SN Computer Science (2024) 5:217
https://doi.org/10.1007/s42979-023-02545-y

ORIGINAL RESEARCH

Early Prediction of Diabetes Using Feature Selection and Machine Learning Algorithms

Jafar Abdollahi[1,2] · **Solmaz Aref**[3]

Received: 13 February 2022 / Accepted: 6 December 2023

Abstract

Diabetes has become one of the most common diseases in middle- and low-income countries. Machine learning (ML) and data mining techniques have recently been used to predict diabetes with a high success rate. As a result, medical professionals seek a dependable method for predicting diagnosis. Of course, the feature selection process may be considered a global combinatorial optimization problem in machine learning. The number of features is reduced, irrelevant, noisy, redundant data are removed, and classification accuracy is acceptable. This work uses particle swarm optimization (PSO) to implement feature selection, followed by performance comparison. After that, three medical datasets are used to compare the performance of several machine learning methods. Standard approaches are used to determine the optimum technique for the three datasets. The best results for three datasets are reported for each scheme. The primary goal is to assess the validity of each algorithm's data classification in terms of efficiency and effectiveness in terms of accuracy, sensitivity, and specificity. Decision Tree, Random Forest, and Naïve Bayes deliver the highest accuracy with the lowest mistake rate, according to the findings of the experiments. Machine learning may classify and determine which instances should be sent to medical for further evaluation and treatment with high accuracy. Using such an algorithm on a global scale could help minimize the number of people diagnosed with diabetes.

Keywords Diabetes · Machine learning · Feature selection · PSO

Introduction

Diabetes is a prominent cause of death across the world [1]. Diabetes can harm one's health if discovered too late [2]. Individuals/families, healthcare institutions, and society bear tremendous financial costs [3]. Furthermore, nearly 30 million Indians have diabetes, with many more at risk [4]. Most people get chronic illnesses due to their lifestyle, eating choices, and lack of physical exercise [5]. Predicting future health outcomes is extremely desired, especially for pre-diabetic patients implementing preventative and intervention measures [6]. Diabetes remission is a hotly disputed concept in contemporary endocrinology [7].

Medical practitioners are looking for an effective diabetes prediction system. Different machine learning approaches can examine data from various angles and synthesize it into meaningful information. If specific data mining techniques are applied to large volumes of data, they will be able to provide us with relevant knowledge [8].

Data mining techniques aid in the machine learning process and are widely used in various critical applications [9]. Many data processing methodologies, decision support systems, and systems that probe deeper into the diseases were discovered in the current literature [10–17]. Several machine learning approaches are used in clinical settings to forecast illness, and they have been demonstrated to be more accurate than the traditional methods for diagnosis [18]. As a result, modern medicine has encountered issues acquiring vast amounts of data, analyzing it, and applying the resulting

This article is part of the topical collection "Pattern Recognition and Machine Learning" guest edited by Ashish Ghosh, Monidipa Das and Anwesha Law.

✉ Jafar Abdollahi
ja.abdollahi77@gmail.com

[1] Department of Computer Engineering, Ardabil Branch, Islamic Azad University, Ardabil, Iran

[2] Young Researchers and Elite Club, Ardabil Branch, Islamic Azad University, Ardabil, Iran

[3] Department of Molecular Biology, Mysore University, Mysore, Karnataka, India

Published online: 20 January 2024

knowledge to solving complicated clinical problems; AI capabilities are required for these goals [19].

Given the importance of diabetes care and the assumption that AI applications for diabetes care are useful tools, and the scarcity of studies examining the use of AI for diabetes care, this study examined AI algorithms and techniques for diabetes care, focusing on machine learning methods. Diabetes outcomes are classified and diagnosed by employing a type of algorithm. This work compares the performance of nine classifiers following Feature Selection using Particle Swarm Optimization (PSO). The most prominent data mining algorithms in the top 10 data mining algorithms research community are LR, NB, C 4.5, DT, RF, SVM, GB, SGDA, and KNN. Our goal is to evaluate the efficiency and effectiveness of these algorithms in terms of accuracy, sensitivity, specificity, and precision.

A significant amount of vital and sensitive healthcare data have been produced due to the tremendous breakthroughs in biotechnology and public healthcare infrastructures. Many intriguing patterns are discovered through intelligent data analysis tools for the early and onset diagnosis and prevention of various fatal diseases. An early diabetes diagnosis can result in more effective therapy. Data mining techniques are widely used for the prediction of disease at an early stage. In this study, diabetes is predicted using significant attributes, the relationship between the various features is also characterized, and a comparison of the proposed approach with the current state-of-the-art techniques is also carried out, demonstrating the proposed method's adaptability in many public applications in healthcare. Moreover, the main contribution of this article is as follows:

- Diabetes prediction models using machine learning performed well.
- A comparison of the findings from the suggested technique with the most pertinent studies carried out following the prior literature.
- We investigate the benefits of feature selection (PSO-ML) for prediction and feature selection.

The article's structure is as follows: In "Related works" Section summarizes related work, "Materials and methods" Section proposes a method, and "Results" Section gives experimental data, including performance evaluation and comparison. The article's "Conclusion and Future Work" Section are presented in the final section.

Related Work

Feature selection (FS) is indeed a tough, challenging, and demanding task due to the large exploration space. It moderates and lessens the number of features. It also eliminates insignificant, noisy, superfluous, repetitive, and duplicate data, and provides reasonably adequate classification accuracy. Present feature selection approaches do face the difficulties like stagnation in local optima, delayed convergence and high computational cost. In machine learning, particle swarm optimization (PSO) is an evolutionary computation procedure which is computationally less costly and can converge quicker than other existing approaches. PSO can be effectively used in various areas, like medical data processing, machine learning and pattern matching, but its potential for feature selection is yet to be fully explored. PSO improves and optimizes a candidate solution iteratively with respect to a certain degree of quality. It provides a solution to the problem by having an inhabitant of swarm particles. By applying mathematical formulas, velocity and position of swarm particles are calculated and these particles are moved in the search space. The movement of individual swarm particle is inclined by its local finest known position and is also directed to the global finest known position in the exploration space. These positions are updated as improved positions, which are found by other particles. These improved positions are then used to move the swarm in the direction of the best solutions. The aim of the study is to inspect and improve the competence of PSO for feature selection. PSO functionalities are used to detect a subset of features to accomplish improved classification performance than using entire features set [20].

In [21] several algorithms are examined on the PIMA Indian dataset and a localized dataset. Principle component analysis (PCA) and PSO are also used in different combinations with classification algorithms. The best results of 79.56% by PCA-LR and 92.43% by PSO-Naive Bayes were achieved on the PIMA Indian and localized datasets. The PSO is also employed by [5], to improve ANN accuracy for diabetes detection. They successfully tried to control the saturation rate of PSO activation function.

Hassan et al. [22] examined a self-organizing map (SOM) optimization algorithm with four metaheuristic algorithms, including PSO, newton-based SOMPSO, SOMHSA (SOM with the Harmony search algorithm), and SOMSwram. The best accuracy of diagnosis of diabetic patients of 80% is achieved on the PIMA Indian diabetes dataset. The four algorithms are also examined on Wisconsin and new Thyroid dataset, and better accuracies than those on the PIMA Indian dataset were obtained. For example, for the new Thyroid dataset, accuracy of 91% through newton-based SOM, and Wisconsin dataset, accuracy of 97% was gained through SOMHSA.

Machine learning methods are now utilized to analyze high-dimensional biomedical data automatically. Some examples of biomedical applications of ML include liver disease diagnosis, skin lesions, cancer categorization, risk

assessment for cardiovascular disease, and analysis of genetic and genomic data [19].

Type 1 and type 2 diabetes exacerbates the negative effects of COVID -19 independently [23]. In [24], the proportional contributions of insulin resistance and beta-cell dysfunction in type 2 diabetes are varied and dependent on demographic, genetic, and clinical factors, with significant interaction with environmental factors [25]. In the case of newly diagnosed DM2, the VERIFY research found that early treatment with metformin–vildagliptin improves long-term glycemic control and can slow disease progression [26]. People with type 2 diabetes diagnosed in adolescence and early adulthood (or with a younger present age) were intrinsically and more prone to retinopathy after accounting for illness duration and other key confounding factors [27]. Simple non-invasive fibrosis scores based on normal blood tests are increasingly examined as screening tools [28].

Miroslav Marinov et al. [29] reviewed 31 articles related to a diabetes diagnosis. This study was classified under the classification, clustering, and association data mining methods. The authors stated that data mining has a bright future in biomedicine. However, there was no detailed classification accuracy comparison.

Anjali Khandgar presented a review to interpret various data mining techniques for diabetes prediction. This study has shown standards for analyzing the parameters of behavior and lifestyle of patients such as emotions, physical activities, eating habits, etc. The retrieved information can be used to check clinical parameters, other prognoses, and treatment planning. However, a comparison of the accuracy of different methods is not mentioned [30].

Preeti Verma et al. [31] reviewed various studies with classification techniques for a diabetes diagnosis. The results showed that the support vector machine (SVM) effectively classifies the diabetic disorder. The accuracy rate obtained using SVM is 96.58%. The authors have not investigated the effects of data preprocessing on the accuracy of the prediction of diabetic patients.

Yu et al. [32] used quantum particle swarm optimization (QPSO) and weighted least squares support vector machine (WLS-SVM) for type 2 diabetes prognosis. Fanicol et al. conducted their study on the same data set and used four algorithms NB, DT, LR, and 274RF. They calculated the performance of each classifier and found that the most successful method was RF with tenfold cross-validation with an accuracy of 97.4% [33, 34]. Zhu et al. [35] reduced the data size by principal component analysis (PCA) in feature extraction methods using random data from 68,994 patients obtained from a hospital in Luzhou, China. Using the obtained features, they achieved an accuracy of 80.84% with RF. In the following, comparing the related work with existing work and its limitations (Table 1).

Table 1 Comparing the limitations of related work with Existing work

Refs.	Authors	Limitation
[23]	Gregory et al.	This study has three important limitations that should be considered First, information was gathered from one academic health system that mostly served urban and suburban populations Second, even though COVID-19 testing was administered to all hospitalized and presurgical patients at VUMC during the prospective study period, our study cannot rule out the possibility that clinicians in the outpatient setting were more likely to test patients with diabetes than those without diabetes because of the belief that those with diabetes were at higher risk Third, even though we thoroughly described our study's risk factors and outcomes for COVID-19 patients, the sample size is still somewhat small compared to studies of the entire community. We could not perform some multivariate regression analyses within the type 1 diabetes group due to the smaller sample size, which made it difficult to characterize COVID-19 outcomes in type 1 diabetic Hispanic patients
[24]	Graham EA et al.	Results showed heterogeneity and evidence of publication bias
[26]	Middleton TL et al.	A cohort that was referred to a hospital outpatient clinic is the subject of the data analysis. Because of this, referral bias may limit generalizability, and causality cannot be established from a cross-sectional study like any other type of study. There may be a survival bias and conflicting risks of death when there are diabetic complications, among other drawbacks
[27]	Alkayyali T et al.	Our results need to be understood in light of several constraints. First, additional compound surrogates, such as APRI and the AST/ALT ratio, were not considered in our investigation, which was entirely focused on the FIB-4 and NFS scores. Although transient elastography—a frequently used, accurate non-invasive diagnostic tool—was not used, all patients had liver ultrasonography. Furthermore, we did not examine how anti-diabetic medications might have impacted our results. Last but not least, because our results—particularly about the idea FIB-4 and NFS values—were obtained in a Turkish population, they might not be generalizable. Additional clinical research is required to determine whether such cutoffs might be population specific
Existing work	Jafar Abdollahi et al.	Information that is either difficult to find or unreliable. Because of this, referral bias may limit generalizability, and causality cannot be established from a cross-sectional study like any other type of study

Particle swarm optimization (PSO) is used to implement feature selection in this work, followed by a performance comparison of machine learning algorithms on three medical datasets. The project is divided into two halves. The first is the feature selection approach, which encourages more relevant traits while discarding the irrelevant for faster and more efficient data classification. The classification algorithms are applied to the obtained features in the second stage to predict.

Machine Learning Algorithms

Machine learning (ML), a subset of artificial intelligence (AI), has expanded significantly in data analysis and computing in recent years, enabling programs to perform intelligently. ML is typically referred to as the most well-liked newest technology in the fourth industrial revolution and gives systems the ability to learn and improve from experience automatically without being specifically programmed (4IR or Industry 4.0). Utilizing cutting-edge smart technologies like machine learning automation, "Industry 4.0" is often the ongoing automation of traditional manufacturing and industrial activities, including exploratory data processing. Thus, to intelligently analyze these data and construct the related real-world applications, machine learning algorithms are the key [36].

In the following sections, we will provide a brief overview of several machine learning algorithms that are the most often utilized and, consequently, the most well-liked ones. Additionally, it aims to emphasize the advantages and disadvantages of machine learning algorithms from the perspective of their applications to help decision-makers make an informed choice when choosing the best algorithm to fulfill a certain application requirement. Table 2 compares the benefits and drawbacks of the algorithm for diagnosing diabetes to previous methods (Table 3).

PSO Algorithms

Many challenging research issues can be formulated as optimization issues. The emergence of big data technology has also sparked a large-scale increase in the complexity and size of optimization challenges. The development of parallelized optimization techniques has become necessary due to the high computing cost of these issues. One of the most well-known swarm intelligence-based algorithms, particle swarm optimization (PSO), is enhanced with resilience, simplicity, and global search capabilities [37]. It has undergone numerous improvements since it was first introduced in 1995. With more knowledge of the method, researchers have created new iterations that address diverse demands, created new applications in various fields, published theoretical analyses of the consequences of the different parameters, and proposed numerous algorithm variations [38]. Ant colony optimization (ACO), particle swarm optimization (PSO), artificial fish swarm (AFS), bacterial foraging optimization (BFO), and artificial bee colony are just a few of the swarm intelligence techniques that have been developed in recent years (ABC). This paper attempts to pick features using PSO. Table 4 compares the effectiveness of feature selection methods based on PSO.

In the previous article [1], we used the genetic algorithm to predict diabetes, and in the comparison, we made with the particle swarm algorithm, we saw that this algorithm has the following advantages over the genetic algorithm and can be successful in predicting diabetes. Therefore, we tried to use this algorithm to predict diabetes. Also, the PSO does not rely on the gradient of the objective function, it is computationally more efficient than Genetic Algorithm. Moreover, it is simple to parallelize. Each particle may be changed concurrently, and since we are manipulating numerous particles to find the best answer, we only need to gather the updated value once per iteration. As a result, PSO may be implemented well using map-reduce architecture. In this article, feature selection using this algorithm is proposed. The results obtained using this method are compared to those obtained using a number of traditional machines learning algorithms, including Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbor (KNN), Naive Bayesian Classifier (NBC), Random Forest Classifier (RFC), and Logistic Regression (LR). The computational findings of our suggested strategy demonstrate that improved prediction accuracy can be attained with significantly fewer features. This work has the potential to be useful in clinical settings and serve as a resource for clinicians.

Difference Between PSO and Genetic Algorithm

Genetic Algorithms (GAs) and PSOs are both used as cost functions, they are both iterative, and they both have a random element. They can be used in similar kinds of problems. The difference between PSO and Genetic Algorithms (GAs) is that GAs does not traverse the search space like birds flocking, covering the spaces in between. The operation of GAs is more like Monte Carlo, where the candidate solutions are randomized, and the best solutions are picked to compete with a new set of randomized solutions. Also, PSO algorithms require normalization of the input vectors to reach faster “convergence” (as heuristic algorithms, both do not truly converge). GAs can work with features that are continuous or discrete. Also, In PSO, there is no creation or deletion of individuals. Individuals merely move on a landscape where their fitness is measured over time. This is like a flock of birds or other creatures that communicate.

Table 2 Advantages and disadvantages of the proposed method in diagnosing diabetes compared to other methods

Row	Algorithm	Advantages	Disadvantages
1	LR	It is now easier to use, analyze, and train logistic regression	If the data is less than the number of features, logistic regression should not be used; otherwise, overfitting may occur
2	NB	The Naive Bayes algorithm performs exceptionally well with categorical input variables compared to numerical data	This method is also notorious for being a poor estimate. As a result, you should not take the "predict probe" probability outputs too seriously
3	KNN	Because the data are a model that will be used as a reference for future predictions, KNN modeling does not require a training period	The high dimensionality complicates the distance calculation technique, making it difficult to determine the distance for each dimension
4	DT	Because it follows a regular protocol for making any call-in real life, it is straightforward to comprehend	For several category labels, the decision tree's process quality might be improved
5	RF	As the number of braids increases, so does the development precision	Predicting many trees takes a lengthy time
6	GB	In comparison to other modes, it is generally more accurate Faster training, particularly on larger datasets; most of them allow categorical features, and some handle missing values natively	Overfitting is a problem that can be remedied by using L1 and L2 regularization penalties. You can also attempt a low learning rate; Models, especially on CPUs, can be computationally expensive and require a long time to train The final models are difficult to interpret
7	SGDA	Because the network processes only one training sample, it is easier to fit into memory Because only one sample is processed at a time, it is computationally efficient It can converge faster for larger datasets because the parameters are updated more frequently The steps toward the minima of the loss function include oscillations that can help get out of the local minimums of the loss function due to frequent updates (in case the computed position turns out to be the local minimum)	The steps taken toward the minima are highly loud due to frequent updates. This can cause the gradient to slant in unexpected directions Furthermore, due to noisy steps, convergence to the loss function minima may take longer Because all resources are used to process one training sample at a time, frequent updates are computationally expensive Because it only interacts with one sample at a time, it lacks the benefit of vectorized operations
8	C4.5	C4.5 is a multi-branch tree with a slow calculation speed, whereas CART is a binary tree with a fast calculation speed C4.5 can only be classified, whereas CART can be classified or regressed CART uses the Gini coefficient as the impurity of the variable, reducing a lot of logarithmic operations; CART uses proxy testing to estimate missing values, whereas C4.5 is divided into different nodes with different probabilities; CART uses the "cost	C4.5 employs a poultry instead of a binary tree, which is more efficient; C4.5 can only be used for classification The entropy model in C4.5 includes many time-consuming logarithmic operations, continuous values, and sorting operations
9	SVM	The possibilities are endless The maximum level of dimensional efficiency	There is a loss of efficiency when there are more than a few samples
10	Proposed Model	Achieve a high level of classification reliability When compared to machine learning models, accuracy and error have improved A comparison of the proposed method's outcomes with another model	Data that are not readily available or are not trustworthy. Consequently, referral bias could potentially reduce generalizability, and, as with any cross-sectional study, causality cannot be inferred

Table 3 Comparison of the performance of other feature selection

Filter methods	Wrapper methods	Embedded methods
Information gain	Recursive feature elimination:	L1 regularization (LASSO)
Chi-square test	Sequential feature selection algorithms	Decision tree
Fisher scored	Genetic algorithms	
Correlation coefficient		
Variance threshold		

Table 4 Comparison of the PSO approach with other feature selection approaches mentioned (filter methods, wrapper methods, embedded and methods)

S.no	Year	Authors	Algorithms/techniques used	Results
1	2020	Tuan Minh Le et al. [39]	PSO-based Filter methods (Adaptive Particle Swam Optimization (APSO) to optimize the Multilayer Perceptron (MLP))	97%
2	2020	Anil Kewat et al. [40]	PSO and Wrapper-Based Feature Selection	The outcomes showed that Particle Swarm Optimization method and Genetic Search method can improve the classification performance and outperformed over the Greedy feature selection technique
3	2018	R. Vanaja et al. [41]	Particle Swarm Optimization with Digital Pheromones (PSODP)	The proposed work shows improvement in classification accuracy with minimal time required compared to the existing feature selection and classification techniques
4	2019	Ratna Patil et al. [20]	PSO-ANN-Based Computer-Aided Diagnosis and Classification of Diabetes	PSO functionalities are used to detect a subset of features to accomplish improved classification performance than using entire features set

Advantages and Disadvantages of Particle Swarm Optimization

Advantages:

- Insensitive to scaling of design variables.
- Easily parallelized for concurrent processing.
- Derivative free.
- Very few algorithm parameters.
- A very efficient global search algorithm.

Disadvantages:

- PSO's optimum local search ability is weak..

Equation for the Objective Function they were Maximizing or Minimizing

We are looking to maximize or minimize a function to find the optimum solution. A function can have multiple local maximums and minimum. However, there can be only one global maximum as well as a minimum. If your function is very complex, then finding the global maximum can be a very daunting task. PSO tries to capture the global maximum or minimum. Even though it cannot capture the exact global maximum/minimum, it goes very close to it. It is the reason we called PSO a heuristic model. Particle Swarm Analysis Fish shoaling and bird flocking social behaviors served as inspiration for Eberhart and Kennedy's [42] PSO stochastic optimization method. Each component of the folk elements is represented by a particle in the PSO, which gives physical characteristics like mass and volume. Each component of the folk elements is represented by a particle in the PSO, which gives physical characteristics like mass and volume.

Let us assume a few parameters first. You will find some new parameters, which I will describe later.

F: Objective function, VI: Velocity of the particle or agent, A: Population of agents, W: Inertia weight, C1: cognitive constant, U1, U2: random numbers, C2: social constant, Xi: Position of the particle or agent, Pb: Personal Best, gb: global Best.

The actual algorithm goes as below:

1. Create a 'population' of agents (particles) which is uniformly distributed over X.
2. Evaluate each particle's position considering the objective function (say the below function)

$$z = f(x, y) = \sin^2 + \sin y^2 + \sin x \sin y \quad (1)$$

3. If a particle's present position is better than its previous best position, update it.
4. Find the best particle (according to the particle's last best places).
5. Update particles' velocities.

$$V_i^{t+1} = W \cdot V_i^t + c_1 U_1^t (P_{b1}^t - P_i^t) + c_2 U_2^t (g_b^t - p_i^t) \quad (2)$$

5. Move particles to their new positions.

$$P_i^{t+1} = P_i^t + v_i^{t+1} \quad (3)$$

6. Go to step 2 until the stopping criteria are satisfied.

The operation of PSO is described by Eqs. (1)–(5).

$$X_i = \left(x_{i1}, x_{i2}, \ldots, x_{iD}\right) \quad (4)$$

$$P_i = \left(p_{i1}, p_{i2}, \ldots, p_{iD}\right) \quad (5)$$

$$V_i = \left(v_{i1}, v_{i2}, \ldots, v_{iD}\right) \quad (6)$$

$$V_{id} = w * v_{id} + c1 * r1 * \left(P_{id} - X_{id}\right) + c2 * r2 * \left(P_{gd} - X_{id}\right) \quad (7)$$

where the current position of a particle is x_{id}, the best of the particle is P_{ID}, the best of the group is paged, the velocity of particle is $\mathrm{vo}_{\mathrm{id}}$, the entire factor is we, the relative influence of the cognitive component is c_1, the relative influence of the social component is c_2, and r_1, r_2 are random numbers. r_1, r_2 are employed to keep the population's change spread between [0, 1], equally. The c_1 and c_2 are the self-recognition t constant and the social component coefficient, as shown in Eq. (5).

$$w = w_{\max} - \frac{w_{\max} - w_{\min}}{\mathrm{iter}_{\max}} \quad (8)$$

where the initial weight is shown by $w_{\max}$, the final weight is shown by $w_{\min}$, the maximum iteration number is shown by $\mathrm{iter}_{\max}$, and the current iteration number is shown by iter.

A particle swarm optimisation operates in this manner. We start with a number of random locations on the plane (call them particles) and let them search for the minimum point in a variety of directions, much like a flock of birds searching for food. Every particle should look around the lowest position it has ever found as well as the lowest point the entire swarm of particles has ever found at each step. We regard the minimal point of the function to be the last point that this swarm of particles has ever investigated after a specific number of iterations.

Assume we have P particles, and we denote the position of particle I at iteration t as $X^i(t)$, which in the example of above, we have it as a coordinate $X^i(t) = (X^i(t), y^i(t))$. Besides the position, we also have a velocity of each particle, denoted as $V^i(t) = (V_y^i(t), y_y^i(t))$. At the next iteration, the position of each particle would be updated as

$$X^i(t+1) = X^i(t) + V^i(t+1) \quad (9)$$

Or, equivalently,

and at the same time, the velocities are also updated by the rule.

$$V^i(t+1) = wV^i(t) + c_1 r_1 \left(\mathrm{pbest}^i - X^i(t)\right) + c_2 r_2 \left(\mathrm{gbest} - X^i(t)\right) \quad (10)$$

where r_1 and r_2 are random number between 0 and 1, constants w, c_1, and c_2 are parameters to the PSO algorithm, and the $\mathrm{best}^{\mathrm{I}}$ is the position that gives the best of (X) value ever explored by particle I and gbest is that explored by all the particles in the swarm.

Note that $\mathrm{pbest}^{\mathrm{I}}$ and X^i (t) are two position vectors and the difference pbest ^{I}-X^i (t) is vector subtraction. Adding this subtraction to the original velocity $Vi(t)$ is to bring the particle back to the position pbest^i. Similar are the differences gbest-X^i (t).

We call the parameter W the inertia weight constant. It is between 0 and 1 and determines how much should the particle keep on with its previous velocity (i.e., speed and direction of the search). The parameters $C1$ and $C2$ are called the cognitive and the social coefficients respectively. They control how much weight should be given between refining the search result of the particle itself and recognizing the search result of the swarm. We can consider these parameters control the tradeoff between exploration and exploitation.

$$V_i^{t+1} = W \cdot V_i^t + c_1 U_1^t (P_{b1}^t - P_i^t) + c_2 U_2^t (g_b^t - P_i^t) \quad (11)$$

W = The parameter W is the inertia weight, and it is a positive constant, this parameter is important for balancing the global search, also known as exploration (when higher values are set), and local search, known as exploitation (when lower values are set).

$W \cdot V_i^t$

- Diversification: searches for new solutions, finds the regions with potentially the best solutions.
- Inertia: Makes the particle move in the same direction and with the same velocity.

$c_1 U_1^t (P_{b1}^t - P_i^t)$ = Personal Influence: Improves the individuals. Makes the particle return to a previous position, better than the current.

$c_1 U_1^t (P_{b1}^t - P_i^t) + c_2 U_2^t (g_b^t - P_i^t)$ = Intensification: explores the previous solutions, and finds the best solution of a give's regions.

$c_2 U_2^t(g_b^t - P_i^t)$ = Social Influence: Makes the particle follow the best neighbor's direction.

If $W = 1$, the particle's motion is entirely influenced by the previous motion, so the particle may keep going in the same direction. On the other hand, if $0 \leq W < 1$, such influence is reduced, which means that a particle instead goes to other regions in the search domain.

Pb1t and its current position pit. It has been noticed that the idea behind this term is that as the particle gets more distant from the Pb1t (Personal Best) position, the difference (Pb1t-Pit) must increase; hence, this term increases, attracting the particle to its best own position. The parameter C1 existing as a product is a positive constant, and it is an individual-cognition parameter. It weighs the importance of the particle's own previous experiences.

The other hyper-parameter which composes the product of the second term is U1t. It is a random value parameter within the [0, 1] range. This random parameter plays an essential role in avoiding premature convergences, increasing the most likely global optima.

The difference (gbt-Pit) works as an attraction for the particles toward the best point until it is found at t iteration. Likewise, C2 is also a social learning parameter, and it weighs the importance of the global learning of the swarm. And U2t plays precisely the same role as U1t.

In the case of $C1 = C2 = 0$, all particles continue flying at their current speed until they hit the search space's boundary.

In cases $C1 > 0$ and $C2 = 0$, all particles are independent.

In cases $C1 > 0$ and $C2 = 0$, all particles are attracted to a single point in the entire swarm.

In case $C1 = C2 \neq 0$, all particles are attracted toward the average of pbest and gbest.

Feature Selection

A preprocessing method called feature selection identifies the main characteristics of a particular situation. It has historically been used to solve various issues, such as analyzing biological data, financial matters, and intrusion detection systems. Medical applications have effectively employed feature selection to reduce dimensionality and better understand the root causes of disease [43]. Traditional feature selection algorithms do not try to capture causal relationships between features; instead, they choose parts based on the correlations between predictive characteristics and the class variable. Since causal linkages suggest the underlying mechanism of a system, it has been demonstrated that knowledge of the causal relationships between elements and the class variable may be useful for developing interpretable and reliable prediction models. As a result, several algorithms have been presented, and causality-based feature selection has increasingly gained more attention [44]. There are three feature selection strategies: filtering, wrapping, and embedded. Also, the comparison of the performance of other feature selection approaches is shown in Table 5.

Filtering Methods

Using an indirect criterion, such as the distance criterion, which shows how well the classes are separated, filtering methods evaluate the accuracy of predictions or classifications. Usually, this technique is applied as a preliminary step. Instead, the features are chosen to be related to the outcome variable based on how well they perform in various statistical tests.

Table 5 Comparing the performance of other feature selection approaches

S.no	Year	Authors	Algorithms/techniques used	Result
1	2018	Yap Bee Wah et al. [45]	This study contrasts the filter and wrapper feature selection approaches to increase classifier accuracy	According to the simulation results, the wrapper technique (sequential forward selection and backward elimination) selected the right features more accurately than the filter method
2	2019	Xing Song et al. [46]	Used six feature ensemble strategies and three machine-learning-based embedded feature selection methods to choose the top-ranked features for forecasting DKD onset and robustness to data disturbances	The weighted mean rank feature ensemble technique combined with the gradient boosting machine (GBM) performed best, with an AUC of 0.82 (95% CI: 0.81–0.83) on internal validation and 0.71 (95% CI: 0.68–0.73 on external temporal validation. The ensemble model identified a set of 440 features from 84 872 distinct clinical features, including 191 labs, 51 visit details (primarily vital signs), 39 medications, 34 orders, 30 diagnoses, and 95 other clinical features, that are both predictive of DKD onset and robust against data perturbations

Wrapper Methods

Wrapper approaches assess a subset of genes throughout the search phase using a search strategy and a learning model. The wrapper methods typically outperform filter methods in classification accuracy due to a learning model. On the other hand, they have a few drawbacks, including a large computational overhead and the potential for overfitting.

Embedded Methods

These techniques choose features during the learning process and are typically given to students. This model also takes advantage of the previous models using different evaluation criteria in different search stages. Filter and wrapper characteristics are combined with embedded methods. Algorithms use these internal feature selection techniques. Compare the performance of other feature selection showed in Table 3.

Differences Between Filter and Wrapper Methods

The following are the key variables between feature selection processes using wrapper and filtering:

- Since filter methods do not require model training, they are substantially faster than wrapper approaches. Wrapper approaches, on the other hand, also cost a lot to compute.
- Cross-validation is used by wrapper techniques, while statistical methods are used by filtering methods to examine a subset of characteristics.
- Wrapper methods may always offer the best feature subset, whereas filtering methods frequently fail to do so.
- Wrapper methods may always offer the best feature subset, whereas filtering methods frequently fail to do so.
- The model is more vulnerable to employing a subset of filter method characteristics when using a set of wrapper method features.

Feature Selection Techniques Using PSO Algorithms

Eberhart and Kennedy devised the PSO, which is a population-based method. PSO is a well-known and successful worldwide search method. It is an excellent technique for feature selection issues, because it is easy to encode features, has a global search capacity, is computationally acceptable, has fewer parameters, and is easier to apply. The PSO is used to choosing characteristics because of the considerations above. The limitations of feature selection approaches mentioned (Filter methods, Wrapper methods, Embedded and methods) are shown in Table 6.

PSO was used to explore and choose a subset of primary components or the principal features throughout the main space. Particles in PSO represent possible solutions in the search space and form a swarm known as a population. The swarm of particles is created by randomly dispersing 1s and 0s. If the primary component is 1, it is chosen, while the main component of 0 is ignored. As a result, each particle represents a different subset of the primary components. The particle swarm is randomly initiated, and then moved in the search or principal space, updating its position and velocity to find the best collection of characteristics [9, 48, 49]. For example, the Parameter's Initialization PSO-SVM is shown in Table 7.

Motivation

Diabetic disease is typically composed because of higher-than-normal blood sugar levels. Instead, the production of insulin may be regarded insufficient. It has been noted in recent days that the percentage of diabetes-affected patients have grown to a larger extent throughout the world. Evidently, this problem must be taken more seriously in the coming days to ensure that the average percentages of diabetes-affected individuals are reduced. Recently, several research teams conducted detailed research on the machine learning platform to determine the precision of each other.

Table 6 Limitations of other feature selection approaches mentioned (filter methods, wrapper methods, and embedded methods) [47]

Feature Selection Method	Strengths	Weaknesses
Filter-univariate	Independent of classifier Fast and scalable Reduce risk of overfitting	Feature dependencies not modeled Interaction with classifier not modeled
Filter-multivariate	Independent of the classifier Less risk of overfitting Can model feature dependencies?	Slower and not as scalable as univariate filters Interaction with classifier not modeled
Wrapper	Model interaction with classifier Model feature dependencies Better performance than filter method	More prone to overfitting Slower than filter and embedded methods The selected features are classifier dependent
Embedded	Model feature dependencies Faster than wrapper method Model interaction with classifier	The selected features are classifier dependent Slower than filter methods

Table 7 PSO-SVM parameters initialization

Parameter	Value
Population size (n)	{70} particles
Size of particles (d)	No. of features
No. Of iterations	200 either
Velocity initialization	Randomly generated
Positions	Randomly generated
Fitness function	See Eq. 2
Inertia weights	$\omega i \in [0.1, 0.8]$
Local best solution	(ngbr*d) matrix initially zero
Global best solution	(d*itr) matrix initially zero
($c1$, $c2$)	(2, 2)
Number of neighborhoods (c)	5
Degree of connectivity (K)	Randomly initialized

Machine learning can be used by parametric modeling of health data, including diabetic patient data sets, to synthesize expertise in the field. In this study, a model is proposed for Prediction diabetes based on Feature Selection and Machine Learning Algorithms. The combined Particle Swarm Optimization (PSO) and machine Learning Algorithms are used to evaluate a set of medical data relating to a diabetes diagnosis challenge. Experiments are performed on the Diabetes Database. The sensitivity, specificity and accuracy metrics widely used in medical studies have been used to assess the effectiveness of the proposed system reliability. The proposed approach has the potential to be applied for effective and early diagnosis of other medical diseases as well.

The machine learning process can be implemented using various machine learning techniques. The most extensively utilized learning techniques are supervised and unsupervised learning. The supervised learning technique is applied when historical data is available for a specific problem. The system is trained using inputs and replies before being applied to predict new data responses. Artificial neural networks, backpropagation, decision trees, support vector machines, and the Nave Bayes classifier are all examples of supervised techniques. An unsupervised learning technique is applied when the available training data are unlabeled. There is no prior information or training offered in the system. The algorithm must analyze and detect patterns in the available data to make judgments or predictions. K-means clustering, hierarchical clustering, principal component analysis, and the Hidden-Markov model are all examples of unsupervised techniques [19].

Also, it reduces the number of features, eliminates useless, noisy, and redundant data, and generates acceptable classification accuracy. The feature selection process can be considered a global combinatorial optimization problem in machine learning. Feature selection is critical in pattern classification, medical data processing, machine learning, and mining applications. A good feature selection strategy based on the number of characteristic analyses for sample classification is necessary to speed up the processing rate, enhance predicted accuracy, and avoid incomprehensibility. This paper implements feature selection using particle swarm optimization (PSO). Machine learning algorithms using the one-versus-rest strategy are used as a PSO fitness function for the classification problem. The selected features are then used to diagnose diabetes using machine learning algorithms.

Material and Method

Stage 1: Collected Dataset

Pima Indians Diabetes Database

The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDKD) includes cost information (donated by Peter Turney). The selection of these instances from a larger database was subjected to many constraints. All patients are of Pima Indian heritage, female, and at least 21 years old. This study uses the type 2 diabetes dataset from (https://www.kaggle.com/kumargh/pimaindiansdiabetescsv). There are 768 instances in this data set, divided into two groups: diabetic and non-diabetic, with eight risk factors: number of pregnancies, 2-h plasma glucose concentration in an oral glucose tolerance test, diastolic blood pressure, triceps skin fold thickness, 2-h serum insulin, body mass index, diabetes pedigree function, and age, as shown in Table 8. Seventy percent of the information is for training purposes, while 30% is for testing purposes. The data include characteristics like Pregnancy, Glucose, Blood Pressure, Skin-Thickness, Insulin, BMI, Diabetes-Pedigree-Function, Age, and Class.

Table 8 Description of the Pima Indian diabetes datasets

Dataset	Sample size	Feature size, including class label	Classes	Presence of missing attribute	Presence of noisy attributes
Description of the Pima Indian diabetes datasets					
Pima Indian diabetes	768	9	2	NO	NO

Diabetes 130-US Hospitals for Years 1999–2008 Data Set

In addition, algorithms were learned using a different dataset. Two types of diabetic records were used: automatic electronic recording equipment and paper records. The automatic gadget had an inbuilt clock that allowed it to timestamp occurrences, whereas the paper records had "logical time windows" (breakfast, lunch, dinner, and bedtime). Breakfast (08:00), lunch (12:00), dinner (18:00), and the rest (18:00) are all recorded on paper (22:00). As a result, paper records have at notionally consistent recording times, but electronic records have more precise time stamps. These data were analyzed to look for indicators linked to readmission of diabetic patients and other outcomes. For this study, the diabetes dataset available in (https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008), which contains 55 useful variables and 100,000 records, was used. Table 9 lists these variables and acronyms. The information is divided into two categories: training and testing. Training accounts for 80% of the data, while testing accounts for 20%.

The dataset represents clinical care in 130 US hospitals and integrated delivery networks for ten years (1999–2008). There are more than 50 elements that can be used to depict patient and hospital outcomes. The following conditions had to be followed for interactions to remove data from the database.

[1] This is a visit to the hospital (a hospital admission).
[2] It is a diabetic encounter, meaning that diabetes was diagnosed during the interaction.
[3] The duration of the stay ranged between 1 and 14 days.
[4] Throughout the meeting, lab tests were being done.
[5] Drugs were administered throughout the exchange.

The information includes:

- Details about the patient's number, race, gender, and age.
- The type of admission and length of hospital stay.
- The medical specialty of the admitting physician.
- The number of lab tests performed.
- The HbA1c test results.
- The diagnosis.
- The number of medications.
- The number of diabetic medications.
- The quantity of outpatient, inpatient, and emergency visits in the year before the hospitalization.
- Etc.…

Diabetes Iraqi Society Data Set

The diabetic data set's structure was discussed. The data were gathered in Iraqi society and came from the Medical City Hospital's laboratory (the specialized center for endocrinology and diabetes—Al-Kindy Teaching Hospital). To construct the diabetes dataset, patient records were created from which data were extracted and entered into a database. Medical information and test results are included in the data. The data attribute reads, "The data comprises medical notes, laboratory analyses, and other related information." The data attribute is the data that consists of.

1. Medical notes,
2. Laboratory analyses, etc.
3. The information that is initially entered into the system are

- The Number of patients
- Blood glucose level
- Age
- Sex
- Creatinine (Cr)
- Body mass index (BMI)
- Urea
- Cholesterol (Chol)
- Fasting lipid profile, including total
- LDL
- VLDL
- Triglycerides (TG)
- Cholesterol
- HBA1C.

And the class (the patient's disease class is also diabetic, non-diabetic, or pre-diabetic). The dataset (https://data.mendeley.com/datasets/wj9rwkp9c2/1) contains 14 useful variables and 1000 records. Table 10 lists these variables and acronyms. The information is divided into two categories:

Table 9 Description of the diabetes 130-US hospitals for 1999–2008 data set

Dataset	Sample size	Feature size, including class label	Classes	Presence of missing attribute	Presence of noisy attributes
Diabetes 130-US hospitals for years 1999–2008 data set					
Diabetes 130-US hospitals for years 1999–2008 data set	100.000	55	Multivariate	Yes	No

Table 10 Description of the diabetes Iraqi society data set

Dataset	Sample size	Feature size, including class label	Classes	Presence of missing attribute	Presence of noisy attributes
Iraqi society					
Iraqi society	1000	14	2	Yes	No

Table 11 Techniques of scale features

$X' \frac{x-\text{mean}(x)}{a}$ X' = New value Mean = original value A = standard deviation	$X' \frac{x-\min(x)}{\max(x)-\min(X)}$ X' = New value X-min(x) original value
Standardization	Normalization

training and testing. Training accounts for 80% of the data, while testing accounts for 20%.

Stage 2: Data Preprocessing

Data processing is transforming data from one format into another that is more usable, desirable, meaningful, and instructive. Machine learning techniques, mathematical modeling, and statistical expertise can all be used to automate this procedure [50–52]. Outliers and missing data were removed from the clinical data. Each case with missing survival information was eliminated from the analysis to develop a credible model. In addition, mean and mode imputation techniques were used to treat the remaining missing data. This was accomplished utilizing Python software and data mining techniques.

Need for Data Preprocessing

The data must be properly prepared to generate better results from the model used in machine learning applications. Some machine learning models need the data to be in a certain format; for instance, the Random Forest method cannot handle null values. Therefore, null values from the initial raw data set must be treated before the algorithm can run. How the data set is organized should also be considered to run various Machine Learning and Deep Learning algorithms simultaneously and select the best of them. The following approaches were employed in this article:

1. Handling Null Values: Every real-world dataset contains a small number of null values. Whether the issue is classification, regression, or any other kind, no model can handle NULL or Nan variables independently, so we must step in.
2. Standardization: This stage in the preprocessing procedure is crucial. We may standardize our data by giving a mean of 0 and a standard deviation of 1. In machine learning, there are two techniques to scale features (Table 11).
3. Data Reduction: A large database may become slower, cost more to access, and be more challenging to store efficiently. In a data warehouse, data reduction seeks to produce a more straightforward version of the data.
4. Rescale Data: Rescaling our data's attributes to the same scale will help various machine learning techniques when our data contains variables of different sizes. This is helpful for machine learning methods that employ gradient descent and other optimization techniques. It is also beneficial for weighted input algorithms like regression and neural networks, as well as distance-based algorithms like K-Nearest Neighbors.
5. Binarize Data (Make Binary): A binary threshold can be used to modify our data. When a value exceeds or equals the threshold, it is indicated with a 1; when it is equal to or less than the threshold, it is marked with a 0. This process of thresholding or binarizing data is known. It can be useful if you have probabilities that you want to convert to precise values. It is also beneficial when you are doing feature engineering and want to add new features that imply something. You can make brand-new binary attributes.

Stage 3: Proposed Method

For effective machine learning model creation. The majority of attributes are typically irrelevant to supervised machine learning categorization. Feature selection and outlier elimination were part of the raw data preprocessing phase. There are several approaches to dealing with outside and inconsistent data. We chose the qualities in our study that had significantly connected data. A feature subset selection based on PSO is proposed in the second stage. After preprocessing and feature selection, the integrated dataset is subjected to classification algorithms.

The project is divided into two halves. The first is the feature selection approach, which focuses on obtaining more relevant This article discusses various approaches and datasets for evaluating the performance of different machine learning algorithms. Figure 1 depicts the study's recommended methodology. This study's methodology is divided into three key steps: data collecting, preprocessing, and classification. The dataset used for the analysis is the diabetes of the study. The proposed method uses data from three

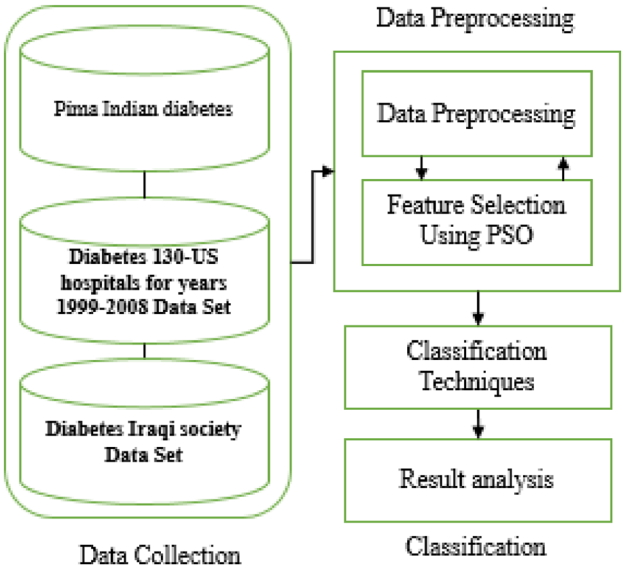

Fig. 1 Methodology followed in the study

different profiles and is based on an integrated methodology. On the other hand, the medical dataset has a lot of missing and irrelevant data that cannot be used for categorization. As a result, the initial phase of the strategy is preparing the dataset using typical imputation techniques in accordance with the data profiles.

In machine learning applications, feature engineering is a critical stage. Modern data sets are defined with several property features while discarding the irrelevant for faster and more efficient data classification. The second stage applies the classification algorithms to the collected parts to produce predictions.

Therefore, the objective of this study Comparison of machine learning algorithms in diagnosing diabetes. Thus, to compare the behavior of LR, NB, KNN, DT, RF, SVM, GB, SGDA, and C4.5, we conducted an experiment evaluating the algorithms' effectiveness and efficiency. Specifically, the research questions we set for the study area:

1. Which algorithm is the most effective?
2. Which one is the most efficient?
3. Which one is the most accurate?

Evaluation of Result

This section presents the results of the information analysis. To apply and evaluate our classifiers, we employed the tenfold Cross-Validation test, a technique for assessing predictive models in which the original set is split into a training sample for training the model and a test set for assessment. After performing the preprocessing and preparation techniques, we visually analyze the data and determine the distribution of values in terms of effectiveness and efficiency.

Table 12 Confusion matrix

Confusion matrix		Classified Ads	
		Negative	Positive
Actual class	Negative	TN	FP
	Positive	FN	TP

The classification cost can be represented by a cost matrix that can identify two types of positive mistakes for classification problems with two categories. The performance measurement is used to determine the effectiveness of the classification method. (FP) In addition, as shown in Table 12, false negatives (FN) and two types of classifications, true-positive (TP) and true-negative (TN), have different costs and benefits.

A confusion matrix is a table that describes how well a classification model (or "classifier") performs on a set of experimental data with known right values. If you have an unequal number of observations in each class or your dataset has quite two categories, classification accuracy alone may be misleading. Calculating a confusion matrix might help you better understand what your classification model gets right and where it goes wrong. The Detail Descriptions of Performance Measures are shown in Table 13.

Accuracy The simplest measure of performance accuracy is classification accuracy, which is defined as the percentage of properly predicted batches obtained using the formula [50–56].

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{12}$$

Sensitivity Real positive rate: If the person's result is positive, the model will be positive in a small percentage of cases, as computed by the formula below.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{13}$$

Propertiesss True-negative rate: If the person's result is negative, the model will likewise have a negative result in certain cases, as computed by the method below.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{14}$$

PPV How likely would a person get diabetes if the model is positive?

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{15}$$

Table 13 Detail descriptions of the performance measures [50–56]

Performance Measure	Description
TP	When the positive samples are classified accurately
TN	When the negative samples are classified accurately
FP	When the negative examples are misclassified
FN	When the positive samples are misclassified
Accuracy	This is the overall percentage classification accuracy resulting from a standard classifier
Sensitivity	It determines the proportion of true-positive samples to total samples and is called the True-Positive Rate
Specificity	It indicates the proportion of true-negative samples out of the total samples and is called the false-positive rate

Table 14 Software requirements

Distribution	Anaconda navigator and Google Colab
Packages	Matplotlib, NumPy, pandas, Sci-kit learn
Language	Python 3.7
IDE	Jupiter Notebook (google collab)

NPV How likely would a person get diabetes if the model is positive?

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}} \tag{16}$$

In this section, we assess the efficacy of all classifiers in terms of the time it takes to build the model, the number of correctly categorized examples, the number of misclassified instances, and accuracy. This article was created with the Python 3.7 programming language in the Jupyter Notebook platform's Anaconda environment. Table 14 shows the implementation details.

Results

Patients' quality of life and life expectancy can benefit from early diabetes diagnosis. Different diabetes detection models [19] have been developed using supervised algorithms. In almost every classification task, the dataset comprises many features. However, because some features are useless and duplicated, they are not required for good classification performance. As a result, classifiers with fewer characteristics but higher classification accuracy are preferred for ease of interpretation. Due to improved representation, the ability to explore huge spaces, being more cost-effective computed, being easier to implement, and requiring fewer parameters, PSO is an excellent technique for feature selection problems. This work compared a particle swarm optimization algorithm and ten machine learning algorithms. The Bayesian information criterion (Accuracy) is proposed as a fitness function. Table 15 shows the feature selection results with the particle swarm algorithm on each data set.

Table 15 Result of feature selection

Diabetes Iraqi society data set	Pima Indian diabetes datasets	Diabetes 130-US hospitals for years 1999–2008 data set
No_Pation	Blood pressure	Gender
Cr	Pregnancies	Age
TG	Blood pressure	Admission type id
BMI	Skin-thickness	Discharge disposition id
	Insulin	Dag 2
		Number diagnoses
		Diabetes Med

All classification techniques were experimented with in "Jupyter Notebook" programming in Python.

- Feature selection Diabetes Iraqi society Data Set:

Number of Features in Subset: 4
Individual: [1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1]
Feature Subset: ['No_Pation', 'Cr', 'TG', 'BMI']

- Feature selection Pima Indian diabetes datasets:

Number of Features in Subset: 4
Individual: [1, 0, 1, 1, 1, 0, 0, 0]
Feature Subset: ['Pregnancies', 'Blood Pressure', 'Skin-Thickness', 'Insulin']

- Feature Selection Diabetes 130-US hospitals for years 1999-2008 Data Set:

Number of Features in Subset: 7
Individual: [1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1]

Feature Subset: ['gender', 'age', 'admission_type_id', 'discharge_disposition_id', 'diag_2', 'number_diagnoses', 'diabetesMed']

As we mentioned in "Materials and methods" Section the selected features are (shown in Table 15) used to diagnose diabetes using machine learning algorithms. We can notice from Table 16 that SVM, SGDA, and C4.5 take about 0.09 s to create their models, whereas NB, KNN, and DT take just 0.01 s. Conversely, the accuracy obtained by RF (98.81%) is healthier than that obtained by LR, NB, KNN, DT, SVM, GB, SGDA, and C4.5, which have an accuracy that varies between 90.00 and 98.01 attempts. It may also be easily seen that RF has the best value of correctly classified instances and lower value of incorrectly classified instances than the opposite classifier. The results are shown in Table 16 and Fig. 2.

The data set has been partitioned into two parts (training and testing). We trained our model with 70% training data and tested it with 30% remaining data. Five models have been developed using supervised learning to detect whether the patient is diabetic or non-diabetic. For this purpose, Logistic Regression (LR), Naive Bayes Classifier (NB), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Support Vector Machines (SVM), Gradient Boosting (GB), and Stochastic Gradient Descent Algorithm (SGDA) algorithm is used.

Figure 2 shows the accuracy of the nine classification models when applied to the dataset. As shown in Fig. 2, the decision trees and random forests perform better than other algorithms. Simulation error is also considered in this study to measure the performance of classifiers better. To do so, we evaluate the effectiveness of our classifier in terms of:

- Kappa statistic (KS) as a chance-corrected measure of agreement between the classifications and the actual classes,

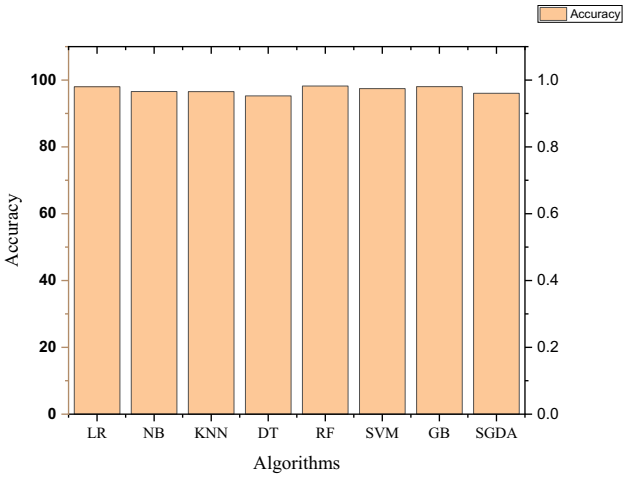

Fig. 2 Accuracy of the classifiers machine learning algorithms

$$k = \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e} \tag{17}$$

- Mean Absolute Error (MAE) as to how close forecasts or predictions are to the eventual outcomes,

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n} \tag{18}$$

- Root Mean Squared Error (RMSE),

$$RMSD(\theta) = \sqrt{MSE(\theta)} = \sqrt{E\left((\theta - \theta)^2\right)}. \tag{19}$$

- Relative Absolute Error (RAE),
- Root Relative Squared Error (RRSP).

KS, MAE, and RMSE are in numeric values. RAE and RRSE are in percentage. The results are shown in Table 17.

Once the predictive model is built, we can check its efficiency. For that, we compare the accuracy measures based on precision, recall, TP rate, and FP rate values for LR, NB,

Table 16 Compared evaluation time to build a model (s) classifiers

Evaluation criteria	Classifiers								
	LR	NB	KNN	DT	RF	SVM	GB	SGDA	C4.5
Pima Indians diabetes database									
Time to build a model (s)	0.02	0.05	0.02	0.01	0.03	0.02	0.05	0.04	0.03
Accuracy (%)	97.26	95.25	96.26	97.88	98.68	95.89	95.05	94.26	94.00
Diabetes 130-US hospitals for years 1999–2008 data set									
Time to build a model (s)	0.05	0.04	0.05	0.04	0.06	0.06	0.05	0.03	0.05
Accuracy (%)	98.65	97.89	97.56	98.00	98.79	97.56	97.26	98.05	95.15
Diabetes Iraqi society data set									
Time to build a model (s)	0.05	0.02	0.06	0.04	0.05	0.02	0.03	0.04	0.06
Accuracy (%)	98.00	96.56	96.52	98.25	98.21	97.43	98.02	96.05	97.25

Table 17 Comparative evaluation Kappa Statistic (KS), Mean Absolute Error (MAE), Root-Mean-Square Error, Relative Absolute Error, and Root Relative Squared Error Classifiers

Evaluation criteria	Classifiers								
	LR	NB	KNN	DT	RF	SVM	GB	SGDA	C4.5
Pima Indians diabetes database									
Kappa Statistic (ks)	0.91	0.89	0.90	0.92	0.93	0.95	0.88	0.89	0.88
Mean Absolute Error (MAE)	0.04	0.06	0.02	0.01	0.08	0.07	0.06	0.04	0.09
Root-Mean-Square Error	0.25	0.16	0.13	0.24	0.18	0.15	0.21	0.16	0.13
Relative Absolute Error	13	12	8.41	11.12	12.15	11.23	08.26	14.65	12.02
Root Relative Squared Error	32	23	12	36	45	223	25	19	24
Diabetes 130-US hospitals for years 1999–2008 data set									
Kappa Statistic (KS)	0.92	0.92	0.92	0.92	0.92	0.91	0.92	0.93	0.89
Mean Absolute Error (MAE)	0.05	0.04	0.06	0.04	0.08	0.04	0.05	0.04	0.06
Root-Mean-Square Error (RMSE)	0.15	0.14	0.19	0.23	0.24	0.21	0.16	0.14	0.18
Relative Absolute Error (RAE) %	11.14	09.51	10.41	09.79	06.25	10.01	07.94	06.15	13.01
Root Relative Squared Error (RRSE) %	15	35	26	18	27	29	34	31	29
Diabetes Iraqi society data set									
Kappa Statistic (KS)	0.92	0.91	0.92	0.93	0.98	0.92	0.93	0.92	0.93
Mean Absolute Error (MAE)	0.04	0.05	0.04	0.06	0.08	0.04	0.06	0.07	0.04
Root-Mean-Square Error (RMSE)	0.15	0.24	0.25	0.14	0.12	0.24	0.18	0.17	0.17
Relative Absolute Error (RAE) %	09.21	11.21	06.33	0.9.21	11.29	16.12	18.12	09.15	13.01
Root Relative Squared Error (RRSE) %	35	42	15	16	19	33	32	25	16

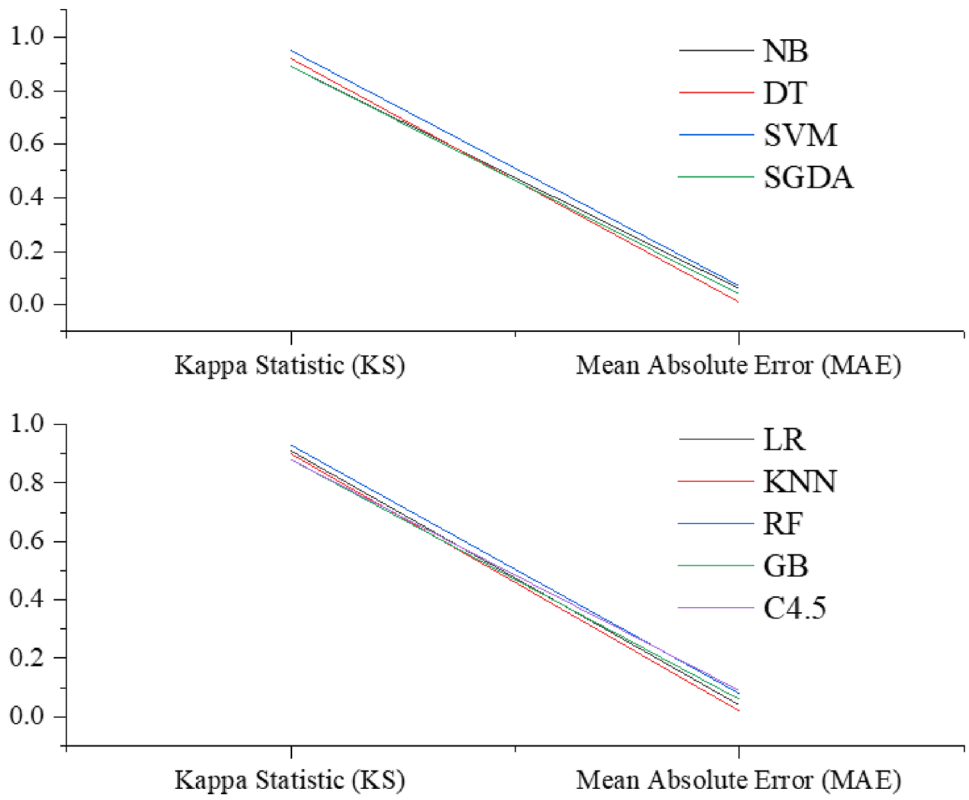

Fig. 3 Training and simulation error Pima Indians diabetes database

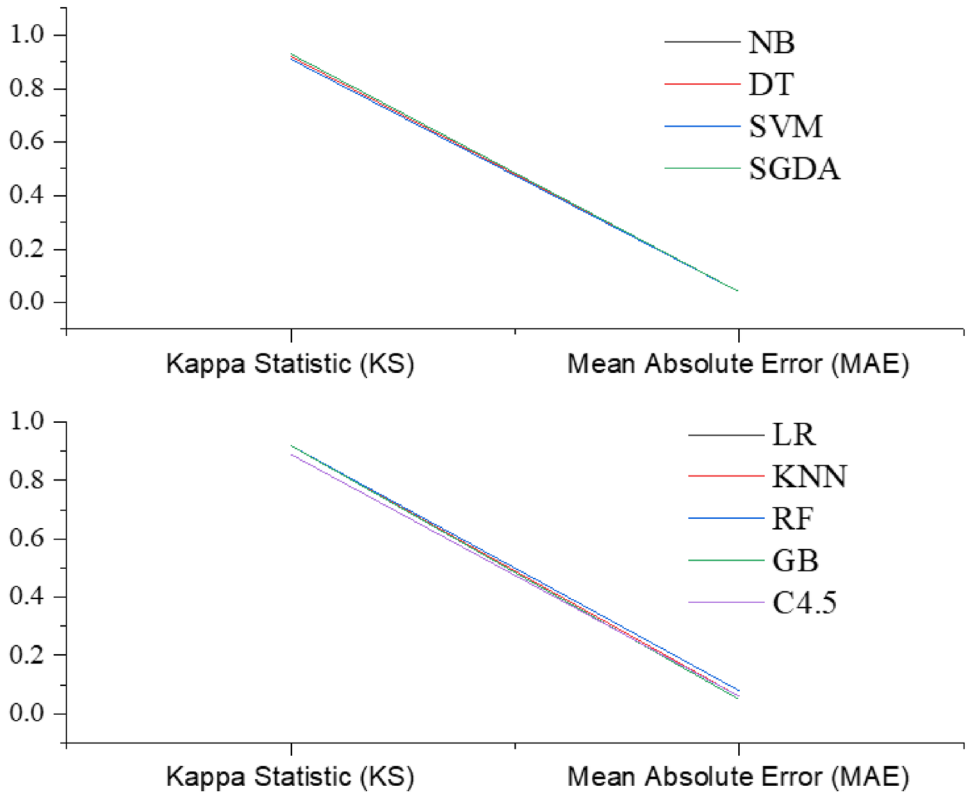

Fig. 4 Training and simulation error diabetes 130-US hospitals for years 1999–2008 data set

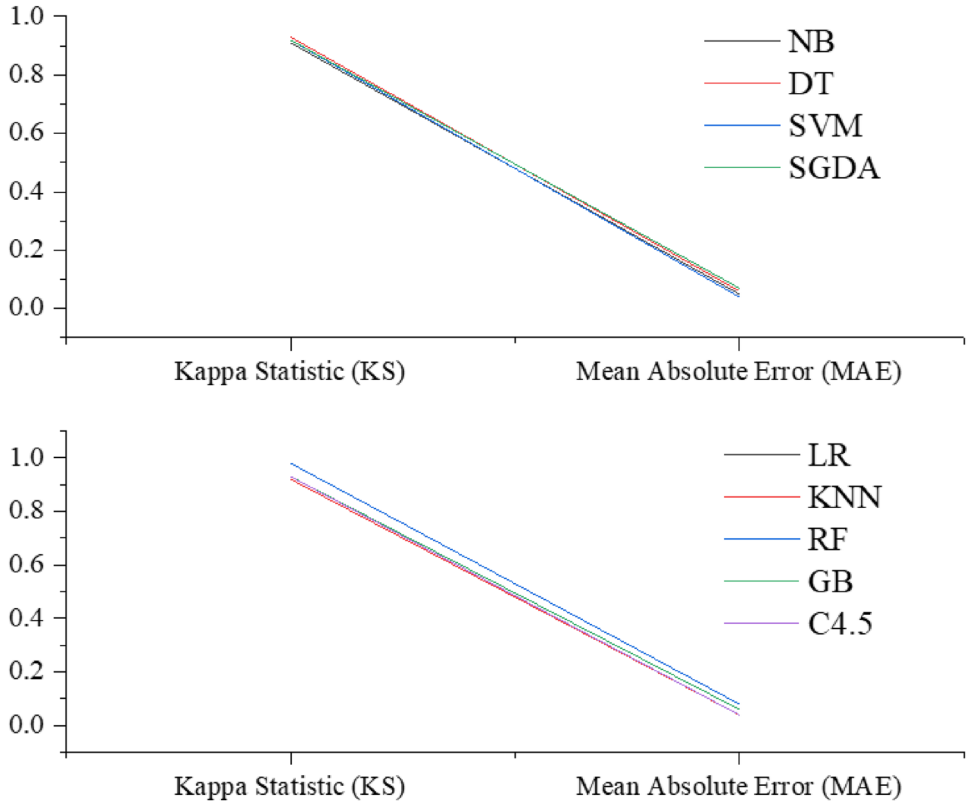

Fig. 5 Training and simulation error Diabetes Iraqi so + . *ciety data set

C, 4.5, DT, RF, SVM, GB, SGDA, and KNN, as shown in Table 17.

Figures 3, 4 and 5 show the results of the mean absolute error (MAE). The square root of the squared error. The mean fundamental error (MAE) measures the error between paired observations expressing the same phenomenon. The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a commonly used measure of the discrepancies between predicted and observed values (sample or population values) by a model or estimator. The RMSD is the quadratic mean of the differences between anticipated and observed values or the square root of the second sample moment of these differences. Interrater dependability is also routinely tested using the kappa statistic.

Table 18 Training and simulation error

	TP	FP	Precision	Recall	F-measure	Class
LR	0.97	0.03	0.95	0.97	0.94	0
	0.96	0.02	0.92	0.96	0.93	1
NB	0.95	0.04	0.93	0.95	0.94	0
	0.96	0.03	0.92	0.96	0.95	1
KNN	0.96	0.03	0.95	0.96	0.94	0
	0.97	0.05	0.96	0.97	0.95	1
DT	0.97	0.3	0.96	0.97	0.94	0
	0.98	0.01	0.99	0.98	0.97	1
RF	0.98	0.01	0.99	0.98	0.98	0
	0.97	0.01	0.98	0.97	0.97	1
SVM	0.95	0.06	0.96	0.95	0.95	0
	0.96	0.05	0.93	0.96	0.92	1
GB	0.95	0.04	0.92	0.95	0.93	0
	0.94	0.03	0.91	0.94	0.93	1
SGDA	0.94	0.03	0.92	0.94	0.93	0
	0.93	0.06	0.94	0.93	0.92	1
C4.5	0.94	0.06	0.93	0.94	0.92	0
	0.92	0.05	0.93	0.92	0.92	1

(Pima Indians diabetes database)

Table 19 Training and simulation error

	TP	FP	Precision	Recall	F-measure	Class
LR	0.98	0.02	0.99	0.98	0.96	Yes
	0.96	0.06	0.95	0.96	0.96	No
NB	0.97	0.04	0.96	0.97	0.97	Yes
	0.95	0.03	0.96	0.95	0.96	No
KNN	0.97	0.04	0.96	0.97	0.95	Yes
	0.96	0.03	0.94	0.96	0.95	No
DT	0.98	0.02	0.96	0.98	0.97	Yes
	0.95	0.06	0.92	0.95	0.95	No
RF	0.98	0.04	0.96	0.98	0.97	Yes
	0.97	0.02	0.95	0.97	0.96	No
SVM	0.97	0.02	0.93	0.97	0.95	Yes
	0.95	0.06	0.92	0.95	0.94	No
GB	0.97	0.03	0.92	0.97	0.94	Yes
	0.98	0.01	0.92	0.98	0.96	No
SGDA	0.98	0.02	0.96	0.98	0.97	Yes
	0.96	0.03	0.94	0.96	0.95	No
C4.5	0.95	0.06	0.98	0.95	0.97	Yes
	0.98	0.02	0.96	0.98	0.97	No

(Diabetes 130-US hospitals for years 1999–2008 data set)

Table 20 Training and simulation error

	TP	FP	Precision	Recall	F-measure	Class
LR	0.98	0.02	0.96	0.98	0.97	N
	0.96	0.06	0.94	0.96	0.95	P
NB	0.96	0.07	0.94	0.96	0.95	N
	0.97	0.05	0.93	0.97	0.95	P
KNN	0.96	0.05	0.92	0.96	0.95	N
	0.95	0.04	0.91	0.95	0.94	P
DT	0.98	0.02	0.94	0.98	0.96	N
	0.97	0.04	0.99	0.97	0.98	P
RF	0.98	0.03	0.96	0.98	0.97	N
	0.99	0.02	0.97	0.99	0.98	P
SVM	0.97	0.01	0.93	0.97	0.95	N
	0.96	0.03	0.98	0.96	0.97	P
GB	0.98	0.05	0.95	0.98	0.96	N
	0.97	0.02	0.95	0.97	0.96	P
SGDA	0.96	0.04	0.92	0.96	0.95	N
	0.93	0.03	0.90	0.93	0.92	P
C4.5	0.97	0.04	0.93	0.97	0.95	N
	0.95	0.06	0.92	0.95	0.94	P

(Diabetes Iraqi society data set)

Tables 18, 19, and 20 show that RF has the most straightforward classification (0.98%) and the lowest warning error rate (0.01). We will also remark that RF has the most straightforward compatibility between the reliability and validity of the data obtained.

We will now study the findings acquired while measuring the efficiency of our algorithms after we have generated the predicted model. The best values were obtained by RF and DT, as shown in Table 14. (99.68%, 99.82%). Based on these findings, we can deduce why the SVM beats the other classifieds.

To diagnose diabetes, the performance of each of the nine models is assessed using parameters such as precision, recall, and F-Measure (Table 21). Tenfold cross-validation is used to avoid the problems of overfitting and underfitting. Our classifier's accuracy reveals how often it is correct to determine whether a patient has diabetes. Precision was utilized to assess the classifier's ability to make accurate positive diabetes predictions. In our research, recall or sensitivity is employed to determine the percentage of actual positive diabetes cases properly detected by the classifier. The capacity of a classifier to distinguish negative diabetes cases is measured by its specificity.

Table 21 Evaluate the efficiency and effectiveness of algorithms in terms of accuracy

	Accuracy								
	Pima Indians diabetes database			Diabetes 130-US hospitals for years 1999–2008 data set			Diabetes Iraqi society data set		
	Holdout	*K*-fold = 5	*K*-fold = 10	Holdout	*K*-fold = 5	*K*-fold = 10	Holdout	*K*-fold = 5	*K*-fold = 10
LR	97.26	98.21	99.65	98.65	98.88	99.16	98.00	98.56	98.98
NB	95.25	97.26	98.88	97.89	98.98	99.86	96.56	97.05	98.88
KNN	96.26	97.56	99.02	97.56	98.86	99.68	96.52	97.86	98.02
DT	97.88	98.65	99.16	98.00	99.59	99.82	98.25	98.65	99.16
RF	98.68	99.16	99.68	98.79	99.00	99.56	98.21	98.65	99.03
SVM	95.89	97.56	99.00	97.56	98.98	99.28	97.43	98.56	98.49
GB	95.05	97.18	98.76	97.26	97.89	98.94	98.02	99.02	99.26
SGDA	94.26	97.25	97.88	98.05	99.14	99.65	96.05	97.25	98.81
C4.5	94.00	96.25	97.02	95.15	96.25	97.02	97.25	98.36	98.09

Discussion

Diabetes is a collection of metabolic illnesses marked by high blood sugar levels caused by a lack of insulin secretion, insulin function, or both. Diabetes-related chronic hyperglycemia is linked to long-term damage, dysfunction, and failure of various organs, including the eyes, kidneys, nerves, heart, and blood vessels. Diabetes must be detected early to maintain a healthy lifestyle. Because diabetes cases are quickly increasing, this disease may cause global concern.

Machine learning (ML) is a computerized method for learning from experience automatically and improving performance to make more accurate predictions. Machine learning techniques are successfully used in various applications, including diagnosis. A machine learning algorithm that develops a classifier system may aid clinicians in identifying and diagnosing diseases at an early stage by generating a classifier system. We will use machine learning classification techniques to improve the speed, performance, reliability, and accuracy of diagnosing this system for a specific ailment. This research focuses on utilizing machine learning approaches to analyze diabetic illness detection.

Kennedy and Eberhart developed particle swarm optimization (PSO) in 1995, a population-based stochastic optimization approach. PSO models species' social behavior, such as bird flocking and fish schooling, to show an autonomously evolving system. PSO refers to each candidate solution as "an individual bird of the flock" or a particle in the search space. Each particle uses memory and the swarm's collective knowledge to choose the best answer (Venter 2002). Each particle has fitness values are maximized using a fitness function and velocities that control particle movement. Each particle adjusts its position during mobility depending upon its own and nearby particle's experiences, selecting the best position it and its neighbor have encountered. The particles follow a current of optimal particles through the problem space [9, 48, 49].

Particle swarm optimization (PSO) was used in a study by Asti Herliana et al. To choose the best diabetic retinopathy feature from a dataset of diabetic retinopathy cases. The selected feature is then further classified via the neural network classification approach. The study's findings indicate a 76.11% improvement in outcome when using neural network-based particle swarm optimization (PSO). According to this study, the classification result has improved by 4.35% when feature selection is used, compared to the prior result of 71.76% when simply utilizing the neural network approach [57].

Using data mining techniques, Xiaohua Li and colleagues published an article on identifying a diabetic patient. Preprocessing, feature selection, and classification are the three steps of the suggested method. With K-means for feature selection, several amalgamations of the Harmony search algorithm, genetic algorithm, and particle swarm optimization algorithm are investigated. The combinations have never been looked at before for applications in diabetes diagnosis. The diabetes dataset is categorized using the K-nearest neighbor algorithm. Sensitivity, specificity, and accuracy have been measured to assess the outcomes. The findings show that the proposed strategy performed better than the earlier methods tested in this paper [58], with an accuracy of 91.65%.

To diagnose various medical conditions, Mohammad Reza Daliri proposes a feature selection technique utilizing a binary particle swarm optimization algorithm. The binary particle swarm optimization's fitness function was implemented using support vector machines. The four databases used to evaluate the suggested technique were the single proton emission computed tomography heart database, the Wisconsin breast cancer data set, the Pima Indians diabetes database and the Dermatology data set. The findings show that using fewer traits could diagnose heart, cancer, diabetes, and erythematosquamous diseases with a higher degree of accuracy. Our approach produced more accurate results when the findings were compared to the F-score and information gain, two classic feature selection techniques. The findings of the suggested method demonstrate a superior accuracy in all but one of the data compared to the genetic algorithm for feature selection. Additionally, the methodology performs better, utilizing fewer characteristics when compared to other methods that employ the same data [59].

Tuan Minh Le et al. [39] suggested a machine learning algorithm to forecast the early onset of diabetes in patients. It is an innovative approach to wrapper-based feature selection that uses Adaptive Particle Swam Optimization (APSO) and Gray Wolf Optimization (GWO) to optimize the Multilayer Perceptron (MLP) and minimize the number of the input characteristics needed. Additionally, they compared the outcomes of this strategy with many well-known machine learning algorithm approaches, including Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbor (KNN), Naive Bayesian Classifier (NBC), Random Forest Classifier (RFC), and Logistic Regression (LR). The computational findings of our suggested method demonstrate that, in addition to requiring significantly less characteristics, higher prediction accuracy can also be attained (97% for APGWO-MLP and 96% for GWO-MLP). This work has the potential to apply to clinical practice and become a supporting tool for doctors/physicians in the following, comparing the related work with the proposed method (Table 22) (Fig. 6).

Numerous expert systems that have been created to improve the accuracy of medical diagnostics assist medical

Table 22 Performance comparison of other feature selection techniques in diabetes diagnosis

Refs.	Authors	Method	Result
[57]	Asti Herliana et al.	Using the diabetic retinopathy dataset, the best diabetic retinopathy feature is chosen using the particle swarm optimization (PSO) approach	The study's findings indicate a 76.11% improvement in outcome when using neural network-based particle swarm optimization (PSO). The results of this study also demonstrate a 4.35% improvement in classification results utilizing the feature selection approach compared to the prior result of 71.76% using the neural network method
[58]	Xiaohua Li et al.	K-means is used to investigate various combinations of the Harmony search algorithm, genetic algorithm, and particle swarm optimization method	The findings show that the proposed approach outperformed the results of the preceding methods evaluated in this article, with an accuracy of 91.65%
[59]	Mohammad Reza Daliri	To diagnose various medical conditions, suggest a feature selection strategy using a binary particle swarm optimization algorithm. The binary particle swarm optimization's fitness function was implemented using support vector machines	The findings show that using fewer traits could diagnose heart, cancer, diabetes, and erythematosquamous diseases with a higher degree of accuracy. Our approach produced more accurate results when the findings were compared to the F-score and information gain, two classic feature selection techniques. The findings of the suggested method demonstrate a superior accuracy in all but one of the data compared to the genetic algorithm for feature selection. Additionally, the methodology performs better, utilizing fewer characteristics when compared to other methods that employ the same data [9]
[60]	Omar S. Soliman et al.	Classification of diabetes mellitus using modified particle swarm optimization and least squares support vector machine	An LS-SVM algorithm is used for classification by finding optimal hyper-plane which separates various classes. The experimental results showed the superiority of the proposed algorithm which could achieve an average classification accuracy of 97.833%
[61]	OO Oladimeji et al.	Classification models for likelihood prediction of diabetes at an early stage using feature selection	Findings The study result show that feature selection helps in getting better model, as it prevents overfitting and removes redundant data. Hence, the study result when compared with previous research shows the better result has been achieved, after it was evaluated based on metrics such as F-measure, Precision Recall curve and Receiver-Operating Characteristic Area Under Curve. This discovery has the potential to impact on clinical practice when health workers aim at diagnosing diabetes disease at its early stage. Originality/value This study has not been published anywhere else
[62]	Seyed RezaKamel et al.	Feature selection using the grasshopper optimization algorithm in diagnosis of diabetes disease	The study result has shown promising accuracy of 97% achieved by the Support Vector Machine (SVM) algorithm
[63]	Jyotismita Chaki et al.	Machine learning and artificial intelligence-based Diabetes Mellitus detection and self-management: A systematic review	This review provides a detailed overview of DM detection and self-management techniques which may prove valuable to the community of scientists employed in the area of automatic DM detection and self-management
	This Paper	Particle swarm optimization (PSO) is employed in this study to implement feature selection, and the classification problem's PSO fitness function is a one-versus-rest machine learning algorithm. Then, a few of these traits are used to diagnose diabetes using machine learning techniques	The findings show that we could diagnose diabetic illnesses more accurately by choosing fewer features. Our approach produced more accurate results when the findings were compared to the F-score and information gain, two classic feature selection techniques. The suggested method outperforms the genetic algorithm for feature selection in terms of accuracy (99.89% for RF using Holdout—MLP, 99.59% for DT using K-fold = 5, and 99.59% for NB using *K*-fold = 10). Additionally, the methodology performs better, utilizing fewer characteristics when compared to other methods that employ the same data. This work has the potential to be useful in clinical settings and serve as a resource for clinicians

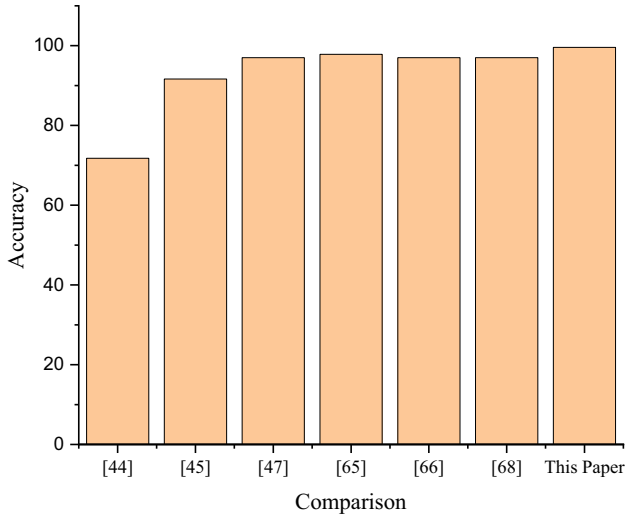

Fig. 6 Performance comparison of other feature selection techniques in diabetes diagnosis

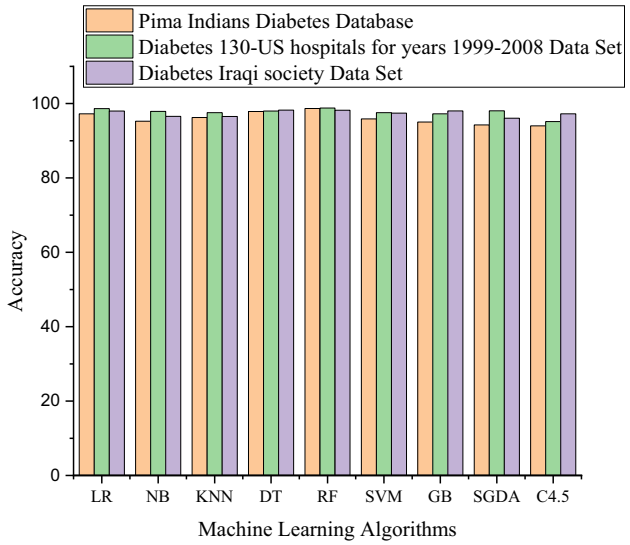

Fig. 7 Evaluate the efficiency and effectiveness of algorithms using Holdout

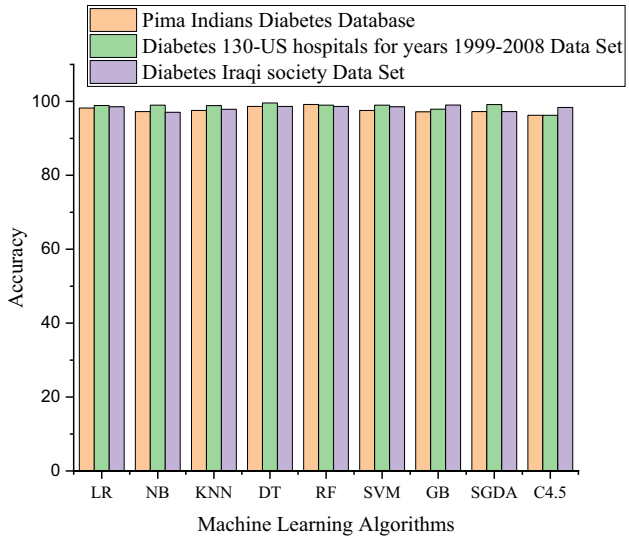

Fig. 8 Evaluate the efficiency and effectiveness of algorithms using K-fold = 5

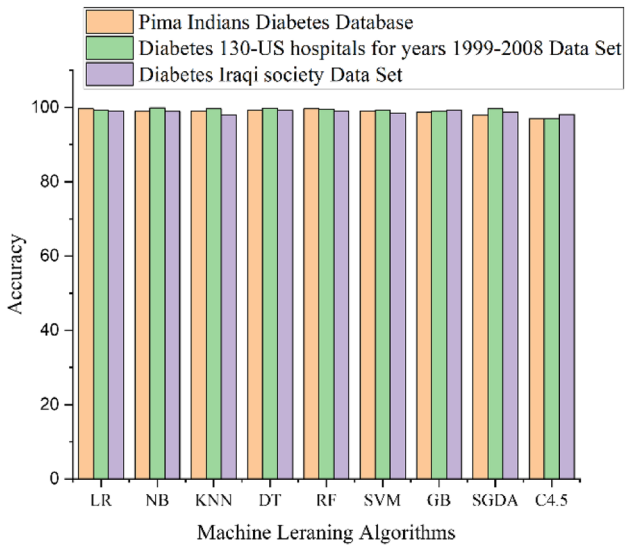

Fig. 9 Evaluate the efficiency and effectiveness of algorithms using k-fold = 10

diagnosis [64]. Table 22 shows that RF, SVM, and C4.5 take about 0.06 s to ble 254uild their model, unlike DT, which takes only 0.01 s. Conversely, the accuracy obtained with RF (98.79%) is healthier than LR, NB, KNN, DT, SVM, RF, GB, SGDA, and C4.5, which have different accuracies between 94.00 attempts to 98.25%. It is also easy to see that RF has the best value of correctly classified instances and lower for incorrectly classified examples than the other classifier.

Figures 7, 8 and 9 show the accuracy of the nine classification models when applied to the dataset. As shown in Fig. 9, the decision trees and random forests have higher performance than the other algorithms.

In summary, RF has demonstrated its effectiveness, efficiency, accuracy, and detectability of support. Compared with a series of diabetes risk prediction research in the literature, our experimental results achieve the best value (99.82%) in diabetes risk prediction classification. RF outperforms other classifiers regarding the accuracy, sensitivity, and specificity in classifying cardiac diabetes. Table 23 shows the performance of machine learning and data mining algorithms with proposed method for classification diabetes.

Table 23 Performance of machine learning algorithms for classification diabetes

S.no	Year	Authors	Algorithms/techniques used	Result (%)
1	2019	Sajida Perveen et al. [65]	Hidden-Markov model (HMM)	86.9
3	2020	Shekharesh Barik et al. [66]	Random forest algorithm	74.10
4	2020	Md EkramulHossain et al. [67]	Six machine learning prediction models	79–88
5	2020	Neha Prerna, and TiggaShruti Garg [68]	Logistic regression algorithm	75.32
6	2021	Minhaz Uddin Emon et al. [69]	Random Forest	98
7	2021	Ram D. Joshi et al. [70]	Utilizing a logistic regression model and decision tree	78.26
8	2021	FayrozaAlaa Khaleel et al. [71]	Logistic Regression (LR), Naïve Bayes (NB), and K-nearest Neighbor (KNN) algorithms	94, 79, and 69
9	2022	This Paper	LR, NB, KNN, DT, SVM, RF, GB, SGDA, and C4.5	94

Through data exchange among intelligent wearables and sensors, the industrial healthcare system has improved the quality of medical services and opened the prospect of implementing enhanced real-time patient monitoring. However, a system of this kind needs to be highly accurate and error-free (Table 24).

Additionally, as is common knowledge, any ML that we employ with data of any kind must be precise, effective, and able to manage data with a wide distribution. A decentralized learning algorithm must be better at managing widely scattered data, since it is more concerned with the distribution of the data. As we saw in the section above of the article, we have several issues with the centralized learning technique on which our majority of traditional models depend. In contrast, swarm learning is a part of the artificial intelligence and machine learning studies where the major focus of swarm learning is to evaluate the behaviors of the decentralized system. We might find it useful to use a decentralized system to get around the drawbacks of centralized learning techniques. The fundamental concept underlying this learning is drawn from the PSO's method of operation.

PSO is a metaheuristic because it can search very huge spaces of potential solutions and makes little to-no assumptions about the problem being optimized. Furthermore, unlike traditional optimization techniques like gradient descent and quasi-Newton methods, PSO does not employ the gradient of the issue being improved, negating the need for the optimization problem to be differentiable. We recommended using metaheuristics like PSO to ensure that an optimal solution is always discovered in a decentralized system. Decentralized AI will also offer a ladder of success that develops from the expansion of knowledge. To exhibit high accuracy and error reduction, we combined PSO with Machine Learning (DL) technique [74–76].

Limitations

The benefits of machine learning techniques are numerous, but they are not without flaws that limit their potential in some respects. For example, many algorithms could be suitable for tackling a particular problem. Similarly, one algorithm may perform well for a given data collection, while others may not. As a result, selecting an acceptable algorithm for a given dataset could be a huge hurdle in bioinformatics, as is deciding on an appropriate feature selection approach. Furthermore, training ML algorithms often necessitates big datasets. These datasets must be unbiased and of good quality. Time is also required for data collection.

Furthermore, ML algorithms require sufficient time to train and test to produce highly reliable outcomes. These

Table 24 Performance of PSO algorithms for feature selection and classification diabetes

Row	Authors	Dataset	Approach	Accuracy (%)
1	Choubey DK et al. [21]	Pima Indian diabetes datasets	PSO-Naive Bayes	92.43
2	Li X et al. [72]	Pima Indian diabetes datasets	PSO-based *K*-means	91.65
3	Santhanam T, and Padmavathi MS [73]	Pima Indian diabetes datasets	*K*-means is used for removing the noisy data and genetic algorithms for finding the optimal set of features with Support Vector Machine (SVM)	96.71
4	Kamel SR, and Yaghoubzadeh R [62]	Pima Indian diabetes datasets	SVM-PSO	93.55
5	This Paper (2023)	Pima Indian diabetes datasets	PSO-SVM	99
		Diabetes 130-US hospitals for years 1999–2008 data set	PSO-SVM	99.28
		Diabetes Iraqi society data set	PSO-SVM	98.49

methods need a significant amount of hardware and resources. In addition, ML algorithms have a hard time confirming their results. As a result, proving that their predictions work in all cases is tough.

The correct analysis and interpretation of the findings generated by ML algorithms are, once again, a major problem in their utilization. Finally, machine learning algorithms are prone to errors. They generate false results when trained with faulty or incomplete data. This can set off a cascade of diagnosis or medication errors that wreak havoc in the healing process. If these problems are detected, detecting the cause of mistakes takes time, and correcting these errors is even more difficult.

Conclusion and Future Work

Detecting the dangers of diabetes at an early stage is one of the world's most pressing health concerns. Machine learning and deep learning have been successfully utilized in medical image and healthcare [52] analysis like whole-slide pathology [54], X-ray [50], diabetes [1, 2], breast cancer [51], heart [53], time series [77], Medicinal Plants [55], stock market [78], Stroke [79], Maximizing the Impact on Social Networks [35], outcome prediction of bupropion exposure [20], etc. This research aims to develop a framework for predicting the likelihood of developing diabetes. This paper compared the outcomes of nine machine learning classification algorithms with various statistical measures. The dataset collected through the UCI site was subjected to tests.

There are also many data processing and machine learning strategies for analyzing medical knowledge. Producing accurate and computationally affordable classifiers for medical applications is a significant challenge in data processing and machine learning. On the diabetes datasets, this study used nine primary algorithms: LR, NB, C 4.5, DT, RF, SVM, GB, SGDA, and KNN. To select the best algorithm—classification accuracy, we sought to analyze the efficiency and efficacy of various algorithms in terms of accuracy, sensitivity, and specificity. Random forest and decision trees performed better than all other algorithms. In conclusion, DT, NB, and RF proved their strength in diagnosing and identifying diabetes and achieved the simplest performance, accuracy, and low error rate.

The findings show that by choosing fewer variables, we could diagnose diabetes illnesses with a higher degree of accuracy. Our method produced more accurate results when the outcomes were compared to the usual feature selection approaches, namely the F-score and the information gain. The accuracy of the suggested method is higher than that of the genetic algorithm for feature selection (99.79% for RF using Holdout—99.59% for DT using K-fold = 5, and 99.86% for NB using K-fold = 10). Additionally, the strategy had a superior performance utilizing fewer features than other methods that employed the same data. This work has the potential to be useful in clinical practice and serve as a tool for doctors and other medical professionals.

In the future, the performance of the machine learning classifier can be improved by feature subset selection using Ant Colony Optimization Algorithm process, and like XGBoost, Extreme Learning Machine, Ensemble Learning Classifiers, and Neural Network.

Author Contributions JA: designed and performed experiments and analyzed data. SA supervised the findings of this work and co-wrote the paper. All authors discussed the results and contributed to the final manuscript.

Funding None.

Data availability https://archive.ics.uci.edu/dataset/34/diabetes.

Declarations

Conflict of Interest None declared.

Ethical Approval Not required.

References

1. Abdollahi J, Moghaddam BN, Parvar ME. Improving diabetes diagnosis in smart health using a genetic-based ensemble learning algorithm. Approach to IoT infrastructure. Future Gen Distrib Syst J. 2019;1:23–30.
2. Abdollahi J, Nouri-Moghaddam B. Hybrid stacked ensemble combined with geneticalgorithms for diabetes prediction. Iran J Comput Sci. 2022;5:1–16.
3. Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Sci Rep. 2020;10(1):1–12.
4. Tigga NP, Garg S. Prediction of type 2 diabetes using machine learning classification methods. Procedia Comput Sci. 2020;167:706–16.
5. Younus M, Munna MTA, Alam MM, Allayear SM, Ara SJF (2020) Prediction model for prevalence of type-2 diabetes mellitus complications using machine learning approach. In: Data Management and Analysis. Springer, Cham, pp 103–116
6. Perveen S, Shahbaz M, Saba T, Keshavjee K, Rehman A, Guergachi A. Handling irregularly sampled longitudinal data and predictive modeling of diabetes using machine learning technique. IEEE Access. 2020;8:21875–85.
7. Kalra S, Singal A, Lathia T. What's in a name? Redefining type 2 diabetes remission. Diabetes Therapy. 2021;12:1–8.
8. Saru S, Subashree S (2019) Analysis and prediction of diabetes using machine learning. Int J Emerg Technol Innov Eng 5(4)
9. Ahmad I. Feature selection using particle swarm optimization in intrusion detection. Int J Distrib Sens Netw. 2015;11(10): 806954.
10. Prasad KS, Reddy NCS, Puneeth BN. A framework for diagnosing kidney disease in diabetes patients using classification algorithms. SN Comput Sci. 2020;1(2):1–6.

11. Chen M, Hao Y, Hwang K, Wang L, Wang L. Disease prediction by machine learning over big data from healthcare communities. Ieee Access. 2017;5:8869–79.
12. Rahman RM, Afroz F. Comparison of various classification techniques using different data mining tools for diabetes diagnosis. J Softw Eng Appl. 2013;6(03):85.
13. Nagarajan S, Chandrasekaran RM. Design and implementation of expert clinical system for diagnosing diabetes using data mining techniques. Indian J Sci Technol. 2015;8(8):771–6.
14. Yıldırım EG, Karahoca A, Uçar T. Dosage planning for diabetes patients using data mining methods. Procedia Comput Sci. 2011;3:1374–80.
15. Garga SB, Mahajanb AK, Kamalc TS (2017) An approach for diabetes detection using data mining classification techniques. Int J Eng Sci
16. Maniruzzaman M, Rahman MJ, Ahammed B, Abedin MM. Classification and prediction of diabetes disease using machine learning paradigm. Health Inf Sci Syst. 2020;8(1):1–14.
17. Shakeel PM, Baskar S, Dhulipala VS, Jaber MM. Cloud based framework for diagnosis of diabetes mellitus using K-means clustering. Health Inf Sci Syst. 2018;6(1):1–7.
18. Choi SB, Kim WJ, Yoo TK, Park JS, Chung JW, Lee YH, Kim DW. Screening for prediabetes using machine learning models. Comput Math Methods Med. 2014;2014:1.
19. Kaur H, Kumari V. Predictive modelling and analytics for diabetes using a machine learning approach. Appl Comput Inf. 2020;18:90.
20. Patil R, Tamane SC (2020) PSO-ANN-based computer-aided diagnosis and classification of diabetes. In: Smart Trends in Computing and Communications: Proceedings of SmartCom 2019, Springer Singapore, pp 11–20
21. Choubey DK, Kumar P, Tripathi S, Kumar S. Performance evaluation of classification methods with PCA and PSO for diabetes. Netw Model Anal Heal Inf Bioinf. 2020;9(1):5.
22. Hasan S, Shamsuddin SM. Multi-strategy learning and deep harmony memory improvisation for self-organizing neurons. Soft Comput. 2019;23(1):285–303.
23. Gregory JM, Slaughter JC, Duffus SH, Smith TJ, LeStourgeon LM, Jaser SS, Moore DJ. COVID-19 severity is tripled in the diabetes community: a prospective analysis of the pandemic's impact in type 1 and type 2 diabetes. Diabetes Care. 2021;44(2):526–32.
24. Graham EA, Deschenes SS, Khalil MN, Danna S, Filion KB, Schmitz N. Measures of depression and risk of type 2 diabetes: a systematic review and meta-analysis. J Affect Disord. 2020;265:224–32.
25. Redondo MJ, Hagopian WA, Oram R, Steck AK, Vehik K, Weedon M, Dabelea D. The clinical consequences of heterogeneity within and between different diabetes types. Diabetologia. 2020;63(10):2040–8.
26. Gómez-Peralta F, Abreu C, Cos X, Gómez-Huelgas R (2020) When does diabetes start? Early detection and intervention in type 2 diabetes mellitus. Revista Clínica Española (English Edition)
27. Middleton TL, Constantino MI, Molyneaux L, D'Souza M, Twigg SM, Wu T, Wong J. Young-onset type 2 diabetes and younger current age: increased susceptibility to retinopathy in contrast to other complications. Diabetic Med. 2020;37(6):991–9.
28. Alkayyali T, Qutranji L, Kaya E, Bakir A, Yilmaz Y. Clinical utility of non-invasive scores in assessing advanced hepatic fibrosis in patients with type 2 diabetes mellitus: a study in biopsy-proven non-alcoholic fatty liver disease. Acta Diabetologia. 2020;57(5):613–8.
29. Marinov M, Mosa ASM, Yoo I, Boren SA. Data mining technologies for diabetes: a systematic review. J Diabet Sci Technol. 2011;5:1549–56.
30. Anjali K. A review on the diagnosis of diabetes mellitus. Int J Digit Appl Contemp Res. 2015;4(1):1–7.
31. Verma P, Kaur I, Kaur J. Review of diabetes detection by machine learning and data mining. Int J Adv Res Ideas Innov Technol. 2016;2:1–5.
32. Yue C et al (2008) An intelligent diagnosis to type 2 diabetes based on QPSO algorithm and WLS-SVM. In: 2008 International Symposium on Intelligent Information Technology Application Workshops
33. Islam MF, et al. Likelihood prediction of diabetes at early stage using data mining techniques. In: Computer vision and machine intelligence in medical image analysis. Springer; 2020. p. 113–25.
34. Rony MAT, Satu MS, Whaiduzzaman M (2021) Mining significant features of diabetes through employing various classification methods. In: 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)
35. Mehrpour O, Saeedi F, Vohra V, Abdollahi J, Shirazi FM, Goss F. The role of decision tree and machine learning models for outcome prediction of bupropion exposure: a nationwide analysis of more than 14,000 patients in the United States. Basic Clin Pharmacol Toxicol. 2023. https://doi.org/10.1111/bcpt.13865.
36. Sarker IH. Machine learning: algorithms, real-world applications and research directions. SN Comput Sci. 2021;2(3):1–21.
37. Lalwani S, Sharma H, Satapathy SC, Deep K, Bansal JC. A survey on parallel particle swarm optimization algorithms. Arab J Sci Eng. 2019;44(4):2899–923.
38. Wang D, Tan D, Liu L. Particle swarm optimization algorithm: an overview. Soft Comput. 2018;22(2):387–408.
39. Le TM, Vo TM, Pham TN, Dao SVT. A novel wrapper–based feature selection for early diabetes prediction enhanced with a metaheuristic. IEEE Access. 2020;9:7869–84.
40. Kewat A, Srivastava PN, Kumhar D (2020) Performance evaluation of wrapper-based feature selection techniques for medical datasets. In: Advances in Computing and Intelligent Systems. Springer, Singapore, pp 619–633
41. Vanaja R, Mukherjee S (2018) Novel wrapper-based feature selection for efficient clinical decision support system. In: International Conference on Intelligent Information Technologies. Springer, Singapore, pp 113–129
42. Eberhart R, Kennedy J. Particle swarm optimization. Proc IEEE Int Confer Neural Netw. 1995;4:1942–8.
43. Remeseiro B, Bolon-Canedo V. A review of feature selection methods in medical applications. Comput Biol Med. 2019;112:103375.
44. Yu K, Guo X, Liu L, Li J, Wang H, Ling Z, Wu X. Causality-based feature selection: methods and evaluations. ACM Comput Surv (CSUR). 2020;53(5):1–36.
45. Wah YB, Ibrahim N, Hamid HA, Abdul-Rahman S, Fong S (2018) Feature selection methods: case of filter and wrapper approaches for maximising classification accuracy. Pertanika J Sci Technol, 26(1)
46. Song X, Waitman LR, Hu Y, Yu AS, Robins D, Liu M. Robust clinical marker identification for diabetic kidney disease with ensemble feature selection. J Am Med Inform Assoc. 2019;26(3):242–53.
47. Biswas S, Bordoloi M, Purkayastha B. Review on feature selection and classification using neuro-fuzzy approaches. Int J Appl Evolut Comput (IJAEC). 2016;7(4):28–44.
48. Koumi F, Aldasht M, Tamimi H (2019) Efficient feature selection using particle swarm optimization: a hybrid filters-wrapper approach. In: 2019 10th International Conference on Information and Communication Systems (ICICS), pp 122–127
49. Feature selection using PSO-SVM (2007) Int J Comput Sci
50. Abdollahi J (2020) A review of Deep learning methods in the study, prediction and management of COVID-19. In: 10th

International Conference on Innovation and Research in Engineering Science
51. Abdollahi J, Keshandehghan A, Gardaneh M, Panahi Y, Gardaneh M (2020) Accurate detection of breast cancer metastasis using a hybrid model of artificial intelligence algorithm. Arch Breast Cancer 22–28
52. Abdollahi J, Nouri-Moghaddam B, Ghazanfari M (2021) Deep neural network based ensemble learning algorithms for the healthcare system (diagnosis of chronic diseases). arXiv preprint arXiv:2103.08182
53. Abdollahi J, Nouri-Moghaddam B. A hybrid method for heart disease diagnosis utilizing feature selection based ensemble classifier model generation. Iran J Comput Sci. 2022;5:1–18.
54. Abdollahi J, Davari N, Panahi Y, Gardaneh M. Detection of metastatic breast cancer from whole-slide pathology images using an ensemble deep-learning method. Arch Breast Cancer. 2022. https://doi.org/10.32768/abc.202293364-376.
55. Abdollahi J (2022) Identification of medicinal plants in Ardabil using deep learning: identification of medicinal plants using deep learning. In: 2022 27th International Computer Conference, Computer Society of Iran (CSICC), pp 1–6
56. Abdollahi J, Mahmoudi L (2022) An artificial intelligence system for detecting the types of the epidemic from X-rays: artificial intelligence system for detecting the types of the epidemic from X-rays. In: 2022 27th International Computer Conference, Computer Society of Iran (CSICC), pp 1–6
57. Herliana A, Arifin T, Susanti S, Hikmah AB (2018) Feature selection of diabetic retinopathy disease using particle swarm optimization and neural network. In: 2018 6th International Conference on Cyber and IT Service Management (CITSM), pp 1–4
58. Li X, Zhang J, Safara F (2021) Improving the accuracy of diabetes diagnosis applications through a hybrid feature selection algorithm. Neural Process Lett 1–17
59. Daliri MR. Feature selection using binary particle swarm optimization and support vector machines for medical diagnosis. Biomedizinische Technik/Biomed Eng. 2012;57(5):395–402.
60. Soliman OS, AboElhamd E (2014) Classification of diabetes mellitus using modified particle swarm optimization and least squares support vector machine. arXiv preprint arXiv:1405.0549
61. Oladimeji OO, Oladimeji A, Oladimeji O. Classification models for likelihood prediction of diabetes at early stage using feature selection. Appl Comput Inf. 2021. https://doi.org/10.1108/ACI-01-2021-0022.
62. Kamel SR, Yaghoubzadeh R. Feature selection using grasshopper optimization algorithm in diagnosis of diabetes disease. Inf Med Unlock. 2021;26: 100707.
63. Chaki J, Ganesh ST, Cidham SK, Theertan SA. Machine learning and artificial intelligence based diabetes mellitus detection and self-management: a systematic review. J King Saud Univ Comput Inf Sci. 2020;32:1158.
64. Biswas R, Vasan A, Roy SS. Dilated deep neural network for segmentation of retinal blood vessels in fundus images. Iran J Sci Technol Trans Electr Eng. 2020;44(1):505–18.
65. Perveen S, Shahbaz M, Keshavjee K, Guergachi A. Prognostic modeling and prevention of diabetes using machine learning technique. Sci Rep. 2019;9(1):1–9.
66. Barik S, Mohanty S, Mohanty S, Singh D (2021) Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques. In: Intelligent and Cloud Computing, Springer, Singapore, pp 399–409
67. Hossain ME, Uddin S, Khan A. Network analytics and machine learning for predictive risk modeling of cardiovascular disease in patients with type 2 diabetes. Expert Syst Appl. 2021;164: 113918.
68. Tigga NP, Garg S (2021). Predicting type 2 diabetes using logistic regression. In: Proceedings of the Fourth International Conference on Microelectronics, Computing and Communication Systems, Springer, Singapore, pp 491–500
69. Emon MU, Keya MS, Kaiser MS, Tanha T, Zulfiker MS (2021) Primary stage of diabetes prediction using machine learning approaches. In: The 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), IEEE, pp 364–367
70. Joshi RD, Dhakal CK. Predicting type 2 diabetes using logistic regression and machine learning approaches. Int J Environ Res Public Health. 2021;18(14):7346.
71. Khaleel FA, Al-Bakry AM (2021) Diagnosis of diabetes using machine learning algorithms. Mater Today Proc
72. Li X, Zhang J, Safara F. Improving the accuracy of diabetes diagnosis applications through a hybrid feature selection algorithm. Neural Process Lett. 2023;55:153–69. https://doi.org/10.1007/s11063-021-10491-0.
73. Santhanam T, Padmavathi MS. Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. Procedia Comput Sci. 2015;47:76–83.
74. Kumar R, Kumar P, Tripathi R, Gupta GP, Islam AN, Shorfuzzaman M. Permissioned blockchain and deep-learning for secure and efficient data sharing in industrial healthcare systems. IEEE Trans Ind Inf. 2022;18:8065.
75. Kumar P, Kumar R, Gupta GP, Tripathi R, Srivastava G. P2tif: a blockchain and deep learning framework for privacy-preserved threat intelligence in industrial iot. IEEE Trans Ind Inf. 2022;18:6358.
76. Kumar P, Kumar R, Gupta GP, Tripathi R. BDEdge: blockchain and deep-learning for secure edge-envisioned green CAVs. IEEE Trans Green Commun Netw. 2022;6:1330.
77. Abdollahi J, Irani AJ, Nouri-Moghaddam B (2021) Modeling and forecasting Spread of COVID-19 epidemic in Iran until Sep 22, 2021, based on deep learning. arXiv preprint arXiv:2103.08178
78. Abdollahi J, Mahmoudi L Investigation of artificial intelligence in stock market prediction studies. In: 10th International Conference on Innovation and Research in Engineering Science
79. Amani F, Abdollahi J, Mohammadnia A, Amani P, Fattahzadeh-Ardalani G. Using stacking methods based genetic algorithm to predict the time between symptom onset and hospital arrival in stroke patients and its related factors. JBE. 2022;8(1):8–23.
80. Khavandi H, Moghadam BN, Abdollahi J, Branch A. Maximizing the impact on social networks using the combination of PSO and GA algorithms. Future Generat Distrib Syst. 2023;5:1–13.