Shahid Bahonar University of Kerman
Journal of Mahani Mathematical Research
Print ISSN: 2251-7952 Online ISSN: 2645-4505

AN OVERVIEW OF DIABETES DIAGNOSIS METHODS ON THE PIMA INDIAN DATASET

F. Heydari, M. Kuchaki Rafsanjani ✉, and M. Sheikh Hosseini Lori

Article type: Research Article
(Received: 29 November 2022, Received in revised form 29 July 2023)
(Accepted: 06 October 2023, Published Online: 18 October 2023)

Abstract. In recent years, data mining and machine learning methods in the medical field have received much attention and have optimized many complex issues in the medical field. One of the problems facing researchers is the appropriate dataset, and the suitable dataset on which different methods of data mining and machine learning can be applied is rarely found. One of the most reliable and appropriate datasets in the field of diabetes diagnosis is the Indian Survey Database. In this article, we have tried to review the methods that have been implemented in recent years using machine learning classification algorithms on this data set and compare these methods in terms of evaluation criteria and feature selection methods. After comparing these methods, it was found that models that used feature selection methods were more accurate than other approaches

Keywords: Diabetes, Machine learning, Data mining algorithms, Detection accuracy, Pima Indian dataset.
2020 MSC: Primary 68T05, 68T09, 94A16.

1. Introduction

Diabetes is a group of metabolic diseases that increase blood sugar levels in the affected person due to insufficient production of insulin in the body and inadequate response of cells to the produced insulin. Despite this disease's high prevalence and spread, no method has been found to treat and eliminate this disease. One of the important applications of data mining is its application in medical diagnosis. When the number of parameters of a disease increases, it also becomes difficult for doctors to diagnose it. For this reason, the use of medical diagnostic tools is strongly felt. Today, medical databases store large amounts of patient information, and extract medical knowledge from this data with data mining methods can help physicians rapidly diagnose the disease. To prevent diabetes, more susceptible people should be identified; for example, people with a family history of diabetes are more susceptible.

Globally, 11.3% of deaths are due to diabetes. About half of these deaths are in people under 60 years of age. According to data from the World Health

✉ kuchaki@uk.ac.ir, ORCID: 0000-0002-3220-4839
DOI: 10.22103/jmmr.2023.20621.1370
Publisher: Shahid Bahonar University of Kerman
How to cite: F. Heydari, M. kuchaki Rafsanjani, and M. Sheikh Hosseini Lori, *An overview of diabetes diagnosis methods on the Pima Indian dataset*, J. Mahani Math. Res. 2024; 13(1): 417 - 441.

Organization, Diabetes caused 4.6 million deaths in 2011 and is projected to be the seventh leading cause of death in 2030. More than 80% of diabetes deaths occur in low- and middle-income countries [11]. The International Diabetes Federation (IDF) has released the latest data on diabetics in 2019, showing that half of the 463 million adults with diabetes are unaware of their disease. Because of this, they are at risk for serious complications associated with diabetes. The highest number of people with diabetes in 2019 is in China, with 116.4 million people, and the number of diabetics in this country is projected to reach 140.5 million by 2030. Moreover, the Middle East and North Africa region have the highest prevalence of the disease, and it is predicted that the prevalence of the disease in this region will reach 13.3% by 2030 [25]. Diabetes mellitus affects about 425 million people worldwide and is projected to increase to 629 million by 2045 [12].

According to research, diabetes can be diagnosed through various methods of machine learning. However, AI-related (Artificial intelligence-related) approaches are more important and increase the accuracy of prediction results. These approaches have advantages such as low cost, fast calculation, and good accuracy. In different decades, machine learning methods and data mining have been used continuously. Many people in the community are unaware of their health status. Thus, using machine learning and data mining techniques can help diagnose patients more quickly. Data mining has performed very well in human life, such as customer relationship management, market analysis, meteorological forecasting, but in medicine. However, the volume of data is very high, by analyzing the initial data of the patient it can help the doctor a lot in diagnosing the disease [3, 8].

When the body cannot produce enough insulin or use insulin correctly, diabetes develops and eventually becomes fatal. On the other hand, artificial intelligence has been used in various sciences and has proven its efficiency, and medical sciences are no exception to this rule, so we decided to adopt the methods used in recent years to diagnose diabetes. Pima Indian dataset has been more used as a dataset and we tried to examine the methods that have used this dataset. So far, various methods for diagnosing diabetes have been proposed using machine learning algorithms and data mining with high diagnostic accuracy. But there are still better models for diagnosing diabetes. These models can be improved by using some tasks such as deleting junk data and selecting or extracting the most important features. One of the most important and famous datasets in the field of diabetes detection is Pima Indian dataset, which has been researched a lot. In this article, the purpose is to review the methods performed on this dataset using machine learning and data mining algorithms:

- In this survey, a new classification is provided for methods of diabetes diagnosis on the Pima Indian dataset.
- An overview of machine learning classification and clustering algorithms is presented.

- Various features and relationships between the features of the Pima Indian diabetes dataset are investigated.
- A comprehensive view of evaluation criteria in diabetes diagnosis methods based on feature selection is provided.

The remainder of this paper is organized as follows: In Section 2, diabetes and all types are explained. In Section 3, the Pima Indian diabetes dataset is presented. In Section 4, the Data mining discussed methods are compared in terms of selection criteria, and also the advantages and disadvantages of each method are stated. In Section 5 various works have been done in Pima Indian diabetes dataset investigated. In Section 6 summary of work done in the field of diabetes diagnosis is provided and finally, the conclusion comes in last Section.

2. Diabetes disease

Diabetes is a chronic disease in which the pancreas is no longer able to make insulin or when the body cannot make good use of the insulin it produces. Insulin is a hormone made by the pancreas that acts as a key to transfer glucose from the food we eat from the bloodstream into the body's cells to produce energy. Inability to produce or use insulin effectively leads to an increase in blood glucose levels (known as hyperglycemia). In the long run, high glucose levels are associated with damage to the body and failure of various organs and tissues [26]. When people with diabetes develop a viral infection, it can be more difficult to treat because of fluctuations in blood sugar levels and the possible complications of diabetes [21].

There are three main types of diabetes: type 1 diabetes, type 2 diabetes and gestational diabetes [21, 26].

`Type 1 Diabetes:` Type 1 diabetes can affect people of any age, but it usually affects children or adults. People with type 1 diabetes need daily insulin injections to control their blood sugar levels. Risk factors for type 1 diabetes are still under investigation. However, having a family member with type 1 diabetes slightly increases the risk of developing the disease. Environmental factors and exposure to some viral infections are also associated with a higher risk of developing type 1 diabetes. The most common symptoms of type 1 diabetes are: abnormal thirst and dry mouth, sudden weight loss, frequent urination, lack of energy, fatigue, persistent hunger, blurred vision, nocturia.

Diagnosis of type 1 diabetes can be difficult, so additional tests may be needed to confirm the diagnosis. People with type 1 diabetes need daily insulin treatment, regular blood sugar monitoring, and a healthy lifestyle to effectively manage their condition. There is currently no effective and safe way to prevent type 1 diabetes

`Type 2 Diabetes:` Type 2 diabetes is more common in adults and accounts for about 90% of all cases of diabetes. In this type of diabetes, the body does not use produced insulin well. The best treatment for type 2 diabetes is a healthy lifestyle, including increased physical activity and a healthy diet. Over time,

however, most people with type 2 diabetes need oral medications or insulin to control their blood sugar levels.

The disease is usually characterized by insulin resistance, because insulin cannot work properly, blood glucose levels rise and more insulin is released. Type 2 diabetes is more commonly diagnosed in the elderly, but is more common in children, adolescents, and middle-aged people due to overweight, inactivity, and poor diet. The most important way to control type 2 diabetes is a healthy diet and increased physical activity. The symptoms of type 2 diabetes are similar to those of type 1 diabetes and, these include excessive thirst and dry mouth, frequent urination, deficiency of energy, fatigue, ulcers with slow healing, recurrent skin infections, blurred vision, tingling or numbness in the hands and feet. These symptoms can be mild or hidden, so people with type 2 diabetes may live with the disease for several years before being diagnosed.

Research shows that in most cases (about 80% according to some studies), type 2 diabetes can be prevented through a healthy diet and regular physical activity. Regular physical activity is essential to help control blood sugar levels. The best way to prevent type 2 diabetes is to combine aerobic exercise (jogging, swimming, cycling) and physical activity.

`Gestational Diabetes Mellitus (GDM):` Gestational diabetes is a type of diabetes that is caused by high blood sugar during pregnancy and has side effects for the mother and her baby. GDM usually disappears after pregnancy, but affected women and their children are at risk for postpartum type 2 diabetes. Women with gestational diabetes are subsequently at higher risk for type 2 diabetes, especially three to six years after giving birth. Exposure to blood sugar in the womb puts babies at risk of being overweight or obese, associated with type 2 diabetes. Many women with GDM experience pregnancy-related complications, including high blood pressure and overweight babies. Many women with GDM experience pregnancy-related complications, including high blood pressure and overweight babies [21]. It is crucial for women with gestational diabetes to carefully monitor their blood sugar levels to reduce the risk of adverse pregnancy outcomes with their health care

3. **Pima Indian Diabetes Dataset**

Pima Indian diabetes dataset is a well-known dataset on type 2 diabetes from the University of California Repository (UCI) [27]. Since 1965, the National Institute of Diabetes and Gastroenterology and Kidney Diseases has continuously monitored the disease due to its high prevalence around Arizona. As a gold standard diagnostic test for diabetes, a two-hour blood glucose test was performed on each person for two years to determine whether the person's blood sugar level was above 200 or not. The person was diagnosed with diabetes between one and five years after the examination. 768 cases were selected for analysis, of which 268 had diabetes. The binary class variable is "0" for

500 samples, including non-diabetics, and it is "1" for 268 samples, including people with diabetes.

Nowadays, most people in the world have a similar lifestyle in which low physical activity and processed foods cause diabetes. Therefore, many people are prone to diabetes. Although there are now larger, more complex diabetes datasets, the Pima Indian Diabetes dataset has remained a benchmark for diabetes classification research. Given the presence of a binary outcome variable, the dataset naturally lends itself to supervised learning and, in particular, logistic regression. However, various ML algorithms have been employed to produce classification models based on this dataset for not being limited to a singular type of model [9].

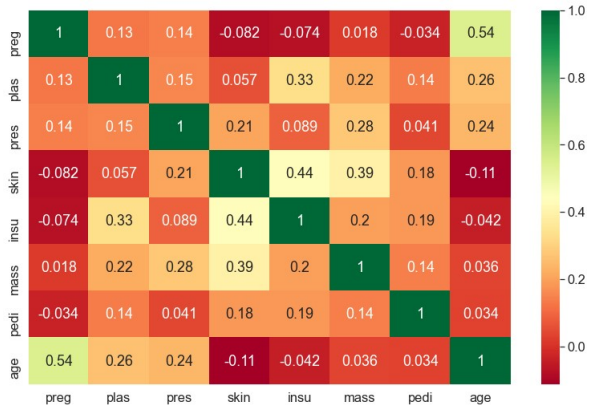

FIGURE 1. Relationship between Features of Indian Pima Database.

Figure 1 shows the relationship between the characteristics of the Pima Indian diabetes dataset. Green indicates more connections between features, and red indicates no connection between features. Figure 2 shows each of the positive and negative classes' sample numbers. Other explanations are about missing values are minimum values, maximum values, number of unique samples in Table1 [16].

4. **Data Mining**

Data mining emerged in the mid-1990s as a new approach to discovering hidden knowledge and analyzing data. The term data mining was first used in late 2009 for medical purposes. Larose's book provides a practical definition of data mining, meaning that data mining is often an analysis of big data to find obscure relationships and summarize data in a way that is Comprehensible to the data owner [34]. Data mining aims to gain new and in-depth insights and understand the large datasets that can be used to support the decision.

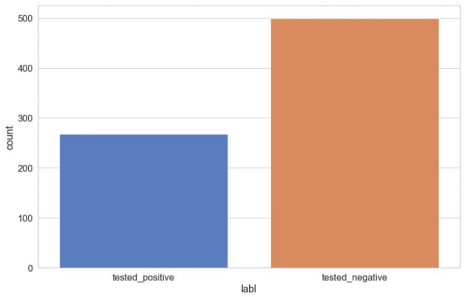

FIGURE 2. Number of positive class and negative class samples.

TABLE 1. Detailed description of examples in the Indian Diabetes Database

Feature Description	Abbreviation	Unit	Type	Percentage of unique samples	Mean	Min
Number of pregnancies	Preg	-	Numerical	2	3.845	0
Plasma glucose concentration	Plas	Mg/Di	Numerical	19	120.895	0
Diastolic blood pressure	Pres	Mm-Hg	Numerical	8	69.105	0
Thickness of triceps skin fold	Skin	Mm	Numerical	5	20.536	0
2-hour serum insulin	Insu	Mm U/Ml	Numerical	93	79.799	0
Body mass index	Mass	Kg/M2	Numerical	76	31.993	0
Function of diabetes derivatives	Pedi	-	Numerical	346	0.472	0.07
Age	Age	-	Numerical	5	33.241	21
Class of Diabetes	Diabetes	-	Nominal	0	-	-

Table 2. Detailed description of examples in the Indian Diabetes Database

Feature Description	Max	Standard	Missing	Distinctive samples
Number of pregnancies	17	3.37	0	17
Plasma glucose concentration	199	31.973	0	136
Diastolic blood pressure	122	19.356	0	41
Thickness of triceps skin fold	99	15.952	0	51
2-hour serum insulin	846	115.244	0	186
Body mass index	67.1	7.884	0	246
Function of diabetes derivatives	2.42	0.331	0	517
Age	81	11.76	0	52
Class of Diabetes	-	-	0	2

As this definition implies, data mining aims to gain new and profound insights and understand large datasets (often stored for operational purposes), which can be used to support decision-making. Data mining can also enable the production of scientific hypotheses from sizeable experimental datasets and from the biomedical literature [37, 40]. Data mining in medical applications is different from other applications.

Data mining processing aims to extract information from a dataset and turn it into a comprehensible structure for further use. This process has become a widespread activity in all areas of medical science research. Data mining involves a series of steps that are used automatically or semi-automatically to extract and discover interesting, unknown, and hidden features of large amounts of data. Data mining has been successful in various fields in human society. However, disease prediction and medical data analysis applications can still be improved. The most crucial advantage of information technology is that a huge stock of previous patient records is constantly maintained and controlled by hospitals for various referrals. This medical data helps physicians examine different patterns in the dataset. Patterns in the dataset may be used to classify, predict, and diagnose diseases [44].

4.1. **Differences between the application of data mining in medicine and other fields.** Medical information is generated primarily through the provision of patient care. Therefore, the extraction of medical data inevitably involves privacy and legal issues. For this reason, data mining in biomedical and healthcare fields is very different from what is done in other fields. This fundamental difference requires a discussion of the uniqueness of data mining in the fields of medicine and health [4, 15].

First, in many cases, the quality of data in biomedicine and healthcare is lower than that found in other fields for many reasons. Second, health care researchers must ensure patients' privacy and use patient data following the Health Insurance Portability and Accountability Act (HIPAA). Health care applications are critical to data mining, ensuring patient safety, maintaining the security and confidentiality of sensitive information for dataset users. Third, there are legal considerations for using health care data. For example, using medical data mining can reveal previously unknown medical errors, which in turn can lead to lawsuits against health care providers.

4.2. **Data mining algorithms.** Before using data mining algorithms, researchers need to understand which data mining algorithm works best for their work based on the dataset they use for their research. They also need to know what types of data mining algorithms are available and how they work. Data mining algorithms are divided into four categories: clustering (unsupervised learning), classification (Supervised learning), semi-supervised learning, and reinforcement learning [18, 43]. In this paper, some classification and clustering algorithms are given. An overview of the data mining classification and clustering algorithms discussed in this paper is given in Figure 3.

4.2.1. *Neural Networks.* A neural network consists of several layers that are connected and interconnected. The first layer is the input, and the last layer is the output, which is connected by a graph of nodes and weighted edges. There are hidden layers between the input and output layers, and the neural network establishes the relationship between input and output [22].

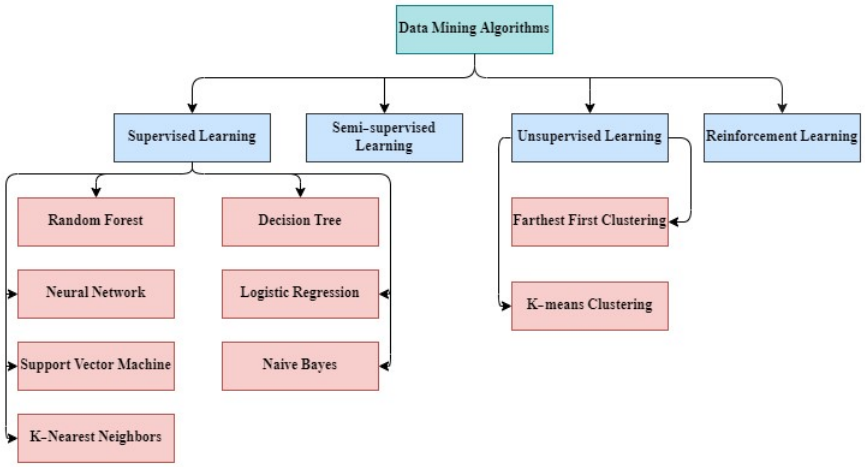

FIGURE 3. Overview of classification and clustering algorithms discussed

4.2.2. *Decision Tree (DT).* Decision tree classification is used as a well-known classification method. The decision tree is a flowchart tree structure in which the inner node represents a property or attribute, the branch represents a decision rule, and each leaf node represents the result. The highest node in a decision tree is known as the root node. Also, decision trees are ideal for creating nonlinear relationships between features and classes. Regularly, an impurity function is determined to evaluate the quality of each node for division, and the Gini Diversity Index is used as a known measure of total performance. In practice, the decision tree is flexible in the sense that it can easily model nonlinear or unconventional relationships. This can interpret the interaction between predictions. The decision tree can also be interpreted well due to its binary structure. However, the decision tree has several drawbacks that tend to overfit data [19, 47]. The decision tree is one of the most popular classification algorithms that has been used to solve several problems in different regions. In addition, unlike other algorithms that may require traffic distribution models and parameters, no prior knowledge of traffic is required.

4.2.3. *Support Vector Machine (SVM).* Support vector machines (SVM) are one of the most popular classifications. A support vector machine is a supervised learning algorithm used to control and solve the classification problem. SVM is a supervised learning algorithm proposed by Corinna Cortes and Vapnik (1995) and is one of the most popular approaches to classification in learning. This program is applied to various programs such as pattern recognition, text classification, image classification, etc [6,10,29]. The basic concept of SVM depends on minimizing structural risk. It uses nonlinear mapping to transfer

the input training pattern to a remarkable high-dimensional space where the optimal separator super plane can be found.

4.2.4. *Naive Bayes.* Naive Bayes is a Bayesian-based probability classifier. The Naive Bayes model is compatible with further analysis of very large datasets. The Naive Bayes classifier is very simple and skilled and performs very well in complex issues. This algorithm has many applications due to its simplicity in various fields such as medicine, pattern recognition, image processing, and weather forecasting. Naive Bayes allows each feature to contribute equally and independently to the other features in the classification process. The problem with this algorithm is that the high simplicity of this algorithm increases the computation but has high accuracy [28, 48].

4.2.5. *K- Nearest Neighbor (KNN).* K-Nearest Neighbor is one of those algorithms that is easy to understand and works amazingly well in practice. It is a non-parametric method used in data mining, machine learning, and pattern recognition. In this algorithm, the units are placed next to each other and respond to the input vector. For one or two dimensions, data clustering is very easy. K-Nearest Neighbor can be used for both classification and regression prediction. The K-Nearest Neighbor algorithm classifies a test sample based on the K-Nearest Neighbor. Tutorial samples are presented as vectors in multidimensional feature space. The space is divided into areas with tutorial samples. A point in space belongs to a class that the most tutorial points belong to that class is in its nearest tutorial sample K. K is a valid positive value and is generally small. K is usually considered an odd number because it prevents equal votes. In both cases, K contains the nearest tutorial sample in the data space, and its output varies depending on the type used in classification and regression. In the classification mode, according to the value specified for K, calculates the distance of the point we want to specify its label with the nearest points, and according to the maximum number of votes of these neighboring points, decides on the label of the desired point. Various methods can be used to calculate this distance. K-Nearest Neighbor is known as a lazy and non-parametric learner because it only stores the trained database, and no general model of the training database is created [23].

4.2.6. *Random Forest (RF).* As its name implies, the Random Forest consists of a large number of individual decision trees that act as a group to make output decisions. Each tree in a random forest determines the class prediction, and the result will be the most predicted class among the decision trees. The reason for this fantastic result of random forest is that the trees protect each other from individual mistakes. Although some trees may predict the wrong answer, many trees correct the final prediction so that the trees can move in the right direction as a group. Random forests, combined with many poor language learners, reduce redundancy in the education complex. Random forests can control a large number of variables in a dataset. Random forest performance is usually

better than the decision tree, but this performance improvement depends partly on the type of data [35]. The main disadvantage of random forests is non-reproducibility because the forest construction process is random [5].

4.2.7. *Logistic Regression Algorithm.* Logistic regression is a classification algorithm used to assign observed samples to a separate set of classes. Unlike linear regression, which produces continuous numerical values, logistic regression modifies its output logistic function using the sigmoid function to return a possible value that can be plotted in two or more separate classes. Logistic regression works well when the relationship in the data is almost linear, but it performs poorly if there are nonlinear relationships between the variables. In addition, more statistical assumptions are needed before using other techniques. In addition, more statistical assumptions are needed before using other techniques [20, 47].

4.2.8. *K-means clustering.* Hartigan and Wong developed the k-means clustering algorithm in 1979 [24]. This algorithm divides the dataset according to a specific cluster k. This algorithm consists of two independent steps. The first step calculates the centers k and in the second step, features the data to the nearest centers based on the Euclidean distance, which is the most common way to calculate. Once the clustering is complete, the new center of each cluster is recalculated. Based on this center, the new Euclidean distance between each center and each data point is calculated, and the point whose Euclidean distance is less than the center assigned to that cluster, and these steps are repeated until the distance between the means of the two consecutive stages is less than the desired level of sensitivity. The k algorithm has a simple average execution, and the disadvantage is that the quality of clustering results depends to some extent on the arbitrary choice of the primary centers, and if the primary centers are selected randomly, different results will be obtained based on the primary centers [45].

4.2.9. *The First Farthest Clustering Algorithm.* Hochbaum and Shmoys [42] proposed the first farthest clustering algorithm. Like K-means clustering, the first clustering algorithm operates in the farthest two steps. The first step is to select the primary centers randomly, and the second step is to assign the data to the cluster centers based on the primary centers. The first clustering algorithm first randomly selects the farthest point, then selects the next center with the farthest point from the current center, and the next centers are selected according to the same procedure. The farthest centers are from the collection of centers that have already been selected. When a preferred center number K is selected, the algorithm assigns all other data points to the cluster that is identified and terminated by the nearest center. Thus, unlike K-means clustering, the farthest first clustering algorithm requires only one call to cluster a set of data points. Since the position of any ordinary feature is not calculated to modify the centers, and all the central points are real data points and not the

geometric centers of the clusters, the farthest clustering algorithm differs from the K-means clustering. Although this algorithm starts with random selection and runs with just one call, it works well in terms of selecting centers [16, 41].

4.2.10. *Genetic Algorithm (GA).* A Genetic algorithm is an iterative process of selecting, crossing, and mutating populations in each iteration called a generation. Each chromosome or individual in a linear string (usually 0 and 1) of constant length is encoded in genetic similarity. In the search space, first of all, individual members of the population are randomly initialized. In the search space, first of all, individual members of the population are randomly initialized. After initialization, each member of the population is solved according to the objective function and a number (value of the objective function) is determined, which indicates the readiness for survival for the person concerned. The GA maintains a fixed number of individuals with the corresponding proportion value. To produce a new generation, people are used in the current population who have a better evaluation function, which allows a better generation to be created. In the production of a new generation with a very low probability of mutation to be able to maintain diversity in the population. In an iterative process, the current generation acts as the parent of the next generation. Thus, it is expected that successive generations have a more appropriate evaluation function. This process is repeated until the desired number of generations is created or the value of the desired evaluation function is obtained [32].

5. Related works on diagnose Diabetes

In recent years, data mining and machine learning methods in the medical field have received much attention and have optimized many complex issues in the medical field. Various works have been done in Pima Indian diabetes dataset. Many researchers have considered the methods of selecting and extracting features to select or extract the most important features. The genetic algorithm has been given much attention for selecting the most important features. Moreover, using hybrid methods, many researchers have improved the accuracy of diagnosis of this disease, in which the support vector machine algorithm has a much better performance than other machine learning algorithms due to the nature of the data. Also, each researcher has considered different methods for data preprocessing.

5.1. **Stacked ensemble-based type-2 diabetes prediction using machine learning techniques.** Rahim et al. [36] proposed a robust approach based on the stacked ensemble method using several machine learning algorithms on the Pima Indian dataset. In this approach, firstly, pre-processing is done on the dataset. The presented model consists of both basic and meta models. The base model includes K-nearest neighbors, Naive Bayes, Random Forest, and Support Vector Machine (SVM) algorithms and meta-models that perform the final prediction through logistic regression using the prediction

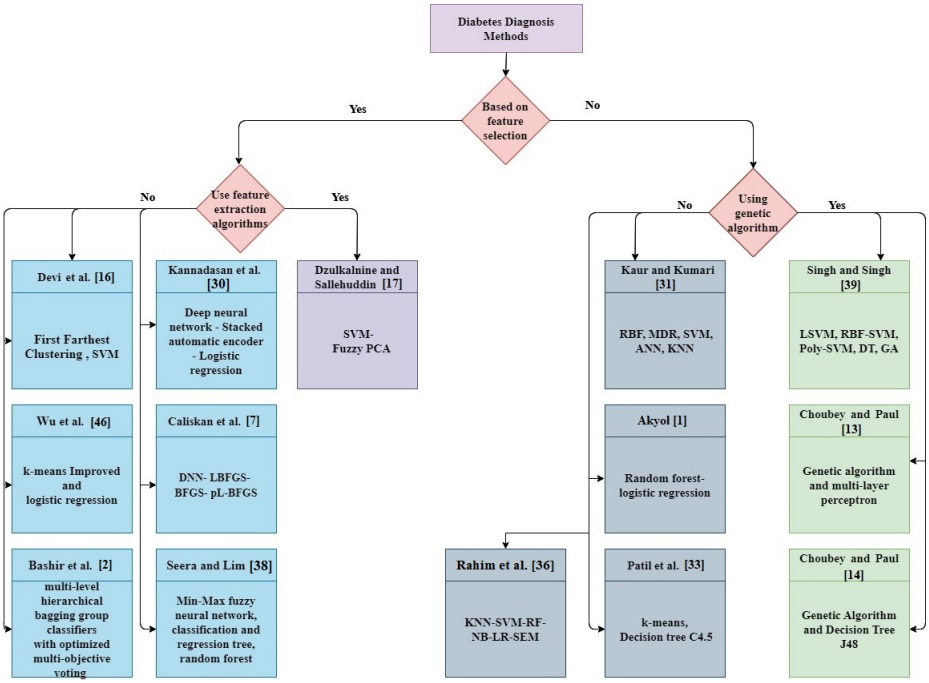

FIGURE 4. An overview of the methods discussed in the diagnosis of diabetes.

outputs of the base models. Each base model is trained on a specific dataset and provides an intermediate prediction. Secondly, the meta-model takes the intermediate prediction as an input feature and provides the final output for the target value. In this method, K-fold cross-validation is used to train the proposed model, and 70% of the data is used for training and the rest of the data is used for testing the model. The presented model has obtained an accuracy of 94.17, precision of 95.92, F-score of 95.43, and recall of 94.95.

5.2. **Evolutionary group learning system based on accumulation and genetic algorithm with NSGA-II-Stacking.** Singh and Singh [39] developed an evolutionary group learning system based on accumulation and genetic algorithms with faulty sorting to predict type 2 diabetes on the Pima Indian diabetes dataset. In this method, the data is first pre-processed, then four types of learners (LSVM, RBF-SVM, Poly-SVM, DT) are used as base-level learners, and teach each of these models with five bootstrapped samples produced by K-fold cross-validation. Then, the NSGA-II genetic algorithm uses 20 basic trained classifiers to select the models. NSGA-II is used to select the model by identifying the composition with the highest accuracy and most minor complexity and predicting training samples. Predictions of training instances

with actual tags constitute Level 1 data and are used for classification training (e.g., KNN). The meta classifier is then used to predict the test samples. To understand the efficiency of the NSGA-II-Stacking model, individual and group models have been selected as benchmark models. The proposed NSGA-II-Stacking model is better than the group and individual benchmark models in terms of evaluation criteria; This model has an accuracy of 83.8%.

5.3. **Hybrid method for diagnosing diabetes using the support vector machine algorithm and the farthest first clustering algorithm.** Using support vector machine algorithms and the first clustering algorithm, Devi et al. [16] Provided the farthest way to diagnose diabetes. Since the dataset used for the proposed model contains outlier data, in the data preprocessing stage, The interquartile range (IQR) is used to identify the outlier data and delete it. The preprocessed data for clustering is then clustered into different algorithms, and the first Farthest clustering algorithm achieves better results. Clustering output is given as input to machine learning classification algorithms, and the support vector machine is more accurate than other classifications. The presented model classifies patients into diabetic and non-diabetic patients with 99.4% accuracy.

5.4. **Improved hybrid model of Support Vector Machine, fuzzy C-Means and Principal Component Analysis (FPCA-SVM-FCM).** Dzulkalnine and Sallehuddin [17] used an improved hybrid model of support vector machine, fuzzy c-means, and principal component analysis on the dataset to diagnose diabetic patients. The proposed model consists of two parts: the fuzzy feature selection section and the missing values placement section. Missing data in the dataset is a major problem in terms of data analysis in various fields of research, especially in the field of medicine. This is because it affects the treatment and diagnosis that the patient should receive. In this research, the f-c-means method replaces the missing data. However, the fuzzy c-means does not take into account important features. Here, fuzzy properties are selected through a combination of fuzzy principal component analysis and Sequential Backward Search (SBS) using a support vector machine, while the missing values placement step is performed using fuzzy c-means. Fuzzy feature selection acts as a filtering method that ranks scores based on the principal component feature. The fuzzy principal component analysis is used to obtain principal component scores for each feature to determine the order in which they relate. The importance of each of these features is determined by the privileges of its respective principal component. The higher the principal component scores, the higher the relevance. In contrast, low scores of the main component indicate less relevance. The fuzzy principal component analysis is used for better accuracy and faster computation time. In the Sequential Backward Search (SBS) step using the support vector machine, the datasets processed and sorted by fuzzy principal component analysis are used as the main support vector input. This step aims to measure the performance of the selected features. In the last

step, the reduced dataset with the related properties obtained from the fuzzy principal component analysis and the support vector machine is given as input to the fuzzy c-means. The highest accuracy of the model is obtained 72.08% when 20% of the data is used as a test set.

5.5. **Deep neural network method using stacked automatic encoders.** Kannadasan et al. [30] Proposed a Deep Neural Network (DNN) method using automated stacked encoders to classify people with diabetes. In this method of stacked automatic encoders, the properties are extracted. Data classification is done using a Soft-max layer. The Smooth Maximum Classifier is a multi-class classifier that uses logistic regression and classifies data. The Soft-max layer uses a classification algorithm that uses extensive logistic regression to classify multiple classes. In this way, inspired by the features of deep networks, from a deep neural network-based framework using automated stacked encoders to classify preprocessing data, the number of pregnancies, which is a numerical feature, is converted to a nominal feature. Moreover, the missing values are replaced by the mean values of those features. The proposed algorithm model consists of three parts: dimensional reduction, elimination of out-of-noise samples, and classification. Dimension reduction in the preprocessed dataset is performed using principal component analysis. The selected principal component is then given to the K-means clustering algorithm to remove outlier and noise samples, and not properly clustered samples are removed from the dataset. Finally, the correctly clustered data is given as input to the logistic regression supervised classification algorithm. Experimental results show that principal component analysis, improved K-means clustering algorithm, and logistic regression classification accuracy are more accurate than other published studies, and logistic regression shows 97.4% accuracy.

5.6. **Hybrid method using improved K-means and logistic regression algorithms.** Alternatively, Wu et al. [46] Proposed a hybrid model for diagnosing type 2 diabetes using data mining and machine learning algorithms. The proposed model consists of two stages. In the first step, noise and outlier data are identified and deleted using the improved K-means algorithm. The correctly clustered data is given as input to the logistic regression algorithm for data classification in the second step. Pima Indian diabetes dataset and Weka tool have been used to implement the proposed model. Before the first stage, the dataset is preprocessed, and two critical pre-processions are performed on the data: First, the number of pregnancies, which is a numerical value, is converted to a nominal value. Second: the missing values of each feature are filled with its mean. The preprocessed data is fed to the improved K-means algorithm, and incorrectly clustered samples are removed from the dataset. The error rate is then calculated, and if it is higher than 75%, it is transferred to the data classification stage; otherwise, the variance is tested within other clusters.

Appropriately clustered data is classified as input to the logistic regression algorithm for the operation. The proposed model accurately diagnoses 95.41% diabetic patients.

5.7. **Predictive modeling and analysis for diabetes using a machine learning approach.** Using machine learning methods, Kaur and Kumari [31] provided predictive modeling and analysis for diabetes. In this method, diabetic patients were classified from the Pima Indian diabetes dataset using R software. The raw data processing step includes feature selection, deletion, and prediction of missing samples using the K-nearest-neighbor algorithm. The Boruta packing algorithm is used to select the essential feature. They used five algorithms to predict diabetes: Radial Basis Function (RBF), support vector machine, K-nearest neighbor, artificial neural network, and Multifactor Dimensionality Reduction (MDR). A linear model of the support vector machine provides the highest accuracy of 89% and 88% accuracy for predicting diabetes, higher than other models used. On the other hand, the K-nearest-neighbor algorithm obtained the recall rate of 90% and the F-score, 88%. The Area Under the Curve (AUC) of the linear model of the support vector machine and the K-nearest neighbor are 90% and 92%, respectively. Therefore, based on all the parameters, the linear model of the support vector machine and the K-nearest neighbor are the two better models for diagnosing diabetes.

5.8. **Partial Limited-memory Broyden-Fletcher-Goldfarb-Shanno (PL-BFGS).** Caliskan et al. [7] Examined the improvement in the performance of deep neural network classifiers with a simple training strategy. First, the dataset is pre-processed, and the internal parameter space of the deep neural network is divided into sections, and these sections are optimized separately. In this method, the Soft-max classification is used to separate the classes. Deep neural network classification is created by cascading a network of automatic stacked encoders with a Soft-max classification layer. Deep neural network classification training is performed using the L-BFGS optimization algorithm. All internal parameters of the network under training are classified as a single parameter vector and using the L-BFGS algorithm, the optimal value of this vector is searched. The simulation is repeated ten times for each L-BFGS strategy and algorithm with a 10-fold cross-validation method for the dataset. The deep neural network trained by L-BFGS has been selected as the source for all sizes of dataset segments with the best results. In order to analyze the performance of the proposed pL-BFGS method and compare it with the L-BFGS method, several simulations and their accuracy values are recorded. Based on the results, it has been observed that the pL-BFGS method can achieve superior performance with high accuracy in almost all cases; the model has achieved an accuracy of 77.09% on the Pima Indian diabetes dataset.

5.9. **Hybrid method of Stability Selection (SS) and Logistic Regression.** Akyol [1] diagnosed diabetic patients using a combination of Stability

Selection (SS) and logistic regression. In this method, first, the data is preprocessed, which consists of two stages. In the first stage, the data is normalized, and in the second stage, the features that reduce the accuracy of the model are removed through the stability selection method. Besides the characteristics of the number of pregnancies, plasma glucose concentration and body mass index are selected because of their great importance in diagnosing diabetes. The selected properties are given to the logistic regression algorithm and random forest. In this method, the aim was to determine the importance of selective characteristics and their function in the accuracy of predicting the diagnosis of diabetes. Stability selection is used to evaluate the importance of selected features. The highest accuracy obtained in this method is 78.57%.

5.10. **Hybrid method using genetic algorithm and multilayer perceptron.** Choubey and Paul [13] proposed a hybrid method for diagnosing diabetes using genetic algorithms and a multilayer perceptron neural network. The proposed model consists of two stages of feature selection and classification. First, the most important features are selected by a genetic algorithm, and in the next step, the selected features for classifying diabetic people from non-diabetic people are given to the multilayer perceptron neural network. The proposed model performs the classification operation using half of the features selected by the genetic algorithm. The multilayer perceptron is a class of Feedforward Neural Networks (FNN) used to classify data. The back Propagation (PB) algorithm is used to better and more accurately adjust the weight gradient used to teach the multilayer perceptron neural network and Feedforward Neural Network (FNN). The proposed model accurately detects 79.13% of diabetic patients.

5.11. **Medical decision support program using a weighted multi-layered group classifier framework.** Bashir et al. [2] proposed a medical decision support program using a new weighted group classification framework. The proposed model, called "HM-BagMoov," works using a set of seven heterogeneous classifications. This model is evaluated in five different datasets of heart diseases, four datasets of breast cancer, two datasets of diabetes, two datasets of liver diseases, and one set of hepatitis data prepared from public repositories. The proposed framework includes data acquisition, preprocessing, classifier training, and multi-level hierarchical bagging group classifiers with optimized multi-objective voting (HM-BagMoov) for disease prediction based on the three-layer approach. The preprocessing module includes feature selection, missing data placement, noise cancellation, and outlier data detection. The proposed framework uses the Criterion-F feature selection method to select the most important and relevant features from the medical dataset. Also, a mobile application called "IntelliHealth" has been prepared based on the presented model. The presented model has achieved an accuracy of 78.21% on the Pima Indian diabetes dataset.

5.12. **Intelligent hybrid system using genetic algorithm and decision tree.** Choubey and Paul [14] identified diabetic patients using genetic algorithms and decision trees in another work. First, the most important features are selected by the genetic algorithm. After selecting the feature by the genetic algorithm, four of the eight features are selected, which in the next step are given to the j48 decision tree classification algorithm to classify diabetic patients. Although the proposed model is faster, it does not have the accuracy and other evaluation criteria. The proposed model has 74.78% accuracy in diagnosing diabetes.

5.13. **Intelligent hybrid system consisting of Min-Max fuzzy neural network, classification and regression tree, and random forest algorithm.** Seera and Lim [38] present a hybrid intelligent system consisting of the Min-Max fuzzy neural network, the classification and regression tree, and the random forest algorithm, and its effectiveness as a decision support tool for classifying medical data has been investigated. The goal of a hybrid intelligent system is to take advantage of the constituent models while reducing their limitations. This model is able to gradually learn from data samples (due to Min-Max neural-fuzzy network), reveal predicted outputs (due to classification and regression tree), and achieve high classification performance (due to random forest). To evaluate the effectiveness of the hybrid intelligent system, three sets of medical data, namely, breast cancer, Pima Indian diabetes, and liver disorders from the UCI repository of machine learning, were used for evaluation. Several useful performance metrics in medical applications include accuracy, sensitivity, and specificity. Experimental results have shown that the hybrid intelligent system effectively performs the task of classifying medical data.

5.14. **Hybrid model using K-means algorithms and C4.5 decision tree.** Patil et al. [33] Proposed a hybrid model using K-means algorithms and the C4.5 decision tree. In this method, the dataset is preprocessed first. Data analysis revealed that zero was used instead of missing values in the data preprocessing stage. Whereas, it is not logical that the value of a variable such as plasma-glucose concentration in living individuals is zero; all samples with a zero plasma-glucose concentration are excluded from the dataset. Also, in this analysis, the number of missing values for the 2-hour serum insulin properties and the thickness of the triceps skinfold is very high (374 and 227, respectively, out of a total of 768 samples), so both of these characteristics were omitted. Similarly, standard score normalization-Z1 is used to normalize the data in this method. For data preprocessing, 625 samples of data remain, which are given as input to the K-means clustering algorithm. Data that is identified as an outlier or noise data is removed from the dataset. After extracting the template, 192 samples were identified as outlier data and removed from the dataset. The remaining data for classification is given to the decision tree algorithm C4.5, and the model is made. Performance evaluation in this method is measured

using accuracy, specificity, and sensitivity. Using 10-fold mutual validation, accuracy and precision were obtained as 84.24% and 92.38%, respectively.

6. Diabetes diagnosis methods comparison

In this section, a summary of the proposed models is given in Tables 3 and 4. The results show that most researchers have chosen Weka tools to predict disease. Moreover, all the methods presented have used the Indian Survey Database to diagnose diabetes. The results show that Devi et al.'s model is more accurate for diagnosing diabetes in terms of evaluation criteria than other proposed methods. However, in this method, selecting the essential features is not used. To demonstrate further efficiency, Wu et al. [46], Bashir et al. [2], Seera and Lim [38] and Caliskan et al. [7] have tested their model with other data sets and the existing models have achieved high accuracy on these data sets. Also, Bashir et al. [2] have designed a module for disease prediction that will be available to doctors. Most of the models have used the methods of selecting or extracting the most important features to obtain the most important features of Pima Indian dataset. One of the disadvantages of the models is the lack of evaluation of the model with the evaluation criteria that the two models of Dzulkalnine and Sallehuddin [17] and Caliskan et al. [7] have not evaluated their model with any evaluation criteria. Also, the model of Dzulkalnine and Sallehuddin [17], in addition to using fuzzy methods, and although it has high complexity, it has obtained low accuracy.

Figure 5 shows the accuracy of the investigated methods on Pima Indian dataset. As shown in the figure, Devi et al. method [16] has higher accuracy.

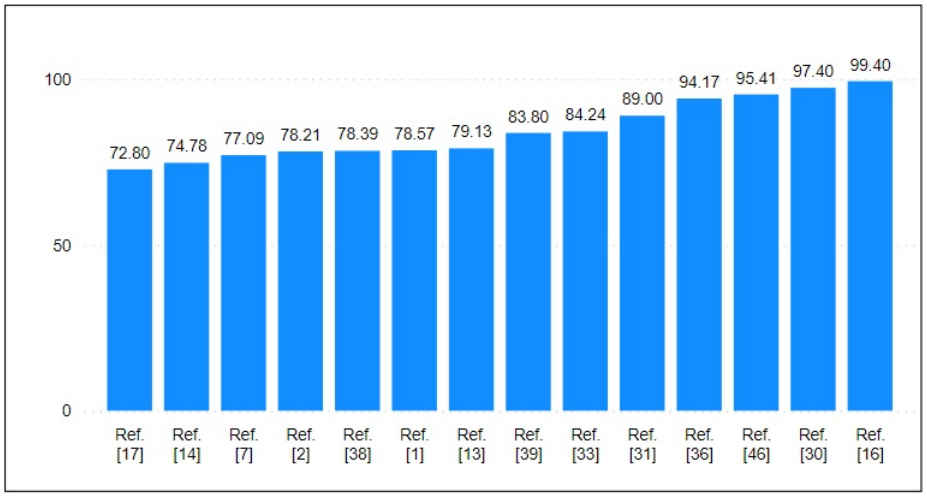

FIGURE 5. Accuracy comparison of reviewed methods.

TABLE 3. Detailed description of examples in the Indian Diabetes Database

Methods \ Criteria	Used	Used tool	Accuracy	F-score	Recall	Precision	Other evaluated criteria
Rahim et al. [36]	KNN-SVM-RF-NB-LR-SEM	-	✓	✓	✓	✓	-
Singh and Singh [39]	LSVM-RBF-SVM-Poly-SVM- DT	MATLAB	✓	✓	✓	-	Specificity-ROC
Devi et al. [16]	First Farthest Clustering-SVM	Weka	✓	✓	✓	✓	ROC-Time
Dzulkalnine and Sallehuddin [17]	Fuzzy SVM - PCA	-	✓	-	-	-	RMSE
Kannadasan et al. [30]	Deep neural network - Stacked automatic encoder - Logistic regression	MATLAB	✓	✓	✓	✓	Specificity
Wu et al. [46]	k-means Improved and logistic regression	Weka	✓	✓	✓	✓	ROC, MCC
Kaur and Kumari [31]	RBF, MDR, SVM, KNN, ANN	R-software	✓	✓	✓	✓	AUC
Caliskan et al. [7]	DNN-LBFGS-BFGS- pL-BFGS	-	✓	-	-	-	-
Akyol [1]	Random forest-logistic regression	-	✓	-	✓	-	ROC-Specificity
Choubey and Paul [13]	Genetic algorithm and multi-layer perceptron	Weka	✓	✓	✓	✓	ROC-MAE

TABLE 4. Detailed description of examples in the Indian Diabetes Database

Bashir et al. [2]	multi-level hierarchical bagging group classifiers with optimized multi-objective voting	-	✓	✓	✓	-	Specificity
Choubey and Paul [14]	Genetic Algorithm and Decision Tree J48	Weka	✓	✓	✓	✓	ROC-MAE
Seera and Lim [38]	Min-Max fuzzy neural network, classification and regression tree, random forest	-	✓	-	✓	-	ROC-Specificity
Patil et al. [33]	k-means-Decision tree C4.5	Weka	✓	-	✓	-	Kappa Statistic-Specificity

7. Discussion, challenges and future direction

Currently, data mining and machine learning methods are employed in different fields such as face and speech recognition, autonomous vehicles, computer games, classification and segmentation of medical images, processing of clinical reports, and medical diagnosis. The main challenge in the application of the data mining and machine learning methods, especially in the medical fields, is to provide sufficient and appropriate data for evaluation. On the other hand, machine learning applications tend to be mistrusted because of their inability to show the internal decision-making process.

One advantage of providing treatments to patients in the early step of their diseases is that they can avoid costly treatments later in life as the disease gets worse day to day. This is made more problematic with a lack of medical physicians in underserved regions.

Diabetes mellitus (type 2 diabetes), or simply diabetes, is a leading non-communicable disease globally. The main issues concerning the diagnosing diabetes are the large number of patients, shortage of healthcare in some countries, and the relatively the diagnostic tests may require. Therefore, considering different datasets for diagnosing diabetes and the abilities of the machine learning methods such as high processing speed, the diabetes can be diagnosed more

quickly and effectively by employing the machine learning techniques. Therefore, as we mentioned before, although there are now larger, more complex diabetes datasets, the Pima Indian Diabetes dataset has remained a benchmark for diabetes classification research. As common problems become clearer, the novel techniques in data science can bring advantages to other fields of science, such as medicine. However, several researches have used different machine learning techniques for diabetes diagnosis on Pima Indian Diabetes dataset such as KNN, SVM, RBF, ANN and DT; And in this paper, some of them were reviewed and compared by various criteria especially accuracy.

Our future work will include developing innovative methods and applying them to other types of medical analysis. So that, one of our objectives is improving the accuracy by using suitable pre-processing techniques for data management and analysis. Also, the combination of the Internet of Medical Things (IoMT) and data mining techniques and machine learning methods can be made available to assistance physicians in the early diagnosis of disease and providing predictive tools for more efficient and timely decision-making.

8. Conclusion

There are different datasets for diagnosing diabetes. In this study, we tried to use articles that used the Pima Indian dataset in their method. Various works have been done in the field of Pima Indian dataset survey. Many researchers have considered the methods of selecting and extracting features to select or extract the essential features. The genetic algorithm for selecting the most critical features has received much attention. Moreover, using combined methods, many researchers have improved the accuracy of diagnosing this disease, in which the support vector machine (SVM) algorithm has a much better performance than other machine learning algorithms due to the nature of the data. Each researcher has also considered different methods for preprocessing data. Most researchers have chosen the Weka data mining tool for their experiments.

References

[1] Akyol, K. (2017). Assessing the importance of attributes for diagnosis of diabetes disease. International Journal of Information Engineering and Electronic Business, 9(5), 1-9.

[2] Bashir, S., Qamar, U., & Khan, F. H. (2016). IntelliHealth: a medical decision support application using a novel weighted multi-layer classifier ensemble framework. Journal of Biomedical Informatics, 59, 185-200. https://doi.org/10.1016/j.jbi.2015.12.001.

[3] Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: current issues and guidelines. International Journal of medical informatics, 77(2), 81-97.

[4] BERGER, A. M., & BERGER, C. R. (2004). Data mining as a tool for research and knowledge development in nursing. CIN: Computers, Informatics, Nursing, vol. 22, no. 3, pp. 123-131.

[5] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324.

[6] Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition, Data mining and knowledge discovery, 2(2), 121-167. https://doi.org/10.1023/A:1009715923555.
[7] Caliskan, A., Yuksel, M. E., Badem, H., & Basturk, A. (2018). Performance improvement of deep neural network classifiers by a simple training strategy. Engineering Applications of Artificial Intelligence, 67, 14-23. https://doi.org/10.1016/j.engappai.2017.09.002.
[8] Chakrabarti, S., Ester, M., Fayyad, U., Gehrke, J., Han, J., Morishita, S., Shapiro, G.P., & Wang, W. (2006). Data mining curriculum: A proposal (Version 1.0). Intensive Working Group of ACM SIGKDD Curriculum Committee, 140, 1-10.
[9] Chang, V., Bailey, J., Xu, Q.A., & Sun, Z. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Computing and Applications. https://doi.org/10.1007/s00521-022-07049-z.
[10] Chapelle, O., Haffner, P., & Vapnik, V. N. (1999). Support vector machines for histogram-based image classification. IEEE transactions on Neural Networks, 10(5), 1055-1064. https://doi.org/10.1109/72.788646.
[11] Chikh, M.A., Saidi, M., & Settouti, N.(2012). Diagnosis of diabetes diseases using an Artificial Immune Recognition System2 (AIRS2) with fuzzy K-nearest neighbor. Journal of Medical Systems, 36(5), 2721-2729. https://doi.org/10.1007/s10916-011-9748-4.
[12] Cho, S., May, G., Tourkogiorgis, I., Perez, R., Lazaro, O., de La Maza, B., & Kiritsis, D. (2018). A hybrid machine learning approach for predictive maintenance in smart factories of the future. proceedings of the IFIP International Conference on Advances in Production Management Systems, Springer, 311-317. https://doi.org/10.1007/978-3-319-99707-0_39. https://doi.org/10.1016/j.ijmedinf.2006.11.006.
[13] Choubey, D. K., & Paul, S. (2016). GA_MLP NN: A hybrid intelligent system for diabetes disease diagnosis. International Journal of Intelligent Systems and Applications, 8(1), 49.
[14] Choubey, D. K., & Paul, S. (2015). $GA_J48graft$ DT: A hybrid intelligent system for diabetes disease diagnosis. International Journal of Bio-Science and Bio-Technology, 7(5), 135-150.
[15] Cios, K. J., Moore, G. W. (2002). Uniqueness of medical data mining. Artificial Intelligence in Medicine, 26(1-2), 1-24. . https://doi.org/10.1016/S0933-3657(02)00049-0.
[16] Devi, R.D.H., Bai, A., & Nagarajan, N. (2020). A novel hybrid approach for diagnosing diabetes mellitus using farthest first and support vector machine algorithms. Obesity Medicine, 17, 100152. https://doi.org/10.1016/j.obmed.2019.100152.
[17] Dzulkalnine, M. F., & Sallehuddin, R. (2019). Missing data imputation with fuzzy feature selection for diabetes dataset. SN Applied Sciences, 1(4), 362. https://doi.org/10.1007/s42452-019-0383-x.
[18] Elavarasan, D., Vincent, D. R., Sharma, V., Zomaya, A. Y., & Srinivasan, K. (2018). Forecasting yield by integrating agrarian factors and machine learning models: A survey, Computers and Electronics in Agriculture, 155, 257-282. . https://doi.org/10.1016/j.compag.2018.10.024.
[19] Esposito, F., Malerba, D., Semeraro, G., & Kay, J. (1997). A comparative analysis of methods for pruning decision trees. IEEE transactions on pattern analysis and machine intelligence, 19(5), 476-491. https://doi.org/10.1109/34.589207.
[20] El-Habil, A. M. (2012). An application on multinomial logistic regression model. Pakistan journal of statistics and operation research, 271-291. https://doi.org/10.18187/pjsor.v8i2.234
[21] Federation, I. D. (2019).IDF Diabetes Atlas 2019. ed: International Diabetes Federation.
[22] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

[23] Guo, G. ,Wang, H. , Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. proceedings of the OTM Confederated International Conferences On the Move to Meaningful Internet Systems, Springer, 986-996. . https : //doi.org/10.1007/978 − 3 − 540 − 39964 − 3 − 62.
[24] Hartigan, J.A., & Wong, M. A. (1979). Algorithm AS 136: a k-means clustering algorithm, J. R. Stat. Soc. Ser. C Appl. Stat. 28 (1), 100–108.
[25] https://diabetesatlas.org/en/sections/demographic-and-geographic-outline.html (Accessed 4 May 2019).
[26] https://idf.org/ (Accessed 8 August 2020).
[27] https://archive.ics.uci.edu/ml/index.php (Accessed 13 August 2020).
[28] Jackins, V., Vimal, S., Kaliappan, M., & Lee, M. Y. (2021). AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. The Journal of Supercomputing, 77(5), 5198-5219. https://doi.org/10.1007/s11227-020-03481-x.
[29] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features, in European conference on machine learning, Springer, 137-142. https://doi.org/10.1007/BFb0026683.
[30] Kannadasan, K., Edla, D. R., & Kuppili, V. (2019). Type 2 diabetes data classification using stacked autoencoders in deep neural networks. Clinical Epidemiology and Global Health, 7(4), 530-535.
[31] Kaur, H., & Kumari, V. (2020). Predictive modelling and analytics for diabetes using a machine learning approach," Applied Computing and Informatics. https://doi.org/10.1016/j.aci.2018.12.004.
[32] Mitchell, M. (1996). Chapter 3: Genetic Algorithms in Scientific Models, An Introduction to Genetic Algorithms. The MIT Press, Cambridge, MA, 85-108.
[33] Patil, B. M., Joshi, R. C., & Toshniwal, D. (2010). Hybrid prediction model for type-2 diabetic patients. Expert Systems with Applications, 37(12), 8102-8108. https://doi.org/10.1016/j.eswa.2010.05.078.
[34] Petricoin, E.F., Ardakani, A.M., Hitt, B.A., Levine, P.L., Fusarob, V.A., Steinberg, S.M., Mills, G.B., Simone, C., Fishmen, D.A., Kohn, E.C., & liotta, L.A. (2002). Use of proteomic patterns in serum to identify ovarian cancer. The lancet, 359(9306), 572-577. https://doi.org/10.1016/S0140-6736(02)07746-2.
[35] Piryonesi, S. M., & El-Diraby, T.E. (2020). Data analytics in asset management: Cost-effective prediction of the pavement condition index. Journal of Infrastructure Systems, 26(1), 04019036. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000512.
[36] Rahim, M.A., Hossain, M.A., Hossain, M.N., Shin, J., & Yun, K.S. (2023). Ensemble-Based Type-2 Diabetes Prediction Using Machine Learning Techniques. Annals of Emerging Technologies in Computing, 7(1), 30–39.
[37] Richards, G., Rayward-Smith, V. J, Sönksen, P. , Carey, S., & Weng, C. (2001). Data mining for indicators of early mortality in a database of clinical records. Artificial Intelligence in Medicine, 22(3), 215-231. https://doi.org/10.1016/S0933-3657(00)00110-X.
[38] Seera, M., & Lim, C. P. (2014). A hybrid intelligent system for medical data classification. Expert Systems with Applications, 41(5), 2239-2249. https://doi.org/10.1016/j.eswa.2013.09.022.
[39] Singh, N., & Singh, P. (2020). Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus, Biocybernetics and Biomedical Engineering, 40(1), 1-22. https://doi.org/10.1016/j.bbe.2019.10.001.
[40] Song, M. (2008). Biomedical ontologies and text mining for biomedicine and healthcare: A survey. Journal of Computing Science Engineering, 2(2), 109-136.
[41] Steinley, D., & Brusco, M. J. (2007). Initializing k-means batch clustering: A critical evaluation of several techniques. Journal of Classification, 24(1), 99-121. https://doi.org/10.1007/s00357-007-0003-0.

[42] Vadeyar, D.A., & Yogish, H. (2014). Farthest first clustering in links reorganization. International Journal of Web and Semantic Technology, 5(3), 17.
[43] Velickov, S., & Solomatine, D. (2000). Predictive data mining: practical examples. proceedings of the 2nd Joint Workshop on Applied AI in Civil Engineering.
[44] Velu, C., & Kashwan, K. (2013). Visual data mining techniques for classification of diabetic patients. proceedings of the 3rd IEEE International Advance Computing Conference (IACC), IEEE, 1070-1075. https://doi.org/10.1109/IAdCC.2013.6514375.
[45] Wu, L., Peng, Y., Fan, J., Wang, Y., & Huang, G. (2021). A novel kernel extreme learning machine model coupled with K-means clustering and firefly algorithm for estimating monthly reference evapotranspiration in parallel computation. Agricultural Water Management, 245, 106624. https://doi.org/10.1016/j.agwat.2020.106624.
[46] Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked, 10, 100-107. https://doi.org/10.1016/j.imu.2017.12.006.
[47] Yoo, I., Alafaireet, P., Marinov, M., Hernandez, K. P., Gopidi, R., Chang, J., & Hua, L. (2012). Data mining in healthcare and biomedicine: a survey of the literature, Journal of medical systems, 36(4), 2431-2448. https://doi.org/10.1007/s10916-011-9710-5.
[48] Zhan, M., Chen, Z. B., Ding, C. C., Qu, Q., Wang, G. Q., Liu, S., & Wen, F. Q. (2021). Machine learning to predict high-dose methotrexate-related neutropenia and fever in children with B-cell acute lymphoblastic leukemia. Leukemia and Lymphoma, 1-12. https://doi.org/10.1080/10428194.2021.1913140.

Farzad Heydari
Orcid number: 0000-0001-8062-4507
Department of Computer Science
Faculty of Mathematics and Computer
Shahid Bahonar University of Kerman
Kerman, Iran
Email address: farzadheydari9390@gmail.com

Marjan Kuchaki Rafsanjani
Orcid number: 0000-0002-3220-4839
Department of Computer Science
Faculty of Mathematics and Computer
Shahid Bahonar University of Kerman
Kerman, Iran
Email address: kuchaki@uk.ac.ir

Masoumeh Sheikh Hosseini Lori
Orcid number: 0009-0009-1579-1734
School of Health
Isfahan University of Medical Sciences
Isfahan, Iran
Email address: masoomeh_shhosseini72@yahoo.com