1 and ROC scores of 99.37% and 99.92%, respectively. Type 2 diabetes is frequently not as dangerous as type 1. Nevertheless, the biggest drawback is that it can also lead to health problems, notably in small blood vessels in the kidneys, nerves, and eyes.

Zhang et al. [16] explores how machine learning (ML) can enhance judgment by examining intricate relationships between many factors with the advent of data mining. Moreover, investigators concentrate on the overall 36,652 suitable contributors from the Henan Rural Cohort Study to investigate the capability of ML methods to forecast the danger of type 2 diabetes mellitus (T2DM) in impoverished Chinese inhabitants. There were six ML computations together with logistic regression (LR), classification regression tree (CART), artificial neural networks (ANN), support vector machine (SVM), random forest (RF), and gradient boosting machine (GBM) used to build models for risk assessment for T2DM. Henceforth, the region over the handset's functioning representative arc, sensitivity, specificity, a positive assessment for prediction, undesirable prognostic rate, and zone below the exactness recollection arc were used to assess the procedure's presentation. Novel methodologies are used to overcome the issue of poor accuracy.

Zhao et al. [17] tends to discuss the predictive significance of integrating the depth of the bronchial and renal basement membrane in diabetic nephropathy (DN)in this paper. Furthermore, around 110 people having type 2 diabetes with operation had confirmed that DN were enrolled in this retrospective analysis between 2011 and 2018. The Renal Pathology Society designations were used to verify the pathological results. As a result, Haas's direct arithmetic average approach and the oblique interception technique were used to calculate the GBM and TBM thickness. Cox proportional hazard modeling was utilized to determine the hazard ratios (HRs) for the effect of mixed GBM and TBM thicknesses to forecast end-stage renal disease (ESRD).

Ahamed, Arya, and Nancy [18] explore the fact that the technological advancements in the medical sector have led to many fresh findings in machine learning. The beginning of numerous illnesses is detected, and trends for illness identification are employed. Diabetes mellitus, fatal cardiac circumstances, and indicative malignancy are among the diseases. Several different procedures have been vital in the forecasting of illnesses. This investigation suggests a machine learning (ML)–based strategy for forecasting diabetic mellitus illness. Moreover, the three efficient classification algorithms like RF, GBM, and LGBM are examined and applied in the suggested investigation. Furthermore, two dissimilar categories of datasets are second hand. They are an accurate dataset of Pima Indians. The precision of every classifier is dignified and compared as an extrapolation module utilizing the ML classifiers LGBM, GB, and RF. Therefore, the total prediction precision is attained for the classifiers LGBM, GB, and RF using the generalized predictive method and the data augmentation method. To further increase the algorithm's precision for forecasting diabetic disease, a comparison among expansion and non-augmentation is also explored for the binary datasets employed. Damage to the kidneys, pregnancy issues, weakening of blood flow, and skin issues are just a few of the type 1 risks and downsides. Unique methods are applied to resolve these problems.

A person with diabetes mellitus (DM) is thought to have a long-lasting illness brought on by having too much sugar in their blood disclosed by Ahamed, Arya, and Nancy [18]. If the disease is left untreated, it can result in serious health concerns and linked disorders like heart attack, damage to the nervous system, issues with feet, kidney and liver destruction, and eye problems. Moreover, gender, chronological age, bloodline, BMI, and blood sugar levels are just a few of the elements that contribute to these issues. Thus, different Machine-Learning (ML) procedures are utilized to anticipate and diagnose the condition to prevent additional health concerns. The kind of diabetes that a person has and the likelihood of developing linked conditions can be used to improve the diabetes prediction process. A pair of datasets, Pima Indians and an experimental assessment dataset, is employed in the study to carry out the task above. Hence, different machine learning (ML) methods are employed, including Random Forest, Light Gradient Boosting Machine, Gradient Boosting Machine, Support Vector Machine, Decision Tree, and XGBoost. Moreover, the performance measurements used here are accuracy, precision, recall, specificity, and sensitivity. Also, methods like data expansion and specimen are utilized in this paper.

An 81-year-old gentleman who had a 30-year past of diabetes mellitus (DM), diabetic retinopathy, and diabetic neuropathy had his renal histology studied. Angiotensin receptor blockers were used with other antihypertensive drugs to keep the patient's blood pressure below the standard range (a value below 140/75 mmHg), according to Shima et al. [19]. Moreover, a rigorous salt intake (no more than 6 g per day) was used to manage edema. Nevertheless, this patient had poor glycemic control with an HbA1c of 8–10%. This is the disadvantage that the investigator disclosed in this paper. The estimated rate of glomerular filtration (eGFR) was 64 ML/min/1.73 m2, and serum creatinine was 0.87 mg/dL. Henceforth, protein production in the urine was 1.5 g per day. Furthermore, the Immunofluorescence microscopy of the renal sample from this patient exhibited linear IgG stains along the GBM. However, light microscopy indicated mostly unbroken glomeruli and the GBM was not enlarged, as demonstrated by microscopy using electrons with a breadth of 288–368 nm (430 nm). Moreover, polar vasculitis was seen near the glomerular vascular pole, and significant arteriolar hyalinosis. This instance shows that persistent hyperglycemia can cause polar vasculitis through angiogenesis. At the same time, diabetic glomerulopathy is limited to a modest change when hypertension and swelling are properly managed.

The surveyed literature contributes to the proposed work by offering valuable insights and methodologies for diabetes detection and management. Rufo et al. [13] emphasize the importance of early diabetes detection in regions with limited healthcare resources and introduce Light Gradient Boosting Machine (LightGBM) for accurate prediction. Ali et al. [14] introduce an ensemble approach using AdaboostM1 to predict diabetes types, while Islam et al. [15] focus on using a Gradient Boosting Machine (GBM) to improve patient detection and emphasize the significance of early diagnosis. Zhang et al. [15] discuss using

machine learning algorithms for risk assessment in diabetes and their potential in impoverished populations. Zhao et al. [16] explore the relevance of depth measurements in diabetic nephropathy prediction using gradient-boosting methods. Ahamed, Arya, and Nancy [17] propose using various machine learning classifiers and data augmentation techniques for diabetes prediction and provide insights into the associated risks. Shima et al. [18] present a case study on the impact of persistent hyperglycemia and its association with renal complications, highlighting the importance of glycemic control. These studies offer diverse approaches and methodologies that can inform and enhance the development of an optimized diabetes mellitus detection model using the GBM-DRU Network with feature engineering and ensemble learning techniques.

3. Formulation of the problem

The description of the issue is to tackle the urgent challenge of reliably predicting diabetes in patients for better clinical decision-making. To this end, an optimized Diabetes Mellitus detection model will be developed utilizing the GBM-DRU Network, feature engineering, and ensemble learning approaches. Public health is significantly impacted by the common and chronic medical illness known as diabetes mellitus. Current models frequently lack clinical usefulness and accuracy [20]. As a result, this research aims to improve prediction accuracy by implementing the GBM-DRU Network, a unique algorithm, together with cutting-edge feature engineering and ensemble learning techniques. By doing this, it hopes to develop a solid and trustworthy tool to help medical professionals make better clinical judgments, ultimately leading to improved management and treatment of diabetic patients [17]. The suggested technique is chartered in Fig. 1. An optimized GBM-DRU (Gradient Boosting Machine - Data Reduction Unit) with Feature Engineering and Ensemble Learning Technique is used to enhance prediction accuracy and facilitate clinical decision-making in this process. Hence, many datasets are presented and used for training and testing purposes. This paper uses a dataset from the Zewditu Memorial Hospital Diabetes Dataset (ZMHDD) to discuss the medical data of diabetes patients to find accurate coding for the data. Then, the pre-processing is done, which will collect the missing data from the database and remove the outliers using the Min-Max Normalization method. Moreover, the topographies such as glucose, pregnancy, blood pressure, skin thickness, insulin, BMI, diabetes, and age are selected using a Data Reduction Unit (DRU). Finally, the prediction of the present or absent DM in patients' bodies is employed by a Gradient Boosting Machine (GBM), which will provide improved predictions and clinical decision-making processes [21].

3.1. Data collection

The dataset used for the testing and training is taken from the Zewditu Memorial Hospital Diabetes Dataset [22]. Approximately the data from 768 patients' pregnancy, glucose, blood pressure, skin thickness, insulin, BMI, diabetes, and age were collected. In this paper, the data records of ten patients are shown in the tabulation below. In the analysis, the dataset comprising information from 768 patients was split, with 70% of the data utilized for training the model and the remaining 30% reserved for testing its predictive performance. The choice of the Zewditu Memorial Hospital Diabetes Dataset (ZMHDD) over other datasets can be justified for several reasons. Firstly, the dataset's source, Zewditu Memorial Hospital in Addis Ababa, Ethiopia, may reflect a population with unique demographic and healthcare characteristics, making it particularly relevant for addressing diabetes detection challenges in regions with limited healthcare resources. Secondly, the dataset's utilization aligns to develop a diabetes detection model that can be applied effectively in areas with constrained capability, potentially filling a critical gap in healthcare infrastructure. Moreover, by selecting ZMHDD, it may have access to a fresh and unique dataset that hasn't been extensively used or explored in previous research, allowing for novel insights and outcomes in diabetes prediction. These reasons collectively support the ' decision to choose ZMHDD as a suitable and valuable dataset for their research on optimizing diabetes detection.

3.2. Pre-processing

Min-max normalization, or feature scaling, is a data preprocessing technique widely employed in machine learning and data analysis. Its primary purpose is to transform numerical data into a standardized range, typically between 0 and 1, by rescaling the values proportionally to their minimum and maximum values within a given dataset or feature. This normalization process is particularly valuable when dealing with features that have different units or varying scales because it ensures that all features contribute equally to the modeling process, preventing dominant features from overshadowing others [23].

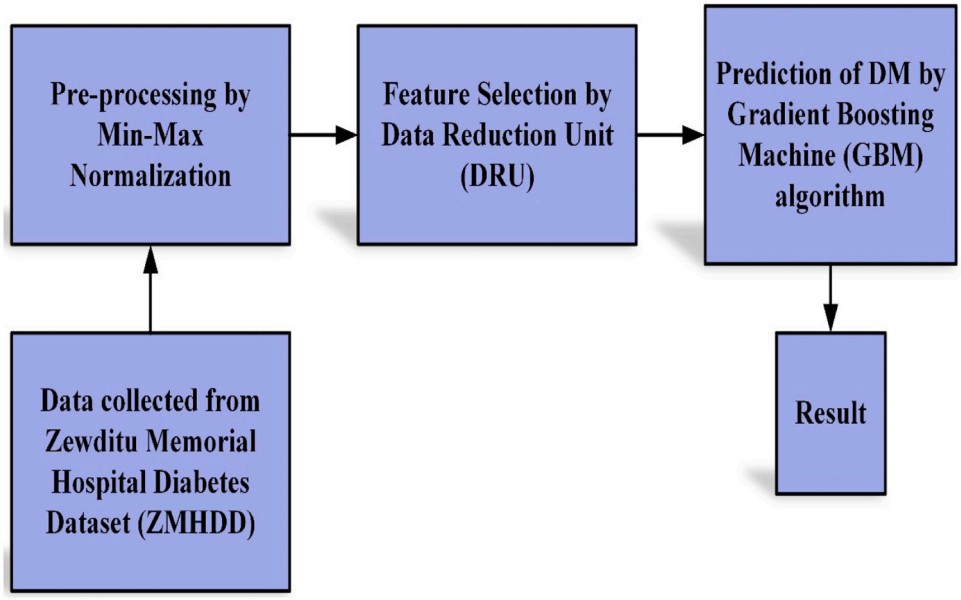

Fig. 1. Optimized GBM-DRU method.

Min-max normalization not only enhances the model's convergence during training but also improves its stability and performance by making it less sensitive to the magnitude of input data. The motivation for normalization in data preprocessing is to bring all numerical features within a consistent and standardized range, typically between 0 and 1, to ensure that they contribute equally to machine learning models. This process eliminates the potential bias that can arise from varying scales and units, enhancing the model's convergence, stability, and overall predictive performance by making it less sensitive to the magnitude of input data [21].

The data preparation involved several processes. This pre-processing method will eliminate the outliers and collect the missing data from the database [24]. When employing min-max normalization, the data were transformed and are given in Eq. (1),

The actual data n is transformed linearly by Min-Max Normalization into the desired interval$\min_{new}, \max_{new}$.

$$n = \min_{new} + (\max_{new} - \min_{new}) * (\frac{n - \min_x}{\max_x - \min_x}) \quad (1)$$

3.3. Feature selection using data reduction unit

A method for locating and displaying the combination of separate variables in some way that have an impact on the dependent variables is called a Data Reduction Unit (DRU). Its main purpose is to discover how different variables connect and how that could impact the result of the system [25]. Moreover, it is superior to conventional systems since it is independent of the model's variables and kind. Although it merges two or more qualities to create one characteristic. The dimensional representation that holds the data that is altered by this transformation. As a result, the algorithm performs better when forecasting a particular class variable. Artificial intelligence uses several DRU enhancements to predict pregnancy, glucose, blood pressure, skin thickness, insulin, BMI, diabetes, and age.

A Data Reduction Unit can aid in the crucial step of feature selection in the development of an optimised diabetes mellitus detection model for better prediction accuracy and clinical decision-making. By using advanced techniques, this unit streamlines the input variables for the model by identifying and retaining the most pertinent features from the dataset. To improve the efficiency and interpretability of the model, the Data Reduction Unit makes sure that only the most informative features are included by applying techniques like principal component analysis (PCA) and recursive feature elimination (RFE). By highlighting the critical elements affecting the diagnosis of diabetes, this enhanced feature set supports clinical decision-making in addition to increasing prediction accuracy. The model's utility in clinical settings is ultimately enhanced by the integration of a Data Reduction Unit, which makes it more reliable, computationally efficient, and better suited to the unique needs of diabetes detection.

3.4. Prediction of DM by ensemble learning technique

An effective machine learning method called ensemble learning combines a group of instructors rather than an individual to predict

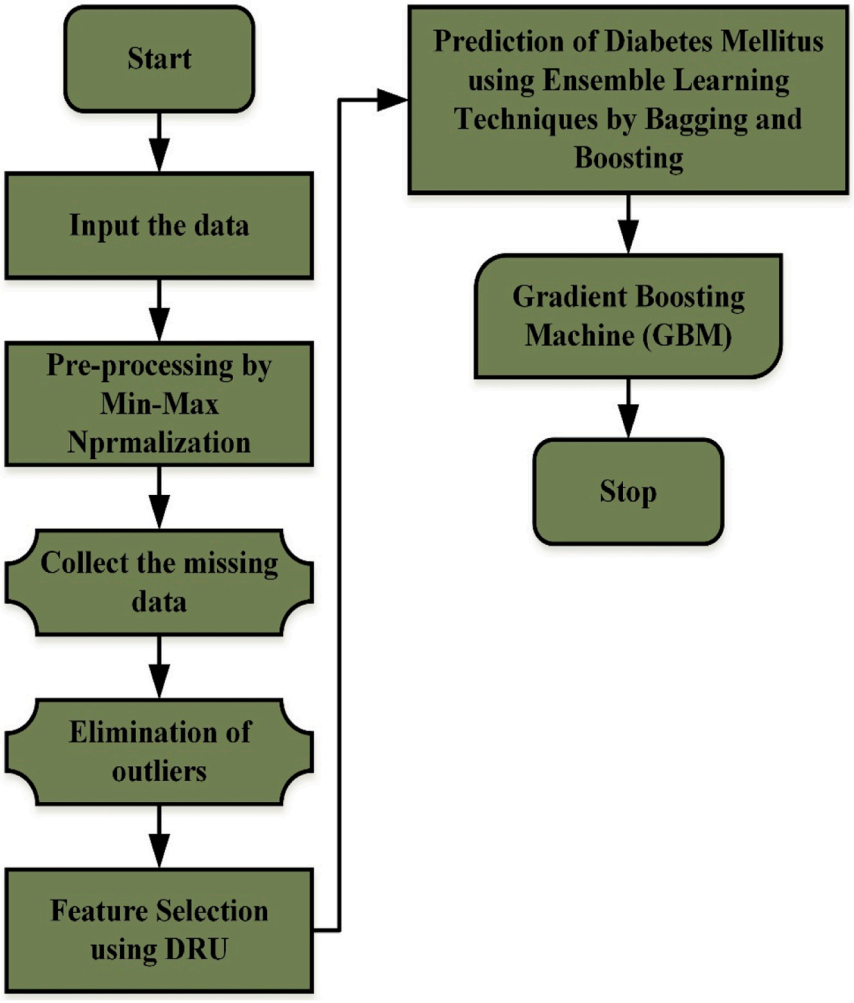

Fig. 2. Flowchart of GBM-DRU model.

unidentified objective qualities. In this framework, a voting system is used to integrate every result from each student and determine the final group's labeling projection. Moreover, the fundamental objective of ensemble instruction is to develop a powerful classifier made up of several learners to get outcomes for classification that tend to be more precise. The four categories of group learning methods are bagging, boosting, voting, and stacking [26]. In this work, the commonly used ensemble learning methods of bagging and boosting are applied to the experiment data and contrasted, as shown in Fig. 2.

3.4.1. Bagging

An ensemble approach called bagging uses bootstrapping to build numerous training sets. Bagging combines multiple models trained on different subsets of the data, reducing overfitting and enhancing robustness. Using randomized and reproducible observations from the initial data set, the bootstrap approach creates several training sets. Moreover, numerous learning prototypes are produced by training every participant in the aggregation framework using these subsections. Once the training groups have been created, the ultimate choice is reached after combining the forecasts regarding every model.

3.4.2. Gradient boosting machine (GBM)

Gradient boosting machine (GBM) is a robust ensemble learning technique used in an optimized diabetes mellitus detection model for better prediction accuracy and clinical decision-making. By using GBM, the model is able to sequentially integrate the strengths of several weak learners, which are typically decision trees, to improve overall predictive performance. This methodology is especially useful for diabetes detection because it allows the model to focus on instances that are essential for precise predictions and correct errors iteratively. Hyperparameters are adjusted during the optimization process to avoid overfitting and provide the best possible balance between sensitivity and specificity. The GBM-based model enhances clinical decision-making by providing more accurate predictions by highlighting features that are relevant to diabetes diagnosis. This strategy works well to improve the effectiveness of diabetes detection systems, which helps medical professionals treat patients with diabetes mellitus or those who are at risk for the disease in a timely and accurate manner.

Gradient boosting sequentially trains weak learners, focusing on the misclassified instances to create a strong ensemble model. This method creates numerous learners by reweighting each sample in the instruction set throughout the learning stage. The augmenting gradient filter is a predictive model created by combining multiple learners who are weak, frequently in the type of decision trees [18]. Therefore, the amount of trees depends on the size of the dataset that was cast off. Furthermore, it is primarily utilized whenever the simulation's bias error has to be reduced. A gradient boosting machine (GBM) will be used to determine the coefficient values. It is necessary to compute the amount of loss function employed to derive the coefficient's value. Moreover, it is computed in Eq. (2),

$$GB_{m+1}(Y) = GB_m(Y) + \gamma_m K_1(y, e_m) \tag{2}$$

$$N_1 = (x_1 - x_1')^2 \tag{3}$$

$$N_1 = (X - GB_m(y))^2 \tag{4}$$

In Eq. (2), the formula represents an update in the boosted GBM, where $GB_{m+1}(Y)$ is the prediction of the model at iteration m+1, $GB_m(Y)$ is the prediction at iteration m, γ_m is the learning rate for iteration m, and $K_1(y, e_m)$ represents the correction factor to be added to the previous prediction to improve the model's accuracy.

In Eq. (3), N_1 calculates the squared difference between two values, x_1 and x_1', suggesting a measure of the squared error or distance between these values.

In Eq. (4), N_1 computes the squared difference between the input data X and the prediction $GB_m(y)$at iteration m, effectively quantifying the squared error between the model's prediction and the actual data, which is often used as a key component in optimizing machine learning algorithms like gradient boosting.

Algorithm for GBM-DRU

Input: *Patient data has been collected*
Output: *Prediction of Diabetes Mellitus*
Load data for provided features
$X = \{X_1, X_2, X_3, \ldots.\}$ *//Data Collection*
Pre-processing *//Min-Max Normalization*
y'' *is given in Eqn. (1)*
Feature Selection *//Data Reduction Unit*
Prediction of DM *// Gradient Boosting Machine*
(GBM)
$GB_{m+1}(Y)$ *is given in Eqn. (2)*
End

4. Results and discussions

The reviewed literature contributes various methodologies and insights to the proposed work on optimizing diabetes mellitus detection. These studies emphasize the significance of early diagnosis and its potential impact on healthcare in regions with limited resources. They introduce machine learning techniques, such as Light Gradient Boosting Machine (LightGBM), Gradient Boosting Machine (GBM), and ensemble methods like AdaboostM1, to improve diabetes prediction accuracy. Furthermore, the literature underscores the importance of accurate prediction and early intervention in preventing diabetes-related complications. Some studies explore the relevance of specific data features and depth measurements in diabetic nephropathy prediction, while others discuss the impact of persistent hyperglycemia on renal complications, providing valuable insights into diabetes management and risk assessment. Collectively, these literature sources offer a diverse array of approaches and methodologies that inform the development of an optimized diabetes detection model using the GBM-DRU Network, feature engineering, and ensemble learning techniques.

The proposed optimized DM detection model utilizing the GBM-DRU

Table 1
Observation of pre-processed scaled data.

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

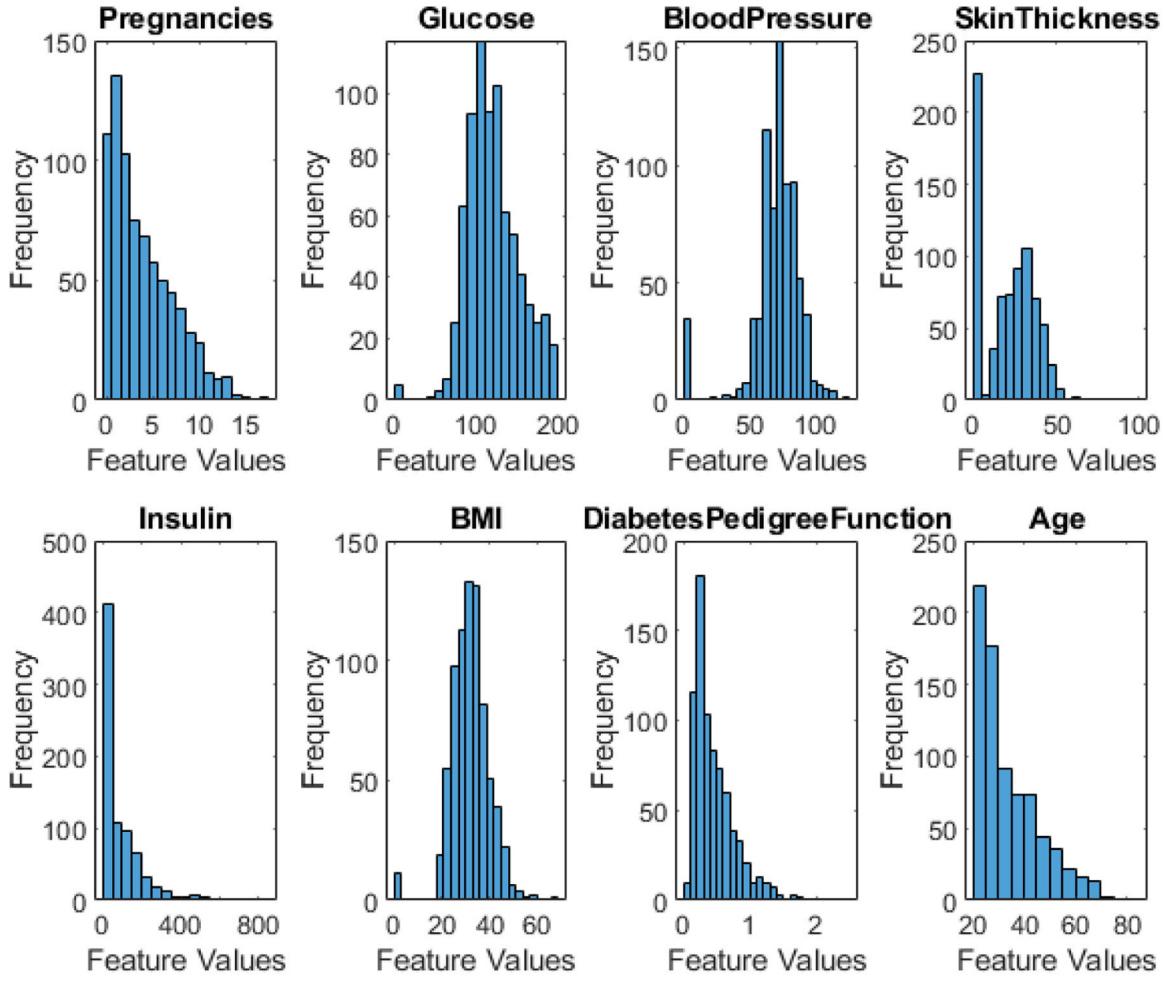

Fig. 3. Histogram features of DM-affected patients in the dataset.

network, feature engineering, and ensemble learning techniques offers an effective solution for accurate DM detection. The planned method can be compared using some features as follows.

4.1. Histogram features

The shape of this distribution illustrates the frequency of a value. Table 1 shows the data records of ten patients, and the resulting Fig. 3 shows the histograms for each characteristic in the data set.

Table 1 provides a dataset of pre-processed scaled data related to diabetes, where each row represents an individual, and columns capture various health attributes. These include the number of pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI, diabetes pedigree function, age, and the outcome (1 for diabetes, 0 for non-diabetes). The data is prepared for analysis, with numerical values scaled for consistency. For instance, the first row describes an individual with 6 pregnancies, a glucose level of 148, blood pressure of 72, skin thickness of 35, no insulin, a BMI of 33.6, a diabetes pedigree function value of 0.627, and an age of 50, diagnosed with diabetes (Outcome = 1). Such datasets are commonly used in machine learning to develop models for predicting diabetes based on these health indicators.

When examining the histograms, it can be noticed that some characteristics, such as blood sugar, blood pressure, skin thickness, and body mass index (BMI), appeared to be regularly handed out. In contrast, other characteristics, such as pregnancy, insulin, diabetes pedigree function, and age, are distributed exponentially. Furthermore, the distribution of age should probably follow a normal pattern. Nevertheless, the restrictions placed on the data collecting may have tipped the scales

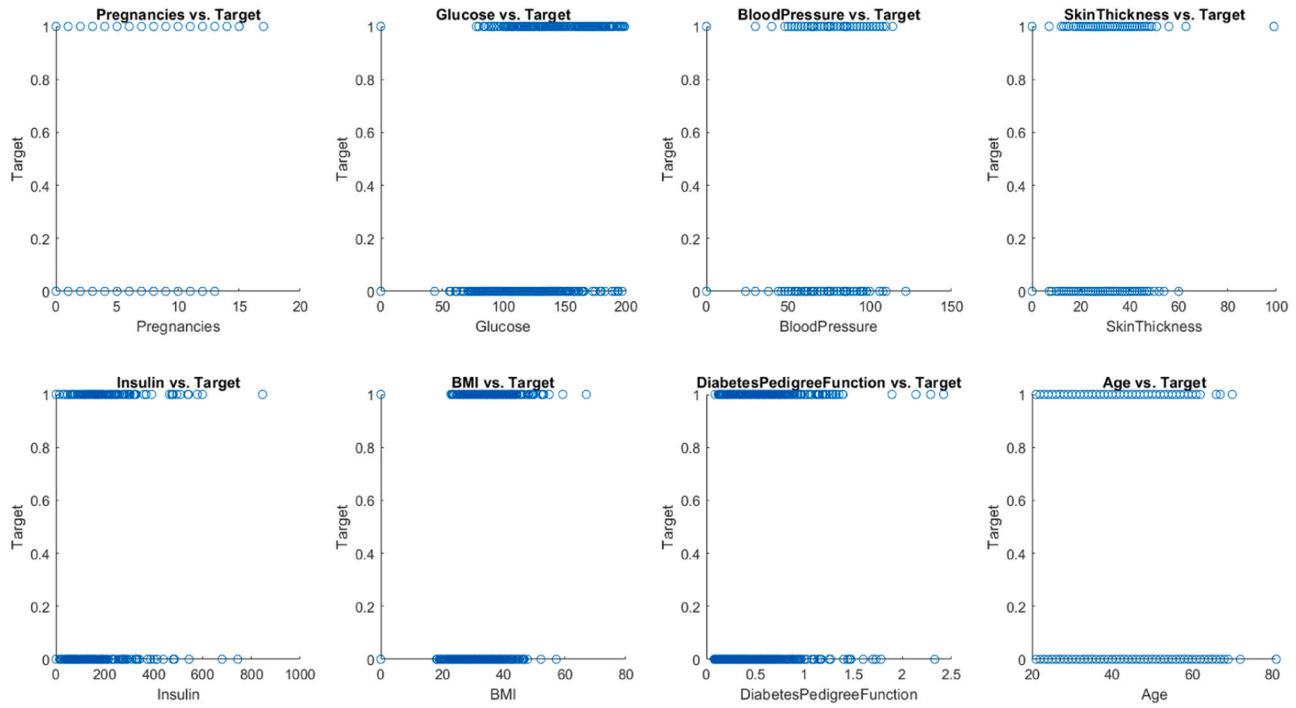

Fig. 4. Pair plot of the given features.

Fig. 5. Pictorial representation of selected features.

[27].

The performance of choosing a subdivision of relevant topographies, or variables, from an extensive data collection in demand to figure algorithms is recognized as feature selection. Supplementary acronyms for feature selection comprise the mutable collection, quality collection, and adaptable subdivision collection. By removing duplicate and unnecessary information, the primary goal of the feature choice is to identify features that accurately describe the data collection.

The Fig. 4 shows the relationship between health care measures and the target variable (0 for nondiabetic, 1 for diabetic) indicating diabetes status. Each scatter plot compares specific health metrics such as pregnancy, glucose levels, blood pressure, skin weight, insulin levels, BMI, diabetes tribe occupation, age, etc. to mean glucose levels and mean BMI that of diabetes shows a positive association with risk of diabetes, whereas age and moderately positive pregnancy rate The association shows the opposite, hypertension, skin obesity, insulin levels, and diabetes family anatomic work shows no significant association with risk of diabetes. This study contributes to our understanding of the impact of these health factors on diabetes risk and any associations between them.

A visual illustration of the relevance of particular factors in affecting the model's predictions is provided by the graphical portrayal of the selected features graph in Fig. 5. Clinicians and data scientists can determine which variables or factors have the most effects on the model's accuracy by looking at feature significance scores or contributions. In turn, this facilitates clinical decision-making by emphasizing the main markers that influence the likelihood of developing diabetes. Additionally, the graphical representation makes it possible to see any outliers or noise in the dataset, allowing for the improvement of feature engineering techniques and the general resilience of the model. Overall, this visual display of a few key aspects is an essential tool for enhancing prediction accuracy as well as assisting healthcare providers in making better clinical judgments on the treatment of diabetic patients.

4.2. *Feature importance ranks*

A distinct risk factor for intrusive disease in diabetic patients is feature importance rating [28]. The procedure involves employing a machine learning algorithm, such as Gradient Boosting Machine (GBM), with the DRU to train a predictive model on the pre-processed scaled data. During the training process, the algorithm assesses the contribution of each feature to the model's predictive performance. Feature importance is calculated based on metrics like information gain or Gini impurity reduction, depending on the specific algorithm used. Subsequently, these importance scores are normalized to generate the Feature Importance Ranks, providing insights into the relative significance of each variable in the diabetes detection model. The inclusion of this procedure in the manuscript enhances transparency and reproducibility, allowing readers to comprehend the factors driving the model's predictions and aiding in the identification of key features influencing diabetes outcomes. This is additionally by the findings about thrombocytopenia inferred from the DRU model in this study. Pro-inflammatory,

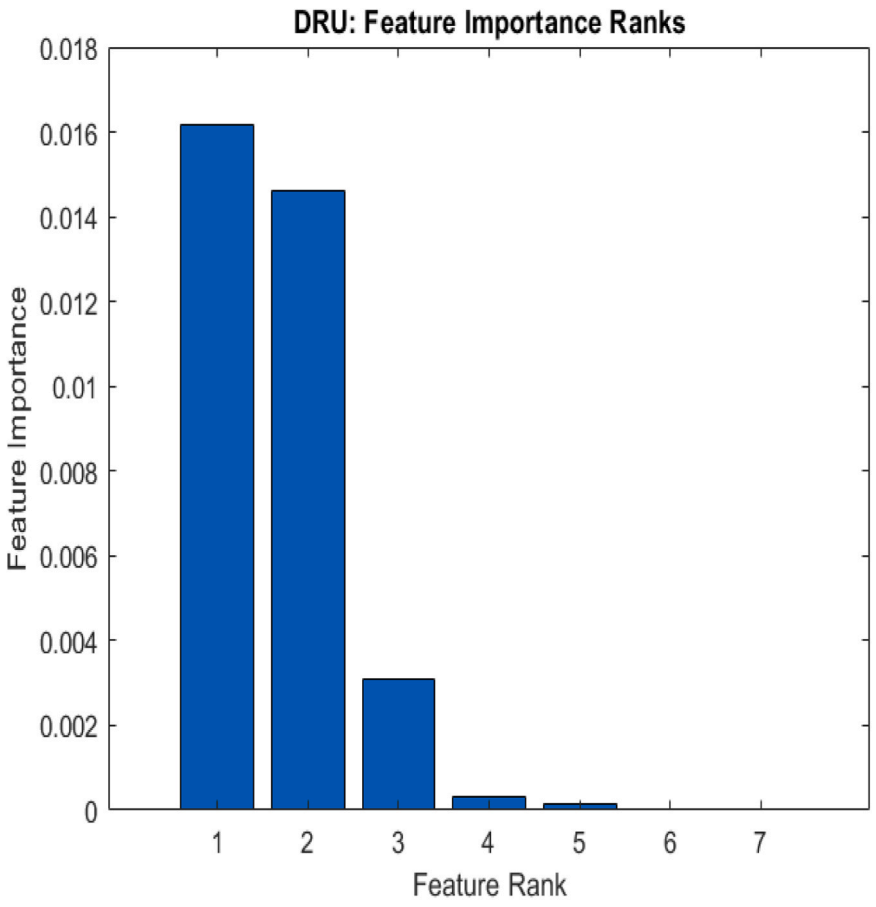

Fig. 6. Feature importance ranks of DRU.

anti-inflammatory, chemokine, antimicrobial, and various other mediators are produced and released by platelets that are triggered to control the body's innate immunological or adaptive immune system reactions. Leukocyte activation and endothelial cell damage are encouraged by the contact of platelets with infections or their consequences, endothelial chambers, and protected chambers.

With feature significance on the y-axis and feature rank on the x-axis, the Feature significance Ranks of the DRU graph in Fig. 6 offer a thorough picture of the relevance of individual features in the diabetes detection model. It can determine the most significant characteristics in

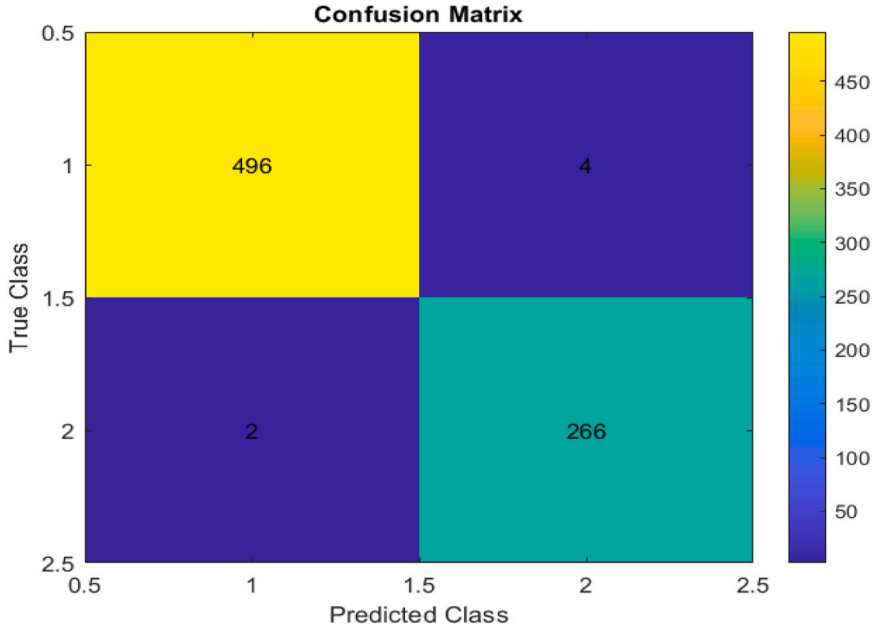

Fig. 7. Confusion matrix.

this graph based on their rankings and accompanying significance ratings. Features with higher rankings and significance values influence the model's decision-making more significantly. Data scientists may concentrate their efforts on perfecting and optimizing the most important variables thanks to this knowledge, which is crucial for feature selection and engineering. Additionally, it helps doctors by emphasizing the patient characteristics that are most important in predicting diabetes, enabling more focused therapeutic treatments and individualized patient care. The graphical representation facilitates more informed decision-making in both clinical practice and research by streamlining the comprehension of complicated model results.

4.2.1. Confusion matrix

The precision and efficacy of a classification model can be evaluated using a confusion matrix, an efficiency measuring tool in analytics and machine learning [29]. It thoroughly analyzes how the algorithm's projections stack against the real ground truth data. The aggregate quantity of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for every group or category in a classification issue is shown in a confusion matrix, which is commonly known as a square matrix.

4.2.2. True positives (TP)

The procedure successfully predicted a positive class.

4.2.3. True negatives (TN)

The procedure successfully predicted a negative class.

4.2.4. False positives (FP)

When an algorithm predicts a class that is positive while the actual class is negative and thus a false positive has occurred.

4.2.5. False negatives (FN)

A Type II error in which the prototype foresaw a corrupt lesson while the real class was positive.

A confusion matrix enables the assessment of a classification algorithm's precision, recall, and accuracy parameters that will give helpful information about how well it performs. These measures, including accuracy, precision, and recall, can be calculated from the numbers in the confusion matrix. The pictorial representation of the confusion matrix is given in Fig. 7.

4.3. Classification report of diabetes mellitus

A report on classification is a frequently used evaluation statistic when evaluating the effectiveness of a model for classification in machine learning. It offers thorough details on how the regression analysis was performed for every category in the dataset. Moreover, a classification report can be used in the context of patients with diabetes to assess how a classification model anticipates whether a patient has diabetes or not. Furthermore, the pictorial representation of the classification report is given in Fig. 8.

4.4. Performance metrics evaluation

The planned method GBM-DRU can be compared using parameters like Accuracy, Precision and Recall.

4.4.1. Accuracy

The accuracy may be defined as the proportion of properly classified illustrations. Accuracy is expressed in Eq. (6),

$$\text{Accuracy} = \frac{T_{\text{Positive}} + T_{\text{Negative}}}{T_{\text{Positive}} + T_{\text{Negative}} + F_{\text{Positive}} + F_{\text{Negative}}} \tag{6}$$

4.4.2. Precision

The ratio of suitable examples between the obtained incidences is precision or positive predictive value. Precision is computed from Eq. (7),

$$\text{Precision} = \frac{T_{\text{Positive}}}{T_{\text{Positive}} + F_{\text{Positive}}} \tag{7}$$

4.4.3. Recall

The proportion of the applicable occurrences that are returned is termed recall or sensitivity. The recall is uttered in Eq. (8),

$$\text{Recall} = \frac{T_{\text{Positive}}}{T_{\text{Positive}} + F_{\text{Negative}}} \tag{8}$$

Fig. 9 displays the performance metrics of Accuracy (99%), Precision (98.5%), and Recall (99%) of the planned method, and the parameters are either compared with other methodologies as follows. Table 2 provides a comparison of various methods based on accuracy, precision, and recall metrics. The Soft Voting Classifier exhibits an accuracy of 79%, precision of 73%, and recall of 70%. The J48 algorithm demonstrates superior performance with an accuracy, precision, and recall of 95%. The Random Forest algorithm also shows impressive results, achieving 97% accuracy, precision, and recall. However, the Naive Bayes classifier yields comparatively lower metrics with an accuracy of 70%, precision of 75%, and recall of 78%. Support Vector Machine (SVM) achieves higher scores with an accuracy of 80%, precision of 85%, and recall of 88%. XGBoost performs well in accuracy (85%) but slightly lower in precision (80%) and recall (80%). The GRM without

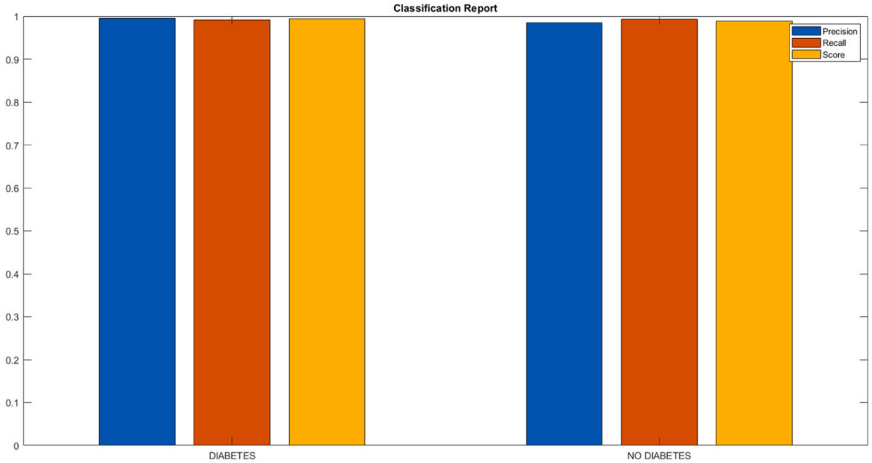

Fig. 8. Classification Report of the diabetes patients.

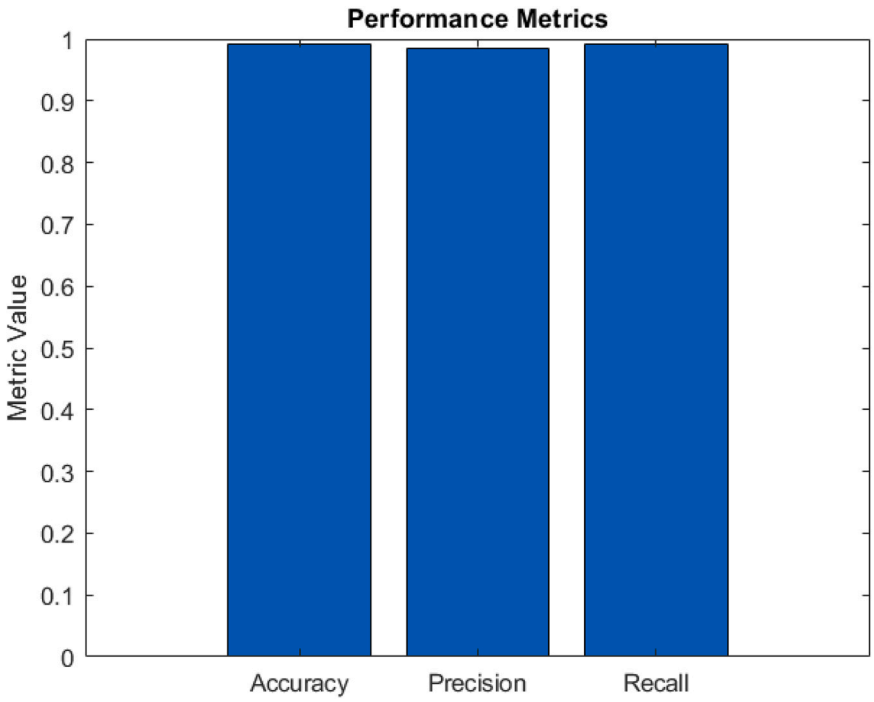

Fig. 9. Performance evaluation of accuracy, precision, and recall.

Table 2
Comparison table of accuracy, precision, recall.

Method	Accuracy (%)	Precision (%)	Recall (%)
Soft Voting Classifier [30]	79	73	70
J48 algorithm [31]	95	95	95
Random Forest algorithm [32]	97	97	97
Navie bayes[33]	70	75	78
SVM	80	85	88
XGBoost	85	80	80
GRM without DRU	97	96	96
Proposed GBM-DRU algorithm	99	98.5	99

DRU achieves high scores across the board with 97% accuracy, precision, and recall. Notably, the proposed GBM-DRU algorithm outperforms all others, achieving the highest accuracy of 99%, precision of 98.5%, and recall of 99%.

Table 2 and Fig. 10 show the comparison and performance evaluation of the highest Accuracy, Precision, and Recall. When comparing those parameters with the following three existing methods, i) Soft Voting Classifier algorithm [30], ii) J48 algorithm [31], iii) Random Forest algorithm [32], the proposed GBM-DRU algorithm produces greater accuracy (99%), greater precision (98.5%) and greater recall (99%).

Fig. 11 comparative analysis employing 10-fold cross-validation, four machine learning models, namely Support Vector Machine (SVM), Naïve Bayes, Random Forest, and a proposed method, were evaluated for their performance. The accuracy scores, expressed as percentages, highlight the effectiveness of each model in handling the given dataset. SVM exhibited consistent accuracy levels of 75%, showcasing its reliability across different folds. Naïve Bayes demonstrated a mean accuracy of 68%, while Random Forest performed slightly better with a mean accuracy of 73%. Notably, the proposed method outperformed the other models significantly, achieving an impressive accuracy of 95%, indicating its robustness in capturing patterns within the data. Furthermore, the proposed method demonstrated a remarkable 99% accuracy in the cross-validation process, showcasing its potential as a highly effective and promising approach for the given task. These findings underscore the superior performance of the proposed method in comparison to SVM, Naïve Bayes, and Random Forest in the specified cross-validation setting.

4.5. Discussions

The GBM-DRU (Data Reduction Unit) network with feature engineering and ensemble learning technique was suggested in this research as a method for developing an optimized diabetes mellitus diagnosis model. The goal was to increase the accuracy of the prediction and simplify clinical judgment while diagnosing diabetes. The gradient boosting machine (GBM) method and a data reduction unit (DRU) are combined in a revolutionary method called the GBM-DRU network. GBM is a potent machine learning method that excels at handling intricate, non-linear data interactions. The dimension of the input features is decreased by including DRU in the GBM structure; this helps to alleviate the problem of dimensionality and enhance model performance.

In the present investigation, feature engineering was quite important. Significant engineered traits related to diabetes mellitus are thoroughly chosen. Understanding of the subject matter and thorough dataset investigation were required for this technique. However, it is more important to advance the representation's capability to grasp the fundamental trends in the data and generate precise predictions by integrating informative and discriminative characteristics. Moreover, the effectiveness of the identification model was further improved using ensemble learning approaches. It has been used to lessen the danger of overfitting by merging various foundation models, such as GBM-DRU, and utilizing each of their distinct capabilities. Using methods like voting or averaging, the ensemble model's predictions were combined after being trained on various subsets of the dataset. This strategy can enhance the model's generalization capacity and boost total prediction accuracy.

In this paper, 768 persons' health records are taken, but in Table 1, the health records of 10 patients are produced. It explores pregnancy,

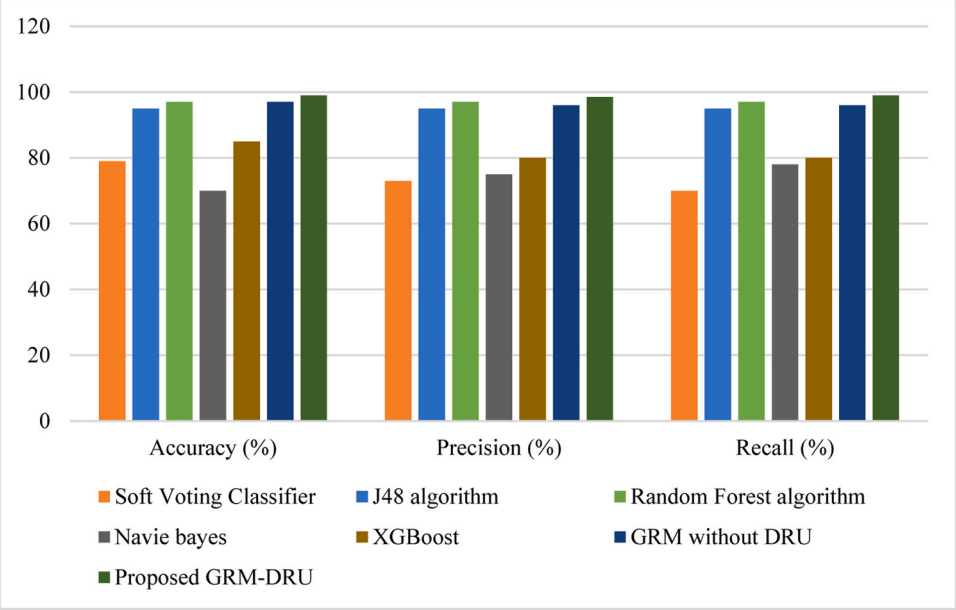

Fig. 10. Performance comparison chart with the highest Accuracy, Precision, and Recall.

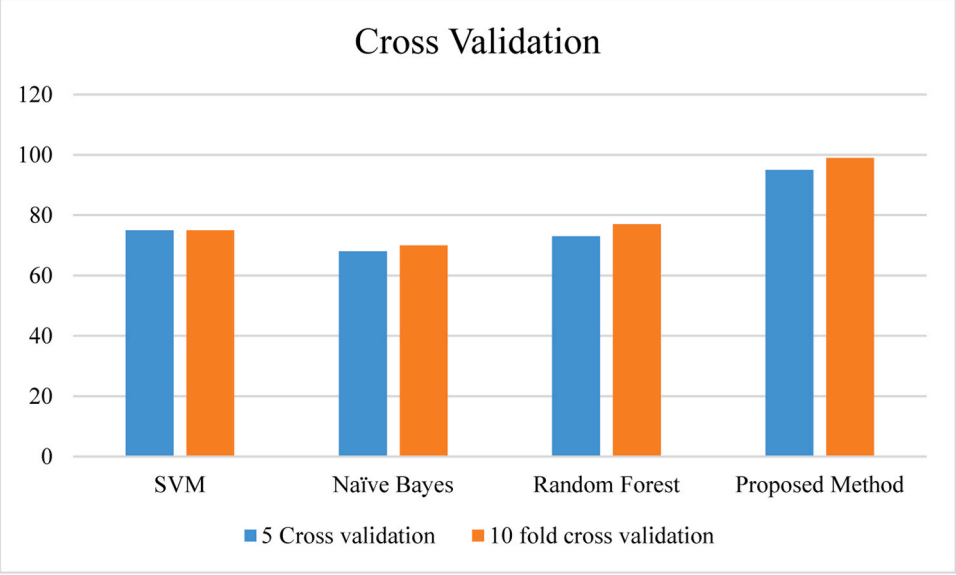

Fig. 11. Cross validation comparison.

glucose, blood pressure, skin thickness, insulin, BMI, diabetes, and age in patients. Then the histogram features are given in Fig. 3, and pair plots are second-hand to define the furthermost separate groups or the finest mixture of topographies to label a construction among the binary factors in Fig. 4. The features are selected using DRU, and their important feature ranks are shown in Figs. 5 and 6. The correlation matrix that explains the correlation coefficient between variables is shown in Fig. 7. The confusion matrix discusses the true class value in Fig. 8. Furthermore, the classification reports that consider the presence and absence of diabetes in patients using their health records with precision, recall, and fi-score are in Fig. 9. Moreover, the accuracy, precision, and recall standards of the proposed methods are shown in Fig. 10. Table 2 provides the comparison table with the higher accuracy, precision, and recall values of the suggested technique GBM-DRU compared with the state-of-the-art methods such as the Soft Voting Classifier algorithm, J48 algorithm, and Random Forest algorithm. The suggested method produces greater accuracy (99%), greater precision (98.5%), and greater recall (99%), and the comparison graph is shown in Fig. 11.

To elucidate the merits of the proposed method and eliminate potential contingencies, the study incorporates the k-fold cross-validation method. This technique involves partitioning the dataset into k subsets, training the model on k-1 folds, and validating it on the remaining fold. This process is repeated k times, ensuring that each subset serves as both training and validation data. By averaging the performance metrics across these iterations, the k-fold cross-validation method provides a robust assessment of the model's generalizability and helps mitigate issues related to overfitting or reliance on a specific data split. Its inclusion in the evaluation protocol contributes to the reliability of the findings, enhancing the study's credibility by providing a more comprehensive understanding of the proposed method's advantages and ensuring its efficacy across diverse data scenarios.

Thus, the consequences of the research showed that the future

approach worked well. In evaluating previous methods, the optimized diabetes mellitus detection algorithm had a much greater prediction accuracy. Therefore, researchers accurately forecast diabetes diagnoses using ensemble learning, feature engineering, and the power of GBM-DRU to extract significant information from the data.

5. Conclusion and future work

Throughout the present study, an improved diabetes mellitus detection model was created using the GBM-DRU network, feature engineering, and ensemble learning methods. Regarding better prediction accuracy and clinical decision-making in diabetes diagnosis, the suggested approach yielded encouraging results. Moreover, investigators can manage high-dimensional data efficiently while minimizing the curse of dimensionality because of the combination of GBM using the data reduction unit. The selection of instructive features and the improvement of the model's capability to grasp the fundamental trends in the data were both made possible by feature engineering. Furthermore, ensemble learning approaches enhanced the performance by mixing various base models even further. Henceforth, the results of this research have significant ramifications for the diabetes diagnosis industry. Furthermore, early treatment and improved outcomes for patients can result in the accurate and prompt identification of diabetes mellitus. The proposed approach may help healthcare providers make clinical decisions that are more well-informed, resulting in more individualized therapies and improved diabetes management. In the future, DM will be predicted using new methods.

Ethical approval

This article does not contain any studies with human participants performed by the author.

Declaration of Competing Interest

The authors declare that there is no conflict of interest.

Acknowledgment

The authors extend their appreciation to the deanship of scientific research at Northern Border University, Arar, KSA for funding this research work through the project number "NBU-FFR-2023–0151". This study is supported via funding from Prince Sattam bin Abdulaziz University project number (PSAU/2024/R/1445).

References

[1] K. Vidhya, R. Shanmugalakshmi, Deep learning based big medical data analytic model for diabetes complication prediction, J. Ambient Intell. Hum. Comput. vol. 11 (11) (2020) 5691–5702, https://doi.org/10.1007/s12652-020-01930-2.
[2] S.A. Alsuhibany, et al., Ensemble of deep learning based clinical decision support system for chronic kidney disease diagnosis in medical internet of things environment, Comput. Intell. Neurosci. vol. 2021 (2021) 1–13, https://doi.org/10.1155/2021/4931450.
[3] R. Liu, et al., Stacking ensemble method for gestational diabetes mellitus prediction in Chinese pregnant women: a prospective cohort study, J. Healthc. Eng. vol. 2022 (2022) 1–14, https://doi.org/10.1155/2022/8948082.
[4] A. Ampavathi, T.V. Saradhi, Multi disease-prediction framework using hybrid deep learning: an optimal prediction model, Comput. Methods Biomech. Biomed. Eng. vol. 24 (10) (2021) 1146–1168, https://doi.org/10.1080/10255842.2020.1869726.
[5] L.J. Marcos-Zambrano, et al., Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment, Front. Microbiol. vol. 12 (2021). ⟨https://www.frontiersin.org/articles/10.3389/fmicb.2021.634511⟩. Accessed: Oct. 07, 2023. [Online]. Available.
[6] Machine Learning Techniques for Screening and Diagnosis of Diabetes: a Survey 3; vol. 26, Teh. Vjesn. doi:10.17559/TV-20190421228262019.
[7] D.-Y. Kim, et al., Intelligent ensemble deep learning system for blood glucose prediction using genetic algorithms, Complexity vol. 2022 (2022) 1–10, https://doi.org/10.1155/2022/7902418.
[8] F.N. Iheagwam, O.T. Iheagwam, M.K. Onuoha, O.O. Ogunlana, and, S.N. Chinedu, "Terminalia catappa aqueous leaf extract reverses insulin resistance, improves glucose transport and activates PI3K/AKT signalling in high fat/streptozotocin-induced diabetic rats, Sci. Rep. vol. 12 (1) (2022) 10711.
[9] B. Tang, Y. Yuan, J. Yang, L. Qiu, S. Zhang, and, J. Shi, Predicting blood glucose concentration after short-acting insulin injection using discontinuous injection records, Sensors vol. 22 (21) (2022) 8454, https://doi.org/10.3390/s22218454.
[10] D. Tomic, J.E. Shaw, and, D.J. Magliano, The burden and risks of emerging complications of diabetes mellitus, Nat. Rev. Endocrinol. vol. 18 (9) (2022) 525–539, https://doi.org/10.1038/s41574-022-00690-7.
[11] S. G, S. Kp, V. R, Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals, Procedia Comput. Sci. vol. 132 (2018) 1253–1262, https://doi.org/10.1016/j.procs.2018.05.041.
[12] G. Geetha, K. Mohana Prasad, Stacking ensemble learning-based convolutional gated recurrent neural network for diabetes miletus, Intell. Autom. Soft Comput. vol. 36 (1) (2023) 703–718, https://doi.org/10.32604/iasc.2023.032530.
[13] D.D. Rufo, T.G. Debelee, A. Ibenthal, and, W.G. Negera, Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM), Diagnostics vol. 11 (9) (2021) 1714, https://doi.org/10.3390/diagnostics11091714.
[14] R. Ali, M.H. Siddiqi, M. Idris, B.H. Kang, and, S. Lee, Prediction of Diabetes Mellitus Based on Boosting Ensemble Modeling, vol. 8867, in: R. Hervás, S. Lee, C. Nugent, J. Bravo (Eds.), Lecture Notes in Computer Science, Ubiquitous Computing and Ambient Intelligence. Personalisation and User Adapted Services, vol. 8867, Springer International Publishing, Cham, 2014, pp. 25–28, https://doi.org/10.1007/978-3-319-13102-3_6. vol. 8867.
[15] R. Islam, S. Banik, K.N. Rahman, and M.M. Rahman, "A Comparative Approach To Alleviating The Prevalence Of Diabetes Mellitus Using Machine Learning," 2023.
[16] L. Zhang, Y. Wang, M. Niu, C. Wang, and, Z. Wang, Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: the Henan Rural Cohort Study, Sci. Rep. vol. 10 (1) (2020) 4406, https://doi.org/10.1038/s41598-020-61123-x.
[17] L. Zhao, et al., Combining glomerular basement membrane and tubular basement membrane assessment improves the prediction of diabetic end-stage renal disease, J. Diabetes vol. 13 (7) (2021) 572–584, https://doi.org/10.1111/1753-0407.13150.
[18] B.S. Ahamed, M.S. Arya, and, A.O.V. Nancy, Diabetes mellitus disease prediction using machine learning classifiers with oversampling and feature augmentation, Adv. Hum. Comput. Interact. vol. 2022 (2022) 1–14, https://doi.org/10.1155/2022/9220560.
[19] N. Shima, et al., Characteristic renal histology of a 81-year-old patient with a 30-year history of diabetes mellitus: a case report, CEN Case Rep. vol. 9 (4) (2020) 338–343, https://doi.org/10.1007/s13730-020-00483-9.
[20] S. Cui, Y. Wang, Y. Yin, T. Cheng, D. Wang, and, M. Zhai, A cluster-based intelligence ensemble learning method for classification problems, Inf. Sci. vol. 560 (2021) 386–409.
[21] Y. Wang, S. Wang, X. Sima, Y. Song, S. Cui, and, D. Wang, Expanded feature space-based gradient boosting ensemble learning for risk prediction of type 2 diabetes complications, Appl. Soft Comput. (2023) 110451.
[22] Y. Su, C. Huang, W. Yin, X. Lyu, L. Ma, and, Z. Tao, Diabetes mellitus risk prediction using age adaptation models, Biomed. Signal Process. Control vol. 80 (Feb. 2023) 104381, https://doi.org/10.1016/j.bspc.2022.104381.
[23] K.R. Tan, et al., Evaluation of machine learning methods developed for prediction of diabetes complications: a systematic review, J. Diabetes Sci. Technol. vol. 17 (2) (2023) 474–489.
[24] O. Odukoya, et al., Development and comparison of three data models for predicting diabetes mellitus using risk factors in a Nigerian population, Health Inf. Res vol. 28 (1) (Jan. 2022) 58–67, https://doi.org/10.4258/hir.2022.28.1.58.
[25] H. Kaur, V. Kumari, Predictive modelling and analytics for diabetes using a machine learning approach, ACI vol. 18 (1/2) (2022) 90–100, https://doi.org/10.1016/j.aci.2018.12.004.
[26] P.Y. Taser, "Application of Bagging and Boosting Approaches Using Decision Tree-Based Algorithms in Diabetes Risk Prediction," in The 7th International Management Information Systems Conference, MDPI, Mar. 2021, p. 6. doi: 10.3390/proceedings2021074006.
[27] "2018 Fourth International Conference on Computing Communication Control and Automation," in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India: IEEE, Aug. 2018, pp. 1–1. doi: 10.1109/ICCUBEA.2018.8697366.
[28] C. Feng, J. Di, S. Jiang, X. Li, and, F. Hua, Machine learning models for prediction of invasion Klebsiella pneumoniae liver abscess syndrome in diabetes mellitus: a singled centered retrospective study, BMC Infect. Dis. vol. 23 (1) (May 2023) 284, https://doi.org/10.1186/s12879-023-08235-7.
[29] A. Mir and S.N. Dhage, "Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare," in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India: IEEE, Aug. 2018, pp. 1–6. doi: 10.1109/ICCUBEA.2018.8697439.

[30] S. Kumari, D. Kumar, and, M. Mittal, An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier, Int. J. Cogn. Comput. Eng. vol. 2 (Jun. 2021) 40–46, https://doi.org/10.1016/j.ijcce.2021.01.001.
[31] D. Pei, Y. Gong, H. Kang, C. Zhang, and, Q. Guo, Accurate and rapid screening model for potential diabetes mellitus, BMC Med Inf. Decis. Mak. vol. 19 (1) (Dec. 2019) 41, https://doi.org/10.1186/s12911-019-0790-3.
[32] 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 (14) (Jul. 2022) 5247, https://doi.org/10.3390/s22145247.
[33] S.-R. Massan, A.I. Wagan, and, M.M. Shaikh, A new metaheuristic optimization algorithm inspired by human dynasties with an application to the wind turbine micrositing problem, Appl. Soft Comput. vol. 90 (May 2020) 106176, https://doi.org/10.1016/j.asoc.2020.106176.