IJWR

Elevating Accuracy: Enhanced Feature Selection Methods for Type 2 Diabetes Prediction

Ghazaleh Kakavand Teimoory[a], Mohammad Reza Keyvanpour[b]*

[a] *Data Mining Laboratory, Department of Computer Engineering Faculty of Engineering, Alzahra University Tehran, Iran; gh.kakavandteimoory@gmail.com*

[b] *Department of Computer Engineering, Faculty of Engineering, Alzahra University, Tehran, Iran; keyvanpour@alzahra.ac.ir*

ABSTRACT

Diabetes, a metabolic disorder, poses significant annual risks due to various factors, requiring effective management strategies to prevent life-threatening complications. Classified into Type 1, Type 2, and Gestational diabetes, its impact spans diverse demographics, with Type 2 diabetes being particularly concerning due to cellular insulin deficiencies. Early prediction is crucial for intervention and complication prevention. While machine learning and artificial intelligence show promise in predictive modeling for diabetes, challenges in interpreting models hinder widespread adoption among physicians and patients. The complexity of these models often raises doubts about their reliability and practical utility in clinical settings. Addressing interpretability challenges is crucial to fully harnessing predictive analytics in diabetes management, leading to improved patient outcomes and reduced healthcare burdens. Previous research has utilized various algorithms like Naïve Bayes, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and decision trees for patient classification. In this study using the Pima dataset, we applied a preprocessing technique that utilized the most important features identified by the Random Forest algorithm and we used an ensemble method combining the SVM algorithm and Naïve Bayes for the model. In the first section of the proposed method, we provided explanations regarding the dataset. In the second section, we elucidated all preprocessing steps applied to this dataset, and in the third section, we evaluated the model using the selected algorithm under investigation. The proposed model, after going through the various stages, was able to report an accuracy of 81.82%, a precision of 82.34%, an AUC of 88.19% and a Recall of 70.68%. Considering the review of similar studies, an improvement of 3.99% in accuracy demonstrates a significant advancement that highlights the benefits of traditional methods in disease prediction. These findings suggest the potential use of web-based applications to encourage both physicians and patients in diabetes prediction efforts.

Keywords— Diabetes Mellitus, Supervised Learning, Ensemble Method, Pima Dataset, E-Health.

1. Introduction

Diabetes stands as one of the foremost metabolic disorders globally. The human body requires a hormone called insulin to convert glucose into energy [1]. If there is any disruption in the energy production process, such as insufficient insulin production, lack of insulin production, or resistance to this hormone, it is recognized as diabetes, termed the "mother of diseases". If this problem arises, the blood sugar level increases significantly, which will lead to health issues for the individual. The disease manifests in three primary types [2]: Type one, triggered by the immune system's attack on insulin-producing cells; The second type, which is characterized by insufficient insulin secretion or, despite the presence of excess insulin in the blood due to insulin resistance, the body cannot utilize the available hormone and the third type, gestational diabetes, occurs when a pregnant mother develops diabetes during pregnancy. This condition usually resolves after pregnancy; however, both the mother and child have an increased risk of developing type 2 diabetes at older ages. Although some sources suggest that

doi http://dx.doi.org/10.22133/ijwr.2024.458872.1218

Citation. G. Kakavand Teimoory, M. Keyvanpour, "Elevating Accuracy: Enhanced Feature Selection Methods for Type 2 Diabetes Prediction ", *International Journal of Web Research*, vol.7, no.2,pp.37-48, 2024, doi: http://dx.doi.org /10.22133/ijwr.2024.458872.1218.

**Coressponding Author*
Article History: Received: 22 December 2023; Revised: 18 March 2024; Accepted: 22 March 2024.

diabetes can be further categorized into additional types based on various factors, such as monogenic diabetes syndromes, neonatal diabetes, diabetes onset during adolescence, exocrine pancreatic diseases like cystic fibrosis and pancreatitis, or as a result of medication and chemical exposure, including the use of certain drugs following organ transplants [3].

Its repercussions extend to a plethora of health complications, including strokes, blindness, kidney diseases [4], cardiovascular ailments, and immune system weakening. The imperative of addressing this ailment cannot be overstated, with projections indicating a staggering surge to nearly 630 million affected individuals by 2045 [5]. Notably, predicting Type 2 diabetes assumes paramount significance, as lifestyle modifications can potentially mitigate its complications to some extent. The significance of timely classification cannot be underscored enough, as many afflicted individuals remain unaware of their condition, jeopardizing their quality of life. Because in some cases, the symptoms of type 2 diabetes can remain hidden for up to 10 years, and the individual may be unaware of their condition. Delayed diagnosis impairs their ability to combat the disease and resume normalcy [6].

As the population grows, meeting individual healthcare needs remains challenging, further complicated by limited access to medical services [7]. Artificial intelligence, particularly machine learning, has emerged as a transformative force in medical diagnostics, continuously evolving[8]. Machine learning techniques show promise, offering encouraging results and fostering competition among researchers [9]. Despite the impressive accuracy of methods like neural networks and deep learning, the challenge of interpretability remains. These models are often seen as opaque "black boxes," which has heightened the need for explainable AI in disease prediction and diagnosis. The rapid advancement of the Internet and social networks has led to a significant increase in data volume and variety [10]. This development allows us to leverage this data for predicting diseases and enhances the accessibility of diagnostic models. This paper introduces a new preprocessing method for the PIMA dataset [11] aimed at improving outcomes. Our goal is to enhance model interpretability by exclusively using conventional machine learning algorithms, which offer reduced complexity. Table 1 summarizes the key advantages and disadvantages of this approach based on a review of previous research [12][13] [14].

Table 1. Advantages and disadvantages of using traditional machine learning approach

Advantages	***Disadvantages***
Simple,understandable and easy to implement	Risk of overfitting to the training data
Requires less data and enables fast and cost-effective training	Limited scalability and higher resource requirements for large datasets
Greater interpretability compared to neural network approaches	Limited predictive accuracy compared to neural networks

One of the key challenges of these methods is the high risk of overfitting to the training data and limited predictive accuracy compared to neural networks. By combining SVM and Naive Bayes, we aimed to mitigate these issues. SVM, with its ability to create robust decision boundaries, helps to reduce the risk of overfitting, while Naive Bayes, known for its simplicity and efficiency, allows us to maintain interpretability and handle smaller datasets effectively. This hybrid approach leverages the strengths of both algorithms, enabling us to overcome the limitations of traditional methods, particularly in scalability and accuracy.

In this section, we provide a clear and technical summary of the key contributions of our paper, outlined in bullet points for clarity:

- **Preprocessing and Feature Selection:** Feature importance was assessed using the Random Forest algorithm, which evaluates each feature based on decision tree performance. Features with the highest importance scores were selected to reduce dimensionality and focus on key variables. Missing values were imputed with the mean of existing values, and feature values were normalized to the range [0, 1] using Min-Max Scaling. These steps enhanced model accuracy by aligning selected features with relevant medical data.

- **Ensemble Methodology:** We employed an ensemble approach combining Support Vector Machines (SVM) and Naïve Bayes. This ensemble method leverages the complementary strengths of both classifiers for its capability with high-dimensional data and Naïve Bayes for its efficiency in probabilistic classification, to improve overall model performance and robustness.

- **Evaluation Metrics:** The ensemble model achieved an accuracy of 81.82%, a precision of 82.34%, an AUC of 88.19%, and a recall of 70.68%. Notably, the accuracy of the model represents a significant improvement over previous methods, underscoring its enhanced effectiveness in distinguishing between diabetic and non-diabetic cases.

Disease detection commonly involves binary classification testing. In this dataset, labels are designated as 1 for diabetic and 0 for nondiabetic (Figure 1). At the outset, an attempt has been undertaken to extract the most essential features from the dataset. One of the key pillars before performing any preprocessing is to conduct a thorough analysis of the dataset and its features [15]. The Random Forest algorithm employs probabilities to discern the most significant features for input [16]. To address the challenge of interpretability, we utilized an ensemble method combining two traditional algorithms. This approach significantly contributed to resolving the issue by aligning the obtained features with medical findings and evaluating their importance. The strength of RF in making predictions lies in combining the outputs of numerous individual decision trees, thereby harnessing the collective power of these weaker learners [17]. Decision tree algorithms are computational frameworks utilized to ascertain outcomes by progressively testing inputs until certainty is attained. Decision tree learning entails constructing a tree from pairs of inputs and corresponding outcomes to approximate the function conveyed by the data. Despite the computational complexity associated with identifying the smallest optimal decision tree for a dataset, Decision tree algorithms are essential in machine learning because of their straightforwardness and ease of understanding. They are classified into classification trees and regression trees based on their predictive objectives, rendering them invaluable tools in predictive modeling [18]. Additionally, the Random Forest algorithm is utilized to assess feature importance, and noteworthy features are selected to form a new dataset [19]. Subsequently, we employed an ensemble approach combining the SVM and Naïve Bayes algorithms for model testing. This combination leverages the strengths of both methods: SVM's effectiveness in handling classification problems and Naïve Bayes' simplicity and efficiency. The use of a polynomial kernel in SVM for more precise separation demonstrated that class label separation in a higher-dimensional space improves the results, further enhancing the model's predictive accuracy.

The paper is organized as follows: Section 2 defines the problem and specifies the tasks. Section 3 reviews the research background and previous studies. Section 4 presents the proposed method, with subsections on the dataset (4.1), preprocessing operations (4.2), and the selected algorithm with its justifications (4.3). Section 5 details the results of the proposed model, Section 6 evaluates the model, and Section 7 provides the conclusion. Figure 2 illustrates the overall process.

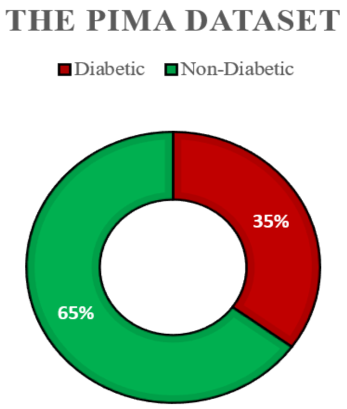


Figure. 1. The diabetic-to-nondiabetic ratio in the PIMA dataset

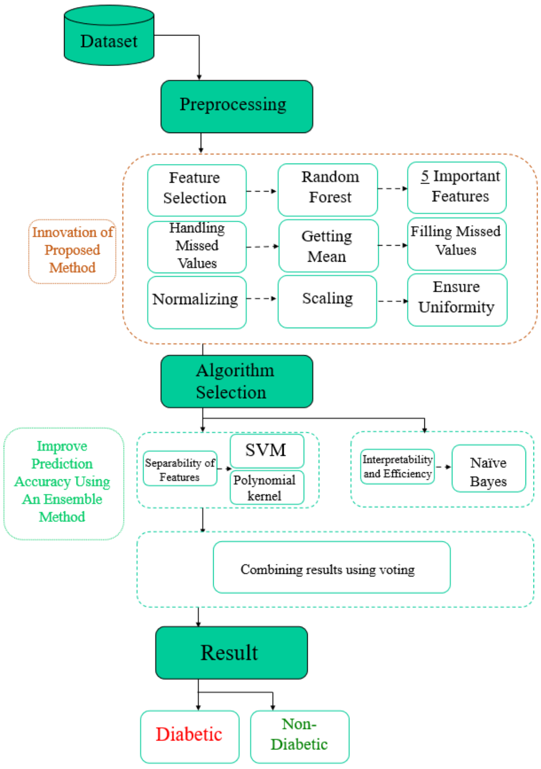


Figure. 2. The diagram of the suggested model

2. Problem Definition

In recent years, the prevalence of Type 2 diabetes has been on the rise, posing significant challenges to public health systems worldwide. With the advent of advanced computational techniques and the availability of datasets, there has been growing interest in leveraging machine learning algorithms to aid in the early detection and management of diabetes. This paper introduces an innovative approach designed to enhance the prediction accuracy of Type 2 diabetes through machine learning techniques. Before delving into the details of our proposed methodology, it is essential to precisely define the problem at hand. The objective is to create a machine learning model that accurately predicts an individual's risk of developing Type 2 diabetes using a dataset comprising demographic, clinical, and diagnostic information. Formally, let Eq (1) represent the dataset, where X_i denotes the feature vector of the ith individual and Y_i represents the binary label indicating whether the individual has Type 2 diabetes (1) or not (0). The objective is to train a predictive model M that can map the input feature vector X to the corresponding label Y as Eq (2), where X denotes the feature space and $Y = \{0,1\}$ denotes the set of possible class labels.

$$D = \{(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)\} \qquad (1)$$

$$M: X \rightarrow Y \qquad (2)$$

3. Related Works

Numerous scholars have contributed to this domain. Chang et al. [11] explored supervised learning techniques, including Random Forest, Naïve Bayes, and J48. Among these, Naïve Bayes notably excelled, particularly in identifying the five most critical features. Oliullah et al. [20] focused on developing accurate diabetes prediction models for early detection in females using various machine learning algorithms. Researchers employed techniques like random forest, XGBoost, and others. The use of Shapley additive explanation (SHAP) also aided model interpretation. Zhang et al.[21] introduced AHDHS Stacking, an ensemble learning framework for diabetes classification using harmony search algorithms. It optimized model performance through feature selection and base learner combinations. Key features identified included age, gender, and glucose levels, making it effective for early diabetes prediction. Kibria et al. [22] employed six algorithms, incorporating a group classifier, to create a transparent diabetes detection model. Their innovative approach, incorporating Shapley explanations, led to a model understandable by physicians. Edeh et al. [23] aimed to outperform others by employing four supervised learning algorithms on the PIMA dataset, with SVM yielding the best model performance. Ahmed Jasim et al. [24] developed optimized machine learning models, including AdaBoost-ET, to predict diabetes using the PIMA Indian Diabetes dataset, attaining high accuracy and surpassing the performance of previous models. Shamim Reza et al. [25] introduces a sophisticated non-linear kernel for SVM models, employing radial basis function (RBF) and RBF city block kernels, to improve Type 2 diabetes classification using the PIMA dataset. This approach outperforms existing kernel functions, offering improved accuracy and robustness for early diabetes prediction in clinical settings. Chatrati et al. [26] proposed a smart home health monitoring system utilizing the SVM algorithm, achieving 75.0% accuracy. Kawarkhe et al. [27] proposed a diabetes prediction model using data preprocessing and Ensemble Classifiers (CatBoost, LDA, LR, Random Forest, GBC), achieving a good accuracy. Performance was evaluated with AUC and ROC metrics, aiming to enhance early-stage diabetes detection and clinical trials. Abdolahi et al. [28] used particle swarm optimization (PSO) for feature selection to improve diabetes prediction accuracy with machine learning models. They evaluated the performance of various algorithms, including Decision Tree, Random Forest, and Naïve Bayes, across three medical datasets, aiming to enhance classification efficiency and effectiveness. Tasin et al. [29] utilized semi-supervised and group learning, along with feature selection, achieving 78% accuracy using the SVM algorithm. Their work aimed to provide explainable artificial intelligence using the LIME and SHAP frameworks. Ahmed Hashim et al. highlight the broad-reaching implications of diabetes across various age demographics and underscore the critical role of early detection in its management. Their study introduces a structured framework for evaluating research on diabetes detection and identification, providing valuable insights derived from analyzing 54 relevant studies to inform future research efforts in this field. Perdana et al. [30] employed the KNN algorithm to categorize patients into diabetic and non-diabetic categories, finding the family history feature insignificant, suggesting its removal without significant impact on outcomes. They reported optimal results with k set to 22. Alnowaiser et al. [31] addresses missing data in diabetes detection datasets by proposing an automated method using a (KNN) imputer and a Tri-ensemble voting classifier model. This approach improves accuracy significantly and outperforms seven alternative machine learning algorithms, highlighting its potential for early diabetes detection and enhancing patient care quality. Febrian et al. [32] focused on patient classification using supervised learning algorithms like KNN and Naïve Bayes, with Naïve Bayes showing superior accuracy at 76.07%.

In the reviewed studies, various techniques were employed to extract important features from the PIMA dataset. However, these features often did not align with medical findings and were used in a purely mechanical and systematic manner. In our paper, we addressed this issue by ensuring that the extracted features are clinically relevant, which led to improved model performance. Furthermore, while ensemble learning can sometimes increase training complexity, our approach was designed to optimize performance while minimizing computational costs.

4. Proposed Method :NBS

In this section, the proposed method of our paper is examined according to Figure 2. In our proposed method (NBS), which is a combination of two algorithms, Naïve bayes and SVM, preprocessing is initially performed on the dataset to prevent harm to the model. Then, based on graphical analysis, model selection is conducted, and the data is trained. In this training, we used a combination of two methods:

SVM and Naïve Bayes. After training each method, predictions are made by combining the results and voting. Finally, the results are evaluated with the test dataset, and the effectiveness of the proposed model is assessed using predefined metrics. We employed the hold-out method to partition the datasets into training and testing sets, with a 70-30 split, where 70% of the dataset is allocated for model training, while the remaining 30% is designated for testing purposes. Our method maintains the simplicity and ease of implementation inherent in traditional approaches. This ensures that our model is not only easy to understand but also straightforward to deploy, making it accessible to a broader range of users without requiring deep technical expertise. The greater interpretability offered by traditional methods is a significant strength of our approach. It allows for better transparency and trust in the model's predictions, which is crucial in many applications, especially those requiring accountability and decision-making justification. By requiring less data, our method enables fast and cost-effective training, making it highly efficient in scenarios where computational resources and time are limited. This efficiency is particularly beneficial for rapid prototyping and iterative development processes. While traditional methods often face challenges with scalability and resource demands for large datasets, our approach incorporates optimizations that mitigate these issues, ensuring better performance and resource management even with increasing data sizes. Though traditional approaches typically have limited predictive accuracy compared to neural networks, our method integrates advanced techniques to enhance accuracy without compromising the interpretability and simplicity advantages. This balanced approach offers a practical solution for achieving reliable performance in real-world applications.

4.1. Dataset

Despite the availability of larger and more complex diabetes datasets, the PIMA Indian Diabetes dataset remains a key benchmark for diabetes classification research [11]. The dataset was selected due to its proven history of yielding strong results in modeling [33]. The dataset was supplied and verified by the National Institute of Diabetes and Digestive and Kidney Diseases [34] and available on (www.kaggle.com). It comprises data from 768 female individuals, each characterized by 8 features. These individuals are classified as either diabetic or nondiabetic. Among them, 500 are nondiabetic, while 268 are diabetic. The dataset includes missing values, which could affect the model's performance. Statistical details [11] regarding this dataset are provided in Table 2. It's important to mention that all values are represented in numerical format and it should be noted that this statistical information is before any preprocessing is conducted. Regarding the features introduced in the table, it is worth mentioning the following explanations. The "Pregnancy" feature indicates the number of pregnancies a person has had. The "Glucose" feature indicates the plasma glucose concentration measured two hours after taking an oral glucose tolerance test. The "Blood Pressure" feature indicates the diastolic blood pressure level. The "Skin Thickness" feature denotes the thickness of the skinfold measurements on the triceps area, expressed in millimeters. The "Insulin" feature displays the level of insulin present in the blood, obtained two hours after ingestion. The "BMI" feature represents the body mass index (BMI) is computed by dividing an individual's weight by the square of their height. Following that, the "Diabetes Pedigree Function" feature reflects the diabetes family history of the individual, and the last feature is the "Age" of the women.

Table 2. Information Of Dataset

Features	***Minimum***	***Maximum***	***Range***	***Missed Values***
Pregnancies	0.00	17.00	[0,17]	0
Glucose	0.00	199.00	[0,199]	5
Blood Pressure	0.00	122.00	[0,122]	35
Insulin	0.00	846.00	[0,99]	374
Skin Thickness	0.00	99.00	[0,846]	227
BMI	0.00	67.10	[0,67.1]	11
Diabetes Pedigree Function	0.078	2.42	[0.078,2.42]	0
Age	21.00	81.00	[21,81]	0

4.2. Preprocessing

To enhance the model and considering the presence of missing values in the clinical data of individuals for disease prediction or diagnosis, as well as the importance of obtaining key features and reducing dimensions, we decided to perform effective preprocessing on the dataset. For this purpose, we first identified and selected the most important features, then filled in the missing values, and subsequently normalized the range of all numbers.

We prioritized feature selection over imputing missing values to ensure its potential impact on the final outcome. Since disease diagnosis and prediction often rely on clinical features gathered from individuals, this study emphasizes the selection of the most crucial features. In previous methods, various techniques were used to extract important features, but discrepancies between these features and medical knowledge have impacted the final model results. This is because certain features may exert a more significant influence on determining an individual's diabetic status. For instance, the impact of blood pressure levels may differ from that of skin thickness. Therefore, we opted to utilize the Random Forest algorithm. The use of an ensemble approach for algorithm selection and the identification of the top five highest-scoring features are key innovations of this paper. These features were rigorously validated against medical findings, demonstrating that integrating domain knowledge through a combined methodology significantly enhances the model's evaluation metrics. By employing this algorithm, we narrowed down the features to five, identifying and selecting the top five most important ones. The significance of these features and the top five most crucial ones are depicted in Figure 3. Additionally, the top 5 features are highlighted in a bolder color.

Random Forest (RF) consists of multiple decision trees, each built from a bootstrapped sample of the training data. Each tree follows a process of iterative partitioning, beginning from the root node and applying the same splitting procedure repeatedly until specific stopping criteria are fulfilled. Figure 4. depicts the procedure of building random forests and identifying the most significant features. After constructing each tree, a score is assigned to each utilized feature to ascertain their importance in the resultant classifications. Based on the ratings generated by this algorithm, five crucial features are singled out. Additionally, the pseudocode utilized in Figure 5. is presented.

Algorithm 1. (Figure 5.) outlines the process employed by the Random Forest algorithm to identify and select important features. The algorithm begins by initializing an empty list to store these significant features (line 1). Next, for each tree within the Random Forest ensemble, the algorithm goes through a training procedure utilizing a bootstrapped dataset, a method in which subsets of the initial data are sampled with replacement (line 3). Following the training of each decision tree, the algorithm proceeds to extract the features deemed important within that particular tree (line 4). These identified features are then appended to the list of important features (line 5). To discern the relative significance of each feature, the algorithm tallies the occurrence count of each feature within the list (line 6). This count provides insights into the frequency with which each feature is identified across the ensemble of decision trees. Subsequently, the algorithm sorts the features based on their occurrence counts in descending order (line 7). This sorting operation arranges the features from most to least frequently identified, allowing for the prioritization of features based on their perceived importance. In the final step of the algorithm (line 8), a predetermined number of top features is selected for further analysis and utilization in subsequent tasks. These selected features represent the subset deemed most critical for the modeling process, based on their prevalence across the ensemble of decision trees. Ultimately, the algorithm concludes by returning the chosen set of selected features (line 9), providing a clear pathway for subsequent analyses or model refinement processes.

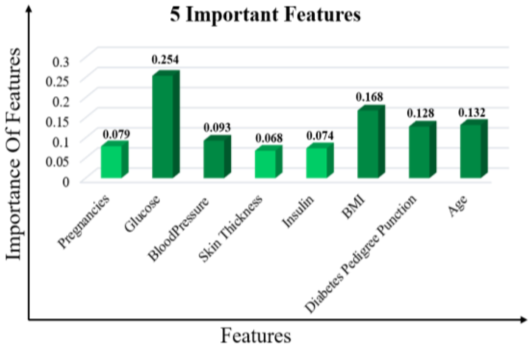


Figure. 3. The importance of each feature is based on Random Forest

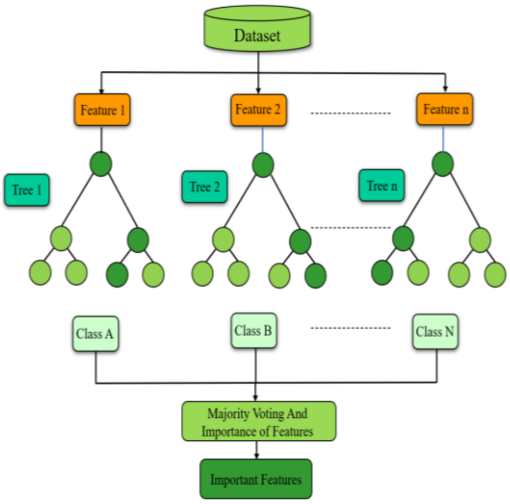
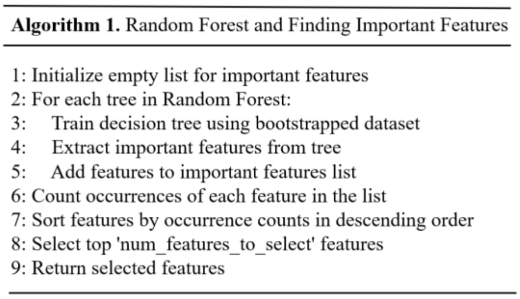


Figure. 4. Constructing Random Trees and Choosing Significant Features


Algorithm 1. Random Forest and Finding Important Features

```
1: Initialize empty list for important features
2: For each tree in Random Forest:
3:     Train decision tree using bootstrapped dataset
4:     Extract important features from tree
5:     Add features to important features list
6: Count occurrences of each feature in the list
7: Sort features by occurrence counts in descending order
8: Select top 'num_features_to_select' features
9: Return selected features
```

Figure. 5. pseudocode of Random Forest

Given the clinical nature of the data collection, encountering missing values is anticipated, a common occurrence in medical settings. In our dataset, missing values were noted (Table 1.). Thus, for enhanced preprocessing and improved model outcomes, we substituted these missing values with the mean and median of other values, stratified by their labels. Our findings suggested that using the mean of other values could yield a more favorable model impact, whereas utilizing the median appeared less suitable in this context.

After completing the aforementioned task, due to the considerable diversity among the values of each feature, we opted to standardize them within a range from zero to one. This standardization procedure aims to minimize the variance between the values, thereby mitigating any adverse effects on the model's performance. Standardization ensures that each feature contributes uniformly to the model, facilitating more reliable and robust predictions. By scaling the features to a consistent range, the model becomes better equipped to discern patterns and relationships within the data, ultimately enhancing its predictive accuracy.

4.3. Algorithm Selection

In our pursuit of a robust and transparent model, we initially focused on conventional supervised learning algorithms, including the Support Vector Machine (SVM), due to their well-established characteristics. However, we also integrated the Naïve Bayes algorithm into our approach, driven by its distinct advantages. One compelling reason behind our selection of SVM lies in its capability to address scenarios where linear separability of features proves elusive, a prevalent difficulty encountered in real-world datasets (Figure 6). This is where the kernel trick of SVM comes into play, offering a versatile solution by allowing for the transformation of feature space into higher dimensions. By employing this technique, we effectively expanded the scope of feature separability, thereby enhancing the discriminative power of our model. Specifically, we opted for a polynomial kernel trick to leverage its ability to capture complex relationships within the data. In this configuration, the polynomial kernel operates with a second-degree polynomial, denoted by Eq (3), where 'x' and 'y' represent input variables, 'c' signifies a constant parameter, 'd' denotes the polynomial degree, and 'alpha' serves as a scaling factor. This choice was strategic, as it enabled us to capture nonlinear relationships among features, contributing to a more nuanced understanding of the underlying data dynamics. In addition, SVM's versatility extends beyond its capacity to handle nonlinear separability; it also offers robustness against overfitting, a common pitfall in machine learning models. This robustness is achieved through the maximization of the margin between the support vectors, which can be mathematically formulated as an optimization Eq (4), where w is the weight vector, b is the bias term, ξi are slack variables that permit some misclassifications, while C is a regularization parameter that balances the trade-off between maximizing the margin and minimizing classification errors. In essence, the decision to leverage SVM was underpinned by its ability to tackle the complexities inherent in our dataset while maintaining transparency and interpretability. Naïve Bayes was selected for its simplicity and efficiency, which proved particularly valuable when working with smaller datasets or when rapid model training and prediction were crucial. The algorithm’s straightforward implementation allowed us to quickly train and evaluate our models.

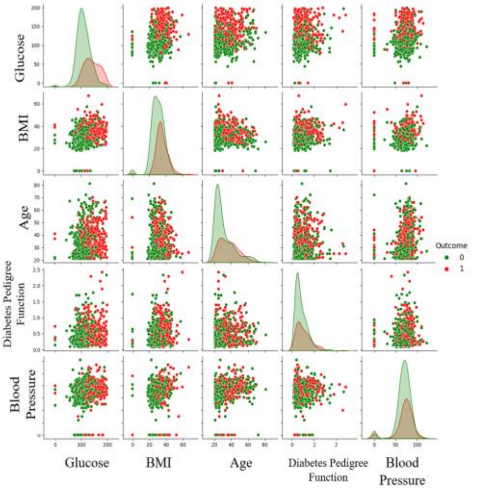


Figure. 6. Evaluating the linear separability of the features

Moreover, Naïve Bayes performs well even with limited data, providing reliable results despite smaller sample sizes. Its transparency makes it highly interpretable, enabling us to easily understand how individual features contribute to predictions. To further enhance our model's performance, we employed an ensemble approach that combines SVM with Naïve Bayes. This ensemble method capitalizes on the strengths of both algorithms: SVM’s capability to manage complex, non-linear relationships through kernel transformations, and Naïve Bayes’ efficiency and simplicity in probabilistic classification. By integrating these techniques, our ensemble model not only improved predictive accuracy but also provided a balanced approach that leveraged the complementary strengths of both algorithms. This ensemble strategy addresses the limitations of each individual method while enhancing overall performance. It allowed us to achieve superior results and gain deeper insights into the data, all while maintaining the desired characteristics of interpretability and efficiency. In essence, the decision to incorporate Naïve Bayes and use it in combination with SVM was driven by a need for simplicity, effectiveness, and robustness, culminating in a model that effectively meets our research objectives.

$$k(x, y) = (\alpha < x, y > +c)^d \tag{3}$$

$$min\omega, b, \varepsilon \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \varepsilon i \tag{4}$$
$$Subject\ to\ Y_i(\langle \omega, x_i \rangle + b) \geq ,1 - \varepsilon_i, \varepsilon_i \geq 0$$

The proposed method was analyzed for its time complexity at each step. Loading the dataset from the CSV file has a complexity of O(n), where n is the number of records. Data normalization using MinMaxScaler has a complexity of O (n.m), where m is the number of features. For the SVM model, training with a polynomial kernel has a complexity of O(n3) in the worst case due to the matrix operations involved. This complexity arises because polynomial kernels require computing and manipulating a higher-dimensional feature space, which can be computationally intensive. Prediction with the trained SVM model has a complexity of O (n.m). In the case of the Naïve Bayes model, training has a complexity of O (n.m), because it involves estimating the probability distributions of features for each class. Prediction with Naïve Bayes also has a complexity of O (n.m), as it involves computing probabilities and making classification decisions based on these probabilities. Combining the SVM and Naïve Bayes models using a Voting Classifier introduces additional complexity. The ensemble method requires performing predictions with each individual model and then combining these results. For the ensemble model, prediction complexity is O(n.(m+k), where k is the number of models in the ensemble (in this case, k=2).

5. Result

As previously outlined, this section delves into the outcomes of the preprocessing techniques applied to the designated dataset, along with the evaluation of the chosen algorithm. These findings have been meticulously examined and subjected to rigorous analysis to ascertain their validity and efficacy in addressing the research objectives. The evaluation metrics employed comprise accuracy, precision, specificity, and AUC. Accuracy (Eq (5)) denotes the ratio of correctly predicted samples to the total sample size, while precision (Eq (6)) signifies the correct identification of diabetic cases, a crucial measure for our study. Recall (Eq (7)) indicates what percentage of actual positive samples (diabetic) were correctly identified by the model, and AUC (Eq (8)) provides insights into the model's efficacy in class discrimination. Additionally, details regarding the parameters of the evaluation metrics are presented in Table 3. Further elucidation of the evaluation metric values can be found in Table 4.

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$AUC = \frac{1 + TPR - FPR}{2} \tag{8}$$

Table 3. serves as a testament to the effectiveness of our proposed model, revealing compelling outcomes that surpass existing benchmarks in the field of disease prediction. With an accuracy rate of 80.09% on the benchmark dataset, our model demonstrates a remarkable capacity to discern patterns and make accurate predictions. This achievement not only solidifies the credibility of our methodology but also positions our approach as a promising contender in the realm of predictive modeling. Moreover, a closer examination of Table 3 unveils notable enhancements in several key evaluation metrics, further bolstering the case for the

Table 3. Definitions of Evaluation Metric Parameters

Parameter	*Definition*
TP	Represents correctly identified positive or diabetic cases.
TN	Denotes correctly identified negative or nondiabetic cases.
FP	Signifies incorrectly predicted diabetic cases.
FN	Indicates incorrectly predicted nondiabetic cases.
TPR	The True Positive Rate, alternatively referred to as Sensitivity or Recall, indicates the ratio of correctly predicted positive cases to the total number of actual positive instances.
FPR	False Positive Rate represents the ratio of inaccurate positive predictions to the total true negative instances.

Table 4. Analysis of Evaluation Metrics Comparisons

Models On PIMA Dataset	*Accuracy*	*Precision*	*Recall*
The proposed model(NBS)	**81.82**	**82.34**	**70.68**
Naïve Bayes [11]	77.83	81.25	86.09
SVM [25]	72.60	74.90	68.00
XGB [35]	76.80	60.78	50.76
AdvanSVM [36]	76.00	74.00	73.5

superiority of our approach. Of particular significance is the substantial improvement in precision, a metric that gauges the model's ability to correctly identify positive cases. This enhancement underscores the robustness of our methodology in effectively distinguishing between instances of the target condition and non-instances, thereby minimizing false positives and maximizing diagnostic accuracy. Beyond accuracy and precision, Table 3. reveals improvements across a spectrum of other evaluation metrics, including AUC and specificity, among others. Each of these metrics offers unique insights into different aspects of model performance, providing a comprehensive assessment of our model's efficacy. Furthermore, the significance of our findings extends beyond mere numerical improvements; it speaks to the real-world implications of our research. By achieving superior performance on a benchmark dataset, our model holds the potential to revolutionize disease prediction practices, offering clinicians and healthcare practitioners a valuable tool for early detection and intervention. In summary, Table 3. encapsulates the success of our proposed model, highlighting its ability to outperform existing benchmarks and deliver tangible improvements in predictive accuracy. These results not only confirm the effectiveness of our method but also open up opportunities for progress in disease prediction and healthcare provision.

The ROC curve graphically represents the discriminatory power of the model, vividly showcasing its effectiveness in distinguishing between positive and negative instances. In Figure 7, the ROC plot reveals a prominent green line positioned significantly above the reference line, indicative of a model with robust discriminatory capabilities. This visual depiction underscores the reliability of our model in accurately categorizing instances, particularly those belonging to the positive class. Moreover, the impressive AUC score of 0.88 further solidifies the model's performance, reflecting a high degree of accuracy in correctly identifying true positives while minimizing false positives. This metric serves as a quantitative validation of our model's efficacy, demonstrating its proficiency in distinguishing between cases with and without the targeted condition. The elevated AUC value not only signifies the model's strength but also instills confidence in its ability to make informed predictions, thereby aiding clinicians and healthcare practitioners in decision-making processes. Overall, the combination of a visually compelling ROC curve and a commendable AUC score corroborates the model's excellence in discriminatory capability, highlighting its potential as a valuable tool in medical diagnostics and predictive analytics.

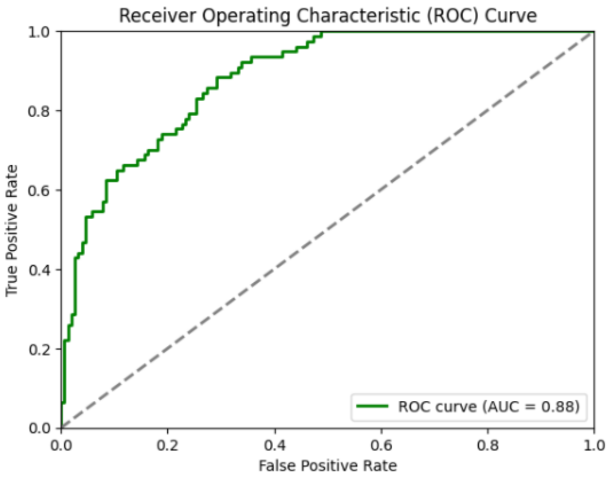


Figure. 7. Assessing Model Discriminatory Ability through ROC Curve Analysis

The proposed method enhances the interpretability and accuracy of diabetes prediction by combining traditional machine learning algorithms in an ensemble approach. Our ensemble model, which integrates SVM and Naïve Bayes classifiers, performs exceptionally well in distinguishing between diabetic and non-diabetic cases, as shown by the ROC curve and AUC score. By leveraging the strengths of SVM and Naïve Bayes, our method improves predictive accuracy while maintaining the interpretability needed for clinical applications. Unlike complex models, our ensemble offers a clear and effective solution without compromising on performance. However, the use of a single dataset, the Pima Indians Diabetes Dataset, may limit the generalizability of our findings. Future research should validate the model on additional datasets to ensure its robustness across different populations.

6. Disscusion

The Pima dataset, characterized by significant missing values, exemplifies the challenges inherent in clinical data collection, particularly in disease-related information gathering. These gaps threaten the accuracy of disease prediction models and necessitate strategic handling. Addressing this issue becomes crucial after identifying the most essential features, as premature imputation could distort their significance and lead to skewed results. In medical scenarios, the impact of features varies considerably; for instance, high blood pressure might be a more critical predictor of diabetes than skin thickness. To navigate this complexity, we employed the Random Forest technique, renowned for its robust feature importance discernment during model execution. Random Forest assigns importance scores to each feature during tree construction and class selection, enabling a data-driven approach to feature ranking. Utilizing these scores, we reduced our feature set to the five most impactful variables, thereby enhancing model accuracy while eliminating less significant features. This reduction was crucial, as our analysis indicated that using fewer than five features would compromise model accuracy. The diverse numerical range within the dataset posed additional challenges to accuracy, prompting the application of data normalization techniques to preserve model variance. In selecting an algorithm, we prioritized both interpretability and enhanced performance, which led us to adopt an ensemble approach combining SVM and Naïve Bayes. This ensemble method was chosen to harness the strengths of both algorithms: SVM's robustness in handling complex, high-dimensional data and Naïve Bayes' efficiency in probability estimation. The primary advantage of our approach lies in its ability to improve predictive accuracy while maintaining interpretability and computational efficiency. The ensemble method effectively balances performance with transparency. By integrating SVM and Naïve Bayes, we capitalize on the SVM’s capacity for complex feature separability, enhanced by the polynomial kernel trick, and the Naïve Bayes’ straightforward probabilistic approach. This combination allows for a more nuanced model that remains interpretable and clinically relevant, which is crucial for medical applications where understanding the reasoning behind predictions is essential. Our decision to use an ensemble approach was informed by the need to address both feature complexity and data volume. The ensemble method provides a robust solution that offers superior predictive performance and clarity in results. It effectively mitigates the limitations of relying solely on a single algorithm by leveraging their complementary strengths. Specifically, SVM excels in capturing non-linear relationships within the data, while Naïve Bayes enhances computational efficiency and probabilistic reasoning. In terms of feature selection and preprocessing, we carefully performed dimensionality reduction and evaluated class separability. Our approach resulted in a model that not only distinguishes diabetic from non-diabetic individuals with high accuracy but also provides a clear rationale for predictions, which is essential in clinical settings. The specificity metric further highlights the effectiveness of our ensemble method in reducing false positives, thereby potentially lowering both social and individual healthcare costs. In conclusion, our study demonstrates that combining traditional machine learning algorithms in an ensemble can effectively enhance predictive performance while maintaining interpretability. By addressing missing values post-feature selection and normalizing data ranges, we have improved model performance and showcased the value of strategic algorithm selection in advancing medical data classification.

7. Conclusion

Diabetes stands as one of the most formidable chronic diseases globally, fraught with various complications and a potential for significant mortality. The ability to predict and diagnose this ailment, particularly its type 2 manifestation, holds paramount importance as it facilitates proactive measures to prevent its associated complications. A delayed diagnosis not only poses a risk to individuals' health but also burdens society with substantial financial costs. To address this pressing need, we delved into one of the cornerstone datasets in this domain and introduced a novel model coupled with preprocessing techniques, resulting in an impressive model accuracy of 81.82%, surpassing existing models. Central to our approach was the identification of critical features, achieved through the utilization of the Random Forest algorithm. This method, through the construction of random trees for each class and subsequent majority voting, effectively pinpointed the most influential features. Consequently, we witnessed notable enhancements in precision (82.34%) and AUC (88.19%) metrics, underscoring the efficacy of our methodology. In our pursuit of model interpretability and enhanced performance, we adopted an ensemble approach combining traditional machine learning algorithms, specifically SVM and Naïve Bayes. This decision was made to leverage the strengths of both algorithms, improving predictive accuracy while maintaining clarity in model interpretation. Our goal is to support the integration of artificial intelligence systems into clinical practice, providing physicians with a robust tool for disease prediction tasks that is both effective and easy to understand. However, we encountered limitations in feature selection, as reducing the feature set to fewer than five could potentially compromise accuracy. Furthermore, we advocate for the adoption of ensemble algorithms to bolster evaluation metrics, thereby furnishing superior outcomes in disease prediction. These approaches necessitate thorough investigations into data preprocessing techniques to ensure optimal performance. Additionally, while acknowledging the potential of deep learning, we emphasize the importance of preserving model interpretability and maintaining control over the complexity of generated layers. Harnessing these methodologies could amplify the efficacy of disease prediction models. By incorporating web application designs tailored for diabetes prediction, we aim to enhance the accessibility and utilization of these invaluable tools in healthcare settings.

Declarations

Funding

This research did not receive any grant from funding agencies in the public, commercial, or non-profit sectors.

Authors' contributions

The contributions of Ghazaleh Kakavand Teimoory and Mohammad Reza Keyvanpour to this paper are as follows:
[GKT]: Conceptualization, Methodology, Writing-Original Draft, Visualization and Data Analysis.
[MK]: Supervision, Project administration, Guidance, Oversight, Expertise, Leadership and Last Edit.

Conflict of interest

The authors declare that no conflicts of interest exist.

Refrences

[1] I. Shaheen, N. Javaid, N. Alrajeh, Y. Asim, and S. Aslam, "Hi-Le and HiTCLe: Ensemble Learning Approaches for Early Diabetes Detection Using Deep Learning and Explainable Artificial Intelligence," *IEEE Access*, vol. 12, pp. 66516–66538, 2024, doi: 10.1109/ACCESS.2024.3398198.

[2] H. El-Sofany, S. A. El-Seoud, O. H. Karam, Y. M. Abd El-Latif, and I. A. T. F. Taj-Eddin, "A Proposed Technique Using Machine Learning for the Prediction of Diabetes Disease through a Mobile App," *International Journal of Intelligent Systems*, vol. 2024, 2024, doi: 10.1155/2024/6688934.

[3] "Classification and diagnosis of diabetes: Standards of Medical Care in Diabetes-2020," *Diabetes Care*, vol. 43, pp. S14–S31, Jan. 2020, doi: 10.2337/dc20-S002.

[4] E. Oghenekome Paul, "Hybrid decision tree-based machine learning models for diabetes prediction," *SCIREA Journal of Information Science and Systems Science*, Jan. 2024, doi: 10.54647/isss120327.

[5] Z. Zhang *et al.*, "A novel evolutionary ensemble prediction model using harmony search and stacking for diabetes diagnosis," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 1, Jan. 2024, doi: 10.1016/j.jksuci.2023.101873.

[6] E. Adua *et al.*, "Predictive model and feature importance for early detection of type II diabetes mellitus," *Translational Medicine Communications 2021 6:1*, vol. 6, no. 1, pp. 1–15, Aug. 2021, doi: 10.1186/S41231-021-00096-Z.

[7] F. Serpush, M. R. Keyvanpour, and M. B. Menhaj, "Remote elderly healthcare: a robust deep learning approach for wearable sensors-based complex activities recognition," *AUT Journal of Modeling and Simulation AUT J. Model. Simul*, vol. 55, no. 1, pp. 109–126, 2023, doi: 10.22060/miscj.2023.21984.5308.

[8] H. Qi, X. Song, S. Liu, Y. Zhang, and K. K. L. Wong, "KFPredict: An ensemble learning prediction framework for diabetes based on fusion of key features," *Comput Methods Programs Biomed*, vol. 231, Apr. 2023, doi: 10.1016/j.cmpb.2023.107378.

[9] S. Mehrmolaei, M. Savargiv, and M. R. Keyvanpour, "Hybrid learning-oriented approaches for predicting Covid-19 time series data: A comparative analytical study," *Eng Appl Artif Intell*, vol. 126, p. 106754, Nov. 2023, doi: 10.1016/J.ENGAPPAI.2023.106754.

[10] M. R. Keyvanpour, B. Pourebrahim, and S. Mehrmolaei, "EADR: an ensemble learning method for detecting adverse drug reactions from twitter," *Soc Netw Anal Min*, vol. 14, no. 1, Dec. 2024, doi: 10.1007/s13278-024-01239-4.

[11] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Comput Appl*, vol. 35, no. 22, pp. 16157–16173, Aug. 2023, doi: 10.1007/s00521-022-07049-z.

[12] M. Badawy, N. Ramadan, and H. A. Hefny, "Healthcare predictive analytics using machine learning and deep learning techniques: a survey," *Journal of Electrical Systems and Information Technology*, vol. 10, no. 1, Aug. 2023, doi: 10.1186/s43067-023-00108-y.

[13] K. R. Tan *et al.*, "Evaluation of Machine Learning Methods Developed for Prediction of Diabetes Complications: A Systematic Review," *https://doi.org/10.1177/19322968211056917*, vol. 17, no. 2, pp. 474–489, Nov. 2021, doi: 10.1177/19322968211056917.

[14] Z. Zrubka *et al.*, "The Reporting Quality of Machine Learning Studies on Pediatric Diabetes Mellitus: Systematic Review," *J Med Internet Res*, vol. 26, no. 1, p. e47430, Jan. 2024, doi: 10.2196/47430.

[15] R. Roshankar and M. R. Keyvanpour, "Spatio-Temporal Graph Neural Networks for Accurate Crime Prediction," in *2023 13th International Conference on Computer and Knowledge Engineering, ICCKE 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 168–173. doi: 10.1109/ICCKE60553.2023.10326223.

[16] S. Gündoğdu, "Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique," *Multimed Tools Appl*, vol. 82, no. 22, pp. 34163–34181, Sep. 2023, doi: 10.1007/s11042-023-15165-8.

[17] J. Hu and S. Szymczak, "A review on longitudinal data analysis with random forest," *Brief Bioinform*, vol. 24, no. 2, pp. 1–11, Mar. 2023, doi: 10.1093/BIB/BBAD002.

[18] H. Blockeel, L. Devos, B. Frénay, G. Nanfack, and S. Nijssen, "Decision trees: from efficient prediction to responsible AI," *Front Artif Intell*, vol. 6, p. 1124553, Jul. 2023, doi: 10.3389/FRAI.2023.1124553/BIBTEX.

[19] X. Yuan, S. Liu, W. Feng, and G. Dauphin, "Feature Importance Ranking of Random Forest-Based End-to-End Learning Algorithm," *Remote Sensing 2023, Vol. 15, Page 5203*, vol. 15, no. 21, p. 5203, Nov. 2023, doi: 10.3390/RS15215203.

[20] K. Oliullah, M. H. Rasel, M. M. Islam, M. R. Islam, M. A. H. Wadud, and M. Whaiduzzaman, "A stacked ensemble machine learning approach for the prediction of diabetes," *J Diabetes Metab Disord*, vol. 23, no. 1, pp. 603–617, Jun. 2024, doi: 10.1007/s40200-023-01321-2.

[21] Z. Zhang *et al.*, "A novel evolutionary ensemble prediction model using harmony search and stacking for diabetes diagnosis," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 1, p. 101873, Jan. 2024, doi: 10.1016/J.JKSUCI.2023.101873.

[22] H. B. Kibria, M. Nahiduzzaman, M. O. F. Goni, M. Ahsan, and J. Haider, "An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI," *Sensors*, vol. 22, no. 19, Oct. 2022, doi: 10.3390/s22197268.

[23] M. O. Edeh *et al.*, "A Classification Algorithm-Based Hybrid Diabetes Prediction Model," *Front Public Health*, vol. 10, Mar. 2022, doi: 10.3389/fpubh.2022.829519.

[24] A. A. Jasim, L. R. Hazim, H. Mohammedqasim, R. Mohammedqasem, O. Ata, and O. H. Salman, "e-Diagnostic system for diabetes disease prediction on an IoMT environment-based hyper AdaBoost machine learning model," *Journal of Supercomputing*, pp. 1–26, Apr. 2024, doi: 10.1007/S11227-024-06082-0/TABLES/4.

[25] M. S. Reza, U. Hafsha, R. Amin, R. Yasmin, and S. Ruhi, "Improving SVM performance for type II diabetes prediction with an improved non-linear kernel: Insights from the PIMA dataset," *Computer Methods and Programs*

in Biomedicine Update, vol. 4, p. 100118, Jan. 2023, doi: 10.1016/J.CMPBUP.2023.100118.

[26] S. P. Chatrati *et al.*, “Smart home health monitoring system for predicting type 2 diabetes and hypertension,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 3, pp. 862–870, Mar. 2022, doi: 10.1016/J.JKSUCI.2020.01.010.

[27] M. Kawarkhe and P. Kaur, “Prediction of Diabetes Using Diverse Ensemble Learning Classifiers,” in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 403–413. doi: 10.1016/j.procs.2024.04.040.

[28] J. Abdollahi and S. Aref, “Early Prediction of Diabetes Using Feature Selection and Machine Learning Algorithms,” *SN Comput Sci*, vol. 5, no. 2, pp. 1–26, Feb. 2024, doi: 10.1007/S42979-023-02545-Y/METRICS.

[29] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, “Diabetes prediction using machine learning and explainable AI techniques,” *Healthc Technol Lett*, vol. 10, no. 1–2, pp. 1–10, Feb. 2023, doi: 10.1049/htl2.12039.

[30] A. Perdana, A. Hermawan, and D. Avianto, “Analyze Important Features of PIMA Indian Database For Diabetes Prediction Using KNN,” *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 12, no. 1, pp. 70–75, Mar. 2023, doi: 10.32736/sisfokom.v12i1.1598.

[31] K. Alnowaiser, “Improving Healthcare Prediction of Diabetic Patients Using KNN Imputed Features and Tri-Ensemble Model,” *IEEE Access*, vol. 12, pp. 16783–16793, 2024, doi: 10.1109/ACCESS.2024.3359760.

[32] M. E. Febrian, F. X. Ferdinan, G. P. Sendani, K. M. Suryanigrum, and R. Yunanda, “Diabetes prediction using supervised machine learning,” *Procedia Comput Sci*, vol. 216, pp. 21–30, Jan. 2023, doi: 10.1016/J.PROCS.2022.12.107.

[33] N. N. N. Nazirun *et al.*, “Prediction Models for Type 2 Diabetes Progression: A Systematic Review,” *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3432118.

[34] N. Katiyar, H. K. Thakur, and A. Ghatak, “Recent advancements using machine learning & deep learning approaches for diabetes detection: a systematic review,” *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol. 9, p. 100661, Sep. 2024, doi: 10.1016/J.PRIME.2024.100661.

[35] G. Dharmarathne, T. N. Jayasinghe, M. Bogahawaththa, D. P. P. Meddage, and U. Rathnayake, “A novel machine learning approach for diagnosing diabetes with a self-explainable interface,” *Healthcare Analytics*, vol. 5, p. 100301, Jun. 2024, doi: 10.1016/J.HEALTH.2024.100301.

[36] M. Ramanna Lamani and T. A. Gondhale, “Enhancing Diabetes Prediction Accuracy with AdvanSVM: A Machine Learning Approach Using the PIMA Dataset,” *Seybold Report Journal*, vol. 19, no. 05, 2024, doi: 10.5110/77.

Mohammad Reza Keyvanpour is a professor at Alzahra University, Tehran, Iran. His academic journey includes a B.S. in software engineering from Iran University of Science & Technology and his M.S. and Ph.D. in software engineering from Tarbiat Modares University. His research spans information retrieval and data mining and his mainly interest is in E-Health; Tehran; Iran;keyvanpour@alzahra.ac.ir

Ghazaleh Kakavand Teimoory her B.S. in software engineering from Information and Communication Technology University, Tehran, Iran, and is currently working toward her master's degree in Software Engineering, actively engaging in research at the Department of Computer Engineering and Data Mining Lab at Alzahra University, Tehran, Iran. Her research interests are in data mining and its applications, Diabetes Prediction, and diseases analysis and E-health, which contribute significantly to the concept and writing of this article; Tehran, Iran; gh.kakavandteimoory@gmail.com