SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V11I3P107

Volume 11 Issue 3, 63-76, March 2024

Original Article

Develop A Hybrid Improved Smooth Support Vector with A Modified Pigeon Search Optimization to Detect the Diabetes Mellitus at Early Stage

D. Arun[1], R. Annamalai Saravanan[2]

[1]*Department of Computer Science, Sankara College of Science and Commerce, Bharathiar University, Tamilnadu, India.*
[2]*Department of Computer Science, Hindusthan College of Arts and Science, Tamilnadu, India.*

[1]*Corresponding Author : arun.divakaran@hotmail.com*

Received: 06 January 2024 Revised: 10 February 2024 Accepted: 07 March 2024 Published: 31 March 2024

Abstract - *Diabetes Mellitus (DM) is a typical metabolic disease in which individuals struggle with high blood sugar (i.e., chronic hyperglycemia). It affects most of the body parts, such as the heart, kidneys, eyes, feet, skin, etc. Several competent Diabetes Diagnosis Systems (DDS) exploit different Machine Learning (ML) algorithms to gain valuable insights from the clinical datasets for DDS and disease management. However, trapping into local optimum solution, lack of privacy, missing values in the input dataset, and deficiency of incremental classification are major issues related to conventional ML-based diabetes classification algorithms. The primary objective of this study is to create an effective DM classification model that can reliably identify patient data as normal or diabetic. A hybrid Improved Smooth Support Vector Machine (ISSVM) with a Modified Pigeon Search Optimization (MPSO) algorithm was developed to detect the DM at an earlier stage. The empirical results reveal that the proposed classifiers outperform other related classifiers of existing systems regarding designated performance measures. These proposed algorithms can support clinicians in enabling the secured and accurate classification of DM with better accuracy and other performance measures.*

Keywords - *DM, Machine Learning, Improved Smooth Support Vector Machine, Modified Pigeon Search Optimization, Diabetes Diagnosis Systems, Performance measures.*

1. Introduction

DM is a long-term and non-communicable disease with relative and/or complete insulin shortage. The source of diabetes can differ significantly, but it always takes in the defect of the pancreas to discharge sufficient insulin, the body's cells do not respond to insulin, or both at some point in the disease period [1]. DM development is sturdily connected with the major disorders in the metabolism of glucose, carbohydrate, protein, and fat [2].

Besides, it is related to declined life expectancy, substantial morbidity owing to diabetes-oriented macrovascular problems (e.g., peripheral vascular disease, stroke, and heart disease), microvascular issues (e.g., nephropathy, neuropathy, and retinopathy), infections, and other consequences (e.g., COVID-19, nonalcoholic fatty liver disease, dental disease, dyslipidemia, and psychosocial problems), and reduced quality of life [3].The frequency of other illnesses that usually accompany diabetes, its distinctive features, and its widespread occurrence of DM makes it one of the current key public well-being and social problems [4]. DM can be classified into 4 categories as given below:

Type 1 DM (T1DM): Juvenile diabetes or T1DM distresses adults and children (i.e., 18 years old or below). It is primarily due to the scarcity of insulin secreted by the β-cells in the pancreas that do not produce sufficient insulin.

According to the National Diabetes Statics Report, T1DM is around 5% to 10% of all identified cases worldwide. Patients with T1DM require regular external insulin treatment to regulate plasma sugar levels. The development of T1DM is related to genetics and ecological aspects.

Type 2 DM (T2DM): It is non-insulin-dependent mellitus and accounts for 90% to 95% of DM cases found in adult individuals. T2DM is driven by insulin opposition, fat, and liver cells, which are scarce in consuming insulin effectively. The hazard of T2DM problems is associated with age, obesity, sloth, unhealthy diet, history of gestational diabetes, heredity, and devastation of glucose metabolism.

Pre-Diabetes (PD): It is a milder form of DM wherein patients have high plasma sugar. It is also called reduced glucose tolerance. Additionally, individuals with PD have a

high risk of having T2DM. A simple laboratory test can diagnose it.

Gestational DM (GDM): GDM is impaired glucose tolerance diagnosed with onset or first recognition during gestation that naturally resolves after delivery. The children of women who had GDM during pregnancy may be vulnerable, suffering from diabetes and being overweight. All the above-mentioned types of diabetes lead to distressing hitches if not managed effectively. There are several difficulties in the effective disease management of DM since economic and personal overheads are associated with DM treatment [5]. Its enduring significance is interpreted into human suffering and economic overheads.

On the other hand, inclusive diabetes care can reduce the progress of problems and healthcare outlays and maximize the quality of life. Hence, there is no qualm that DM demands inordinate attention. At the same time, ML-based DDS approaches can help individuals make a primary decision about DM based on their regular screening test results, and they can act as a reference for medical professionals [6].

Diabetes and heart disease are closely related. These days, diabetes is thought to be the primary risk factor for heart disease on its own. Controlling blood sugar is important for diabetes, while controlling blood pressure and cholesterol is important for heart disease. An attribute shared by both illnesses is insulin resistance [7]. It raises the risk of both heart disease and type 2 diabetes. Diabetes type 1 and type 2 each pose a separate risk for Coronary Heart Disease (CHD). Indeed, it may be reasonable to state that diabetes is a cardiovascular illness from the perspective of cardiovascular medicine [8].

According to the European Public Health Alliance, 41% of deaths are caused by circulatory illnesses, heart attacks, and strokes. Heart and circulatory system disorders are Australia's top cause of death, accounting for 33.7% of fatal cases, according to the Australian Bureau of Statistics. Researchers have been employing data mining techniques to assist medical practitioners in diagnosing heart disease, driven by the annual global increase in heart disease patient mortality and the abundance of patient data available to extract valuable knowledge [9].

As mentioned earlier, DM has become a vital concern in the medical domain, and it is important to determine appropriate solutions to combat the disease. It is important to intervene not only to treat but also to prevent and make an early disease diagnosis. The methods applied in ML, data mining, or any other domain of AI achieve prognostic modeling in DM detection, that is, the utilization of data and knowledge to calculate impending outcomes using previous data. These techniques are employed for the timely prediction and detection of DM to minimize morbidity rate [10].

The most general signs of DM include hyperglycemia, anomalous metabolism, and related risk for particular problems affecting the nervous system, kidneys, and eyes, which are the main parts of the human body. Signs are employed to collect clinical data, and the classification is carried out according to those symptoms [11].

Essentially, the process of diabetes detection is a data categorization problem. Data categorization is generally described as the method of labeling hidden data patterns. According to ML, Classification is the process of arranging a collection of data samples into appropriate groups according to the learning objective of a subset of the dataset whose correct class has been determined.

Using cutting-edge technologies, numerous academics and doctors have developed various DDS to identify DM [12]. As technology continues to progress, the development of such systems is becoming more effective. It is often a difficult undertaking to predict diabetes disease. These methods take advantage of technological developments that make machine learning possible [13].

Creating fresh implications from the early statistics, whether or not human interaction is required, presents a problem in developing such systems, which focuses on a distinct set of data patterns (i.e., the existing data). Based on each ML technique's classification performance and accuracy, the best one is usually chosen to create DDS [14].

There has been a colossal impact on the healthcare sector with the proliferation of cutting-edge technologies, including artificial intelligence, data mining, machine learning, the Internet of Medical Things, etc. These techniques have demonstrated their effectiveness in disease prediction from massive amounts of healthcare data [15]-exploiting information and statistics to calculate imminent results from past experiences.

Diabetes care could be made more successful by extending its reach through the application of machine learning techniques. In four key areas of diabetes care-medical decision support, prognostic population risk stratification, automatic retinal screening, and patient self-management tools-ML techniques are widely applied. It resulted in a change in the way data-driven healthcare was provided to control diabetes. It has revolutionized how diabetes is detected, treated, and avoided, potentially lowering the 8.8% global prevalence [16].

Efficient ML approaches have been employed to develop algorithms to support frameworks in diagnosing diabetes in the early stages of its subsequent complications. ML enables a constant and problem-free system for monitoring biomarkers and symptoms. Several ML approaches are proposed, including supervised, unsupervised, reinforcement, and

learning approaches. This is useful as ML methods rely on data [17].

ML could save substantial personal impact since a huge volume of disease-related information is fed into the diagnosis system. In the ML approach, methods are developed based on this data, and a more accurate output is delivered according to the input data. Some might analyze features of clinical scans, whereas others use blood test data collected from patients. The criteria vary according to the multiple indications of the disease. With several established techniques, Researchers have experimented with several techniques and adjusted numerous hyperparameters to obtain results that are most suitable for real-time use [18].

The SVM-based classification includes two phases including learning and testing. In the learning phase, this classification algorithm detects the predictive features by training data samples. During the testing process, the system classifies the newly arrived data sample according to those features and labels them to the equivalent category [19]. SVM is a non-parametric classifier that may use both linear and non-linear functions to address regression and classification issues [20]. This classification algorithm's fundamental idea is to divide the newly created testing dataset into appropriate classes based on the identified learning dataset.

1.1. Problem Statement

Accurate disease detection and efficient disease management are two important issues in the medical industry and positively impact an individual's health condition. Healthcare organizations are not exposed much to the advanced information and communication technological contexts. They have exploited the same clinical procedures for years. This frame of mind created a frozen culture, and it was very difficult to utilize new procedures in the medical sector. Large pandemic outbreaks like COVID-19 also impose new technical hitches for medical professionals, healthcare providers, investigators, and the public, as they generate a large volume of confidential data in the clinical decision support system [21].

Conventional approaches for data processing, investigation, storage, retrieval, and management are insufficient for the zettabytes of clinical data generated every year. Additionally, since clinical datasets become more assorted and intricate, cutting-edge data mining and computing methods are mandatory for useful data. Moreover, outcome reproducibility and data-communication strategies remain important deliberate problems. The majority of the medical systems, for instance, intensive care units, diagnosis divisions, case administration units, medical trials, etc., all have their tactics for collecting data. Until today, most of them utilize paper folders, paper charts, and faxes to exchange information [22].

As mentioned previously, DM has been a severe global health hazard for many decades. Numerous studies have observed the unabated growing number of people with diabetes worldwide. It is a non-communicable syndrome characterized by higher plasma sugar levels and related to complaints of fat, protein, and carbohydrate disorders. Moreover, it worsens almost all organs of the body. The applications used for diabetes management mandate continuous medical treatment to reduce the long-haul risk and prevent lethal complications [23].

1.2. Objectives

1. To explore existing DDSs using various ML techniques to more precisely classify individuals into normal and abnormal persons.
2. To select an appropriate classifier for DDS to classify the input samples with better classification accuracy.
3. To develop ISSVM with MPSO classifier to handle parameter optimization problems effectively and to evade algorithms trapping into sub-optimal value.
4. To show the superiority of the recommended approaches by relating their enactment with other existing techniques to evaluation metrics.

1.3. Research Gaps

Optimization techniques are capable of calculating the ideal initial variables; nevertheless, their results are not always optimal due to their tendency to find the best global solution instead of getting trapped in local optimum values. Furthermore, it takes longer to reach the ideal solution when bio-inspired optimization methods are used to find the ideal starting parameters, and frequently, iterative application of the method is required to achieve the desired parameters. According to this survey, the traditional diabetes classification models have the following drawbacks:

(i) An unbalanced class population and a small database.
(ii) Inaccurate classification results.
(iii) A lack of a suitable policy for initial point assignment and Artifact sensitivity.

The starting variables must have the best possible answers to maintain the model's correctness. Implementing novel bio-inspired optimization methods for a diabetes classification framework has received little attention.

2. Related Works

Several researches have been conducted to develop artificial intelligence-based computer-aided DDS [24]. Certain diseases have similar symptoms, the diagnostic system is frequently problematic, and optimization techniques are necessary to deal with these problems successfully. It is a data classification challenge to detect diabetes. The challenge of classifying a database into relevant groups to the learning outcome of a group of the dataset to the right category was

identified as categorization from the perspective of machine learning.

To produce smart DDS and enhance classification efficacy, metaheuristic optimization techniques are widely used [25]. The motivation for most metaheuristic techniques comes from physical, biological, or natural principles, and they attempt to emulate these at a fundamental level by applying various limitations. For each iteration, an arbitrary candidate solution is generated by every metaheuristic optimizer, which then repeatedly strives to attain the best outcome [26].

Researchers and doctors have recently become very interested in ML methods because of the availability of large databases merged with enhanced techniques and systems with higher processing capacities. ML is the technical field that deals with how computers learn from previous performance. The terms "machine learning" and "artificial intelligence" are sometimes used synonymously by researchers, although they are not synonymous because intelligence is defined as the capacity for learning.

Specifically, AI aims to develop a smart agent or assistant that utilizes different machine learning approach-based solutions. Machine learning aims to create computer methods that can recognize and develop from their previous statistics [27]. A more detailed definition of machine learning can be found in [28].

At the moment, ML techniques are successfully used for large, especially high-dimensional dataset regression, categorization, clustering, or dimensionality reduction procedures. Additionally, ML has proven to be exceptionally skilled in several fields, such as image processing, driverless vehicles, and online gaming. Therefore, effective machine-learning approaches power many aspects of our daily lives, including fraud detection, credit score reporting, image and speech recognition, web searches, email/spam filtering, etc. [29]. As per Sharma et al. (2021), machine learning techniques can be divided into three main categories:

1. Supervised or predictive learning, in which the system infers a function from a classified learning dataset.
2. Descriptive or unsupervised learning; in which the training stage tries to infer the format of an unknown.
3. Reinforcement learning, in which the system interacts with a dynamic environment to gather data shown in Figure 1 [30].

Diabetes is predicted using a variety of algorithms, such as K-Nearest Neighbours (KNN), Random forests, SVM, Naïve Bayes, Decision trees, and LR. A common diabetes dataset is used for the comparison analysis. Among the characteristics employed in the current dataset are the number of pregnancies, skin thickness, glucose, insulin, blood pressure, BMI, diabetes pedigree function, age, and outcome. Only female patients who are at least 21 years old can receive a diagnosis under the existing system. The effectiveness of existing systems varies; thus, it is important to confirm their correctness using relevant datasets before installation, which increases the time complexity of the deployment [31].

High-dimensional, complicated databases are effectively categorized using machine learning techniques. They make it possible for researchers and medical professionals to examine clinical information for unknown patterns to predict outcomes and reduce the costs associated with categorizing serious illnesses [32]. Real-world clinical databases with various features and external parameters are used to train machine-learning techniques. Many ML techniques are used in various research projects for the early diagnosis and categorization of diabetes. To diagnose diabetes, [33] proposed a DT-based categorization algorithm. On the other hand, a problem with sharp bounds befalls conventional DT algorithms. This study also proposes a fuzzy-based computational smart method for removing sharp edges to improve classification accuracy. It uses 336 experimental samples with a 75.8% classification accuracy assessed using MATLAB R2018b software [34].

The American Diabetes Association's recommendations were followed when constructing the dataset. 4900 data instances are utilized in the learning phase to train the classification technique and assess the results. The remaining 100 data instances are then put through testing. The empirical evaluation shows that because of the Fine k-NN's accuracy in classifying data objects, it is the best technique.

In the context of diabetes diagnosis, [35] examined the effectiveness of six ML algorithms, including J48, Hoeffding Tree, NB, Multilayer Perceptron (MLP), and RF. The World Health Organization (WHO) classification standards for diabetic patients are based on eight features that the authors have chosen to build the feature vector. In this paper, the classification efficacy of PIDD is analyzed using the Weka tool. All eight traits statistically eliminated the null hypothesis, suggesting that these attributes can efficiently differentiate between non-diabetic and diabetic individuals. Using the Hoeffding Tree technique, this work achieves a maximum of 77% accuracy and 77.5% recall [36].

To forecast diabetes patients, [37] used four classifiers: NB, DT, AdaBoost, and RF. Seven features are considered while evaluating the classifiers' performances on an actual dataset. Predictive accuracy is 94.25% when the RF-based classification method and LR-based attribute selection are combined. [38] Investigated how ML techniques affected DDS. The classification method, the database, the characteristics selection techniques with four possible attributes, and their implementation are the basis for a thorough study carried out by the writers. Compared to alternative methods, the authors demonstrated that SVM-

based classification models produced better classification outcomes.

SVM is a non-parametric classifier that uses both linear and non-linear cost functions to address regression and classification problems. According to [39], this discriminative classification method was developed to identify biological signal abnormalities because of its superiority and capacity to manage high-dimensional, non-linear databases in the medical field.

The categorization technique's primary idea is to use the training dataset to distinguish between unknown testing data and appropriate classifications. SVM can be used to categorize both linear and non-linear data in DDS. To vary learning datasets into a higher dimension, non-linear correlations are used. Following the modifications to the training data, the linearly optimal splitting hyperplane is looked for. The SVM categorization method uses accurate margins of different classes to speed up training and testing [40].

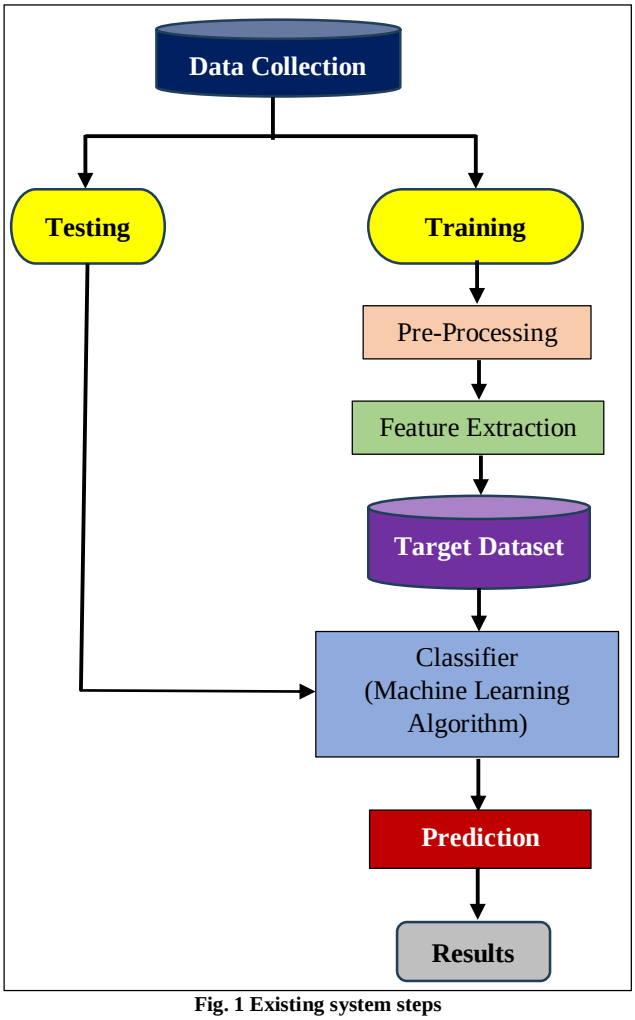


Fig. 1 Existing system steps

3. Proposed System

3.1. Proposed Architecture

Ten times stratified cross-validation was employed to examine the predicting effectiveness of the ISSVM-MPSO methods. Each iteration of the samples resulted in the division of the samples into ten subsets, each of which was tested once and nine of which were utilized to train the technique.

Furthermore, to provide a more accurate estimation of the actual model efficiency, the process was carried out three times with modifications to the sample combinations in these subgroups for a total of thirty executions. Three performance measures-specificity, accuracy, and sensitivity-compiled in the literature were employed to assess the outcomes. The percentage of positives (diabetes) correctly diagnosed is known as the sensitivity or true positive rate.

In contrast, the percentage of negatives (normoglycemic) accurately categorized was known as the true negative rate or specificity. The definition of accuracy is the total quantity of samples that are accurately recognized, taking into account true and false negatives. Figure 2 illustrates a proposed process for a fast, non-invasive, sustainable, portable, and non-invasive saliva-based diabetes testing platform.

3.2. Improved Smooth Support Vector Machine (ISSVM)

Diabetes disease identification through suitable analysis of the diabetes datasets is a significant classification problem. Different diabetes detection methods using artificial intelligence, especially ML methods, have been developed and enhanced using diabetes databases. This study aims to create an effective insulin diagnosis algorithm by utilizing the ISSVM classification algorithm.

Researchers have used ML algorithms to develop efficient DDS, which improve the enactment of the diabetes management system significantly. Numerous studies exploit the ISSVM classifier to diagnose diabetes. Even though ISSVM is widely used for discriminating the inherent attributes of various datasets for non-linear problems, its performance is hampered by the attributes of the designated variables.

ISSVM is a non-parametric approach that can use both linear and non-linear variables to address classification and regression issues. It creates a discriminating classification technique to identify biological sign anomalies because of its significant capacity to manage high-dimensional and non-linear databases used in the medical sector. The primary idea of the categorization method is to separate the hidden data into the appropriate classes according to the learning set of some renowned data. To solve binary categorization problems, ISSVM creates a hyperplane that optimizes and discriminates data instances into two categories.

$$w^t.i + b = 0 \quad (1)$$

A coefficient vector in Equation 1 is perpendicular to the hyperplane. The distance between the origin and the point in the database is denoted by the word b. The major goal of the ISSVM is to calculate the value of b and w. To create an optimal hyperplane, ||w||2 should be reduced under the constraint of $j_x(w^t.i + b = 0) \geq 1$ as given in Figure 3. Therefore, the optimization problem is modeled as:

$$minimizing\ \frac{1}{2}\|w\|^2 \quad (2)$$

$$\text{Subject to } j_x(w^t.i + b = 0) \geq 1, \quad x = 1,2,\ldots n \quad (3)$$

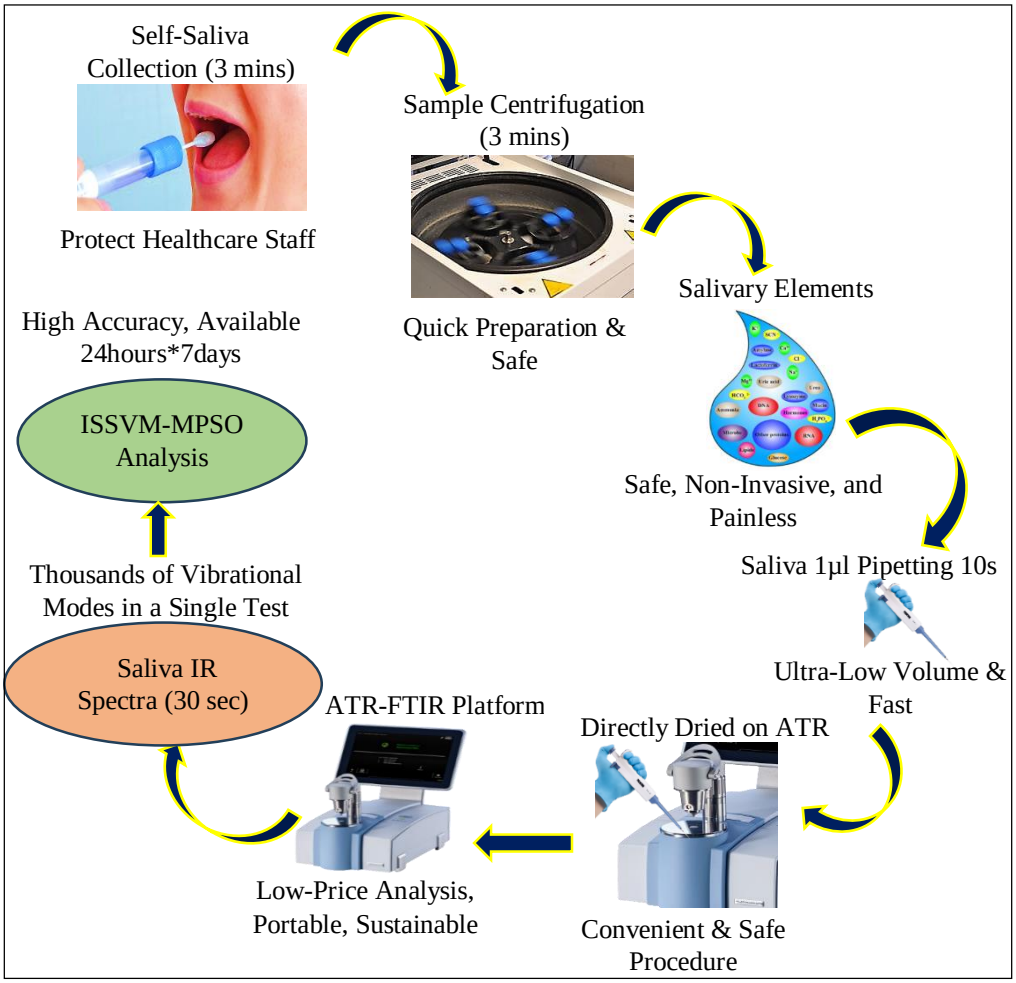


Fig. 2 Proposed architecture

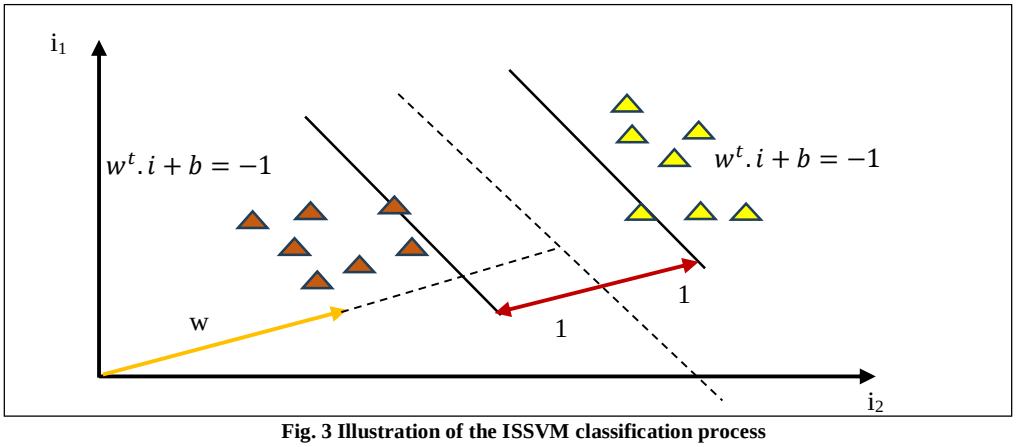


Fig. 3 Illustration of the ISSVM classification process

ISSVM uses Lagrange multipliers to solve the linear problem. In this algorithm, the support vectors are the data points laid on the judgment margin.

$$w = \sum_{x=1}^{n} \alpha_x y_x i_x \quad (4)$$

Where α_x signifies the language multipliers. Once w is calculated, the value of b can be computed using Equation 5.

$$j_x(w^t . i_x + b - 1) = 0 \quad (5)$$

The linear discriminating operation can be defined as given in Equation 6.

$$\hat{j} = sgn(\sum_{x=1}^{n} \alpha_x j_x i^T i_x + b) \quad (6)$$

ISSVM implements the kernel trick to solve a non-linear problem. Then, the decision function can be defined as Equation 7.

$$\hat{j} = sgn(\sum_{x=1}^{n} \alpha_x j_x k(i_x, i) + b) \quad (7)$$

Typically, any positive definite kernel operations, including the Gaussian function $k(i_x, i) = exp(-\gamma \|i - i_x\|^2)$, and the polynomial function $k(i_x, i) = (i^T i_x + 1)^d$ satisfy Mercer's limitation. This section only provides a short note on SVM. For more information, research provides a complete illustration of the ISSVM notions.

3.3. Modified Crow Search Optimization (MCSO)

MSCO is extensively employed in many technical applications, especially in the domain of optimization methods. Crows are the preeminent intelligent birds in the world. They have the biggest ratio of brain-to-body weight. They can identify people and interact with each other when an enemy comes. Similarly, they can use tactics, maturely share information, and remember their caching locations (caching positions) across seasons.

Crows are well-known for recognizing their caching locations, watching other birds, and thieving foodstuffs to the owner's leaves. After a crow was performed theft, it would give additional emplacements like changing their caching locations to avoid being an impending target. The crows utilize their knowledge of a thief to examine the activities of other birds and determine the optimal way to protect their food from theft. The basic assumptions of this CSA were (i) crows live in clusters, (ii) they recollect their caching locations, (iii) they watch other birds steal food, and (iv) They protect their caching locations from being larceny.

3.4. Proposed ISSVM with MCSO

A new ISSVM-MCSO-based DDS model is implemented in this work to solve the attribute extraction issue of DDS. The proposed ISSVM-MCSO method includes 2 stages. In the first phase, the optimal attribute selection process smears to the designated attributes to improve the predictive performance and decrease the number of selected features concurrently.

The ISSVM classifier was implemented in the second phase to realize higher classification accuracy by applying the optimal attribute subset. Attribute selection is a vital process for improving the performance of the classifiers, especially in high-dimensional datasets. It is a significant pre-processing method to remove redundant and unsuitable attributes. In this research, a principal component analysis-based dimensionality decrease approach is designated to convert the large size set of attributes into reduced ones, therefore resolving the relationship issue.

The dataset has been standardized in the area of [−1, +1] before executing the classification method. The k-fold Cross-Validation (CV) is employed to achieve a more accurate outcome assessment of the proposed framework. The intended study adopts k equal to 10. This denotes that the whole database is subdivided into 10 parts. The mean error across all 10 trials was computed. The virtue of the method was that all evaluated datasets are autonomous, and the reliability of the results can be enhanced. It is worth mentioning that repetition of the 10-fold CV would not generate acceptable solutions for validation due to the uncertainty of information. Consequently, all the outcomes are specified on a mean of 10 runs to achieve accuracy.

3.4.1. Algorithm 1: MCSO -ISSVM Algorithm

Input: Data on Diabetics,
Output: Prediction of disease; Compute the fitness function of crows; Compute the memory of crows
Step 1: Initialize the population of crows (n)
Step 2: Analyze the data k;
{
Step 3: for x ranges from 1 to k, do
{
Step 4: for y ranges from 1 to max_iteration, do //training and testing to perform the maximum iteration times of ISSVM
{
Step 4.1: Split(data, 0.4) = [training; testing] // data set split into 60% training and 40% testing
Step 4.2: Ground_Truth_Density = Classifier (X_{xy}, ISSVM parameter)
Step 4.3: Loss_function = Ground_Truth, Disease Density
Step 4.4: Calculate_Residual = Classifier(ISSVM)
Step 4.5: Update_Residual();
}
Step 5: Update Loss_Function();
Step 6: While (i< Imax)
{
Step 7: for each crow
{

Step 8: Define the value of PA
Step 9: Select the value of the rand
Step 9.1: if rand ≥ PA
{
Step 9.2: Update the position of the crows
}
Else
{
Step 9.3: Select the position of the agent randomly
}
}
Step 10: The likelihood of new positions
Step 11: Calculate the fitness of each crow
Step 12: Update the memory of the crows
Step 13: i=i+1
}
Step 14: return the value of the location of the crow
}}

In this algorithm, the performance of exploration and exploitation processes is typically controlled by PA. Lessening PA, this approach tries to explore local space. At the same time, the lesser the value of PA the exploitation is increased.

4. Results and Discussions

This work adopts Statistics and Machine Learning Toolbox in MATLAB R2018b software to simulate classification and optimization algorithms more precisely and to implement an effective diabetes identification model. MATLAB is a very dominant simulator widely used for modeling the behavior of the system to classify and predict DM. It is particularly used for formulating and solving mathematical models efficiently. MATLAB was introduced as a multi-paradigm proprietary programming language and computational intelligence platform. Users can build DDS models with higher classification performance within MATLAB through an integrated development environment. The user can design various user interfaces, perform matrix operations, plot data and functions, implement algorithms, etc.

The database does not provide a different set of data samples for evaluation and verification. As a result, purposefully divided the current dataset into 70% for testing and 70% for learning. Moreover, the complete data samples are divided into 10 parts (10% of each) because this research employed a 10-fold CV. Currently, 90% of the residual samples are divided for training, and one fold (10%) is applied for assessment. Every slide in the data will be tested exactly once, thanks to the use of a 10-fold CV.

In this work, the proposed ISSVM-MCSO model is implemented using MATLAB R2018b. There are 8 attributes used in this database (i.e., blood sugar level, Two-hour serum insulin, diastolic blood pressure, triceps skinfold thickness, the function of diabetes nutrition, the number of times pregnant, body mass index, and age).

Table 1 displays the statistical analysis of each data sample in this database. The area of binary parameters was confined to ‘1’ or ‘0’. The output parameter ‘0’ denotes a negative outcome (i.e., non-diabetic), and ‘1’ displays a positive outcome for DM (i.e., diabetic).

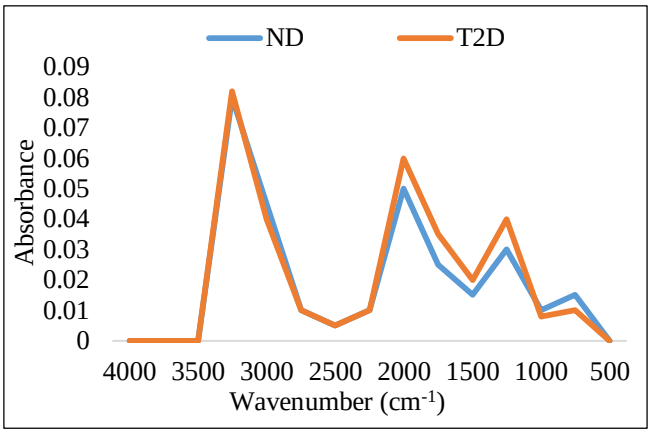
Table 1. Statistical analysis of PIDD

Index Value	Attribute	Mean	Maximum/Minimum
F1	Plasma Sugar Level	121.5	0/198
F2	Body Mass Index (kg/m2)	33	0/67.4
F3	Two-Hour Serum Insulin (mu U/ml)	78.2	0/844
F4	Age (Years)	34.1	22/82
F5	Diastolic Blood Pressure (mm Hg)	69.8	0/123
F6	Diabetes Nutrition Function	0.6	0.079/2.43
F7	Triceps Skinfold Thickness (mm)	21.2	0/98
F8	No. of Time Pregnant	3.6	0/18

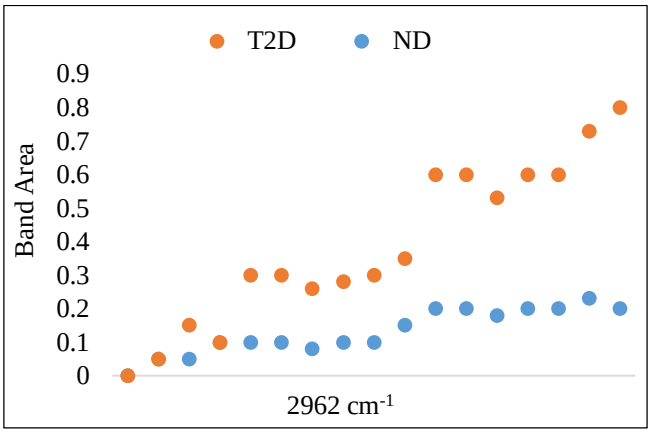
Table 2. Statistical analysis of DTD

Index	Feature	Mean	Maximum/Minimum
F1	A blood sugar test is taken at any time	11.55	7.5/14.2
F2	Age	34	22/82
F3	A blood sugar test is usually taken in the morning or 8 hours after a meal	6.26	3.8/9.3
F4	Class	0.1	
F5	Plasma glucose while fasting	12.75	0/55
F6	Type	Normal, Type I & II	
F7	Plasma glucose 90 mins after a meal	7.18	4.4/8.9
F8	HbA1c	44.22	29/65

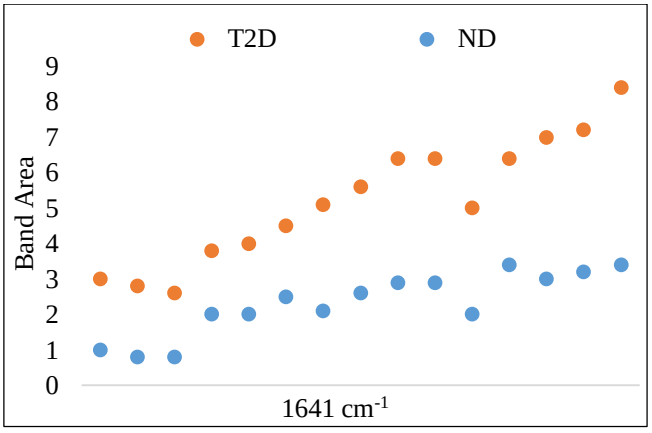
The clinical data samples DTD are gathered from the Data World repository. This dataset contains 1099 data samples with 8 features (e.g., plasma sugar test results generally taken in the morning or 8 hours after a meal, glycated hemoglobin (HbA1c), and blood glucose test arbitrarily taken at any time, plasma glucose while fasting, and plasma glucose 90 minutes after a meal, type, age, and class). Table 2 displays the statistical analysis of each sample of this database. Infrared salivary spectra were obtained for both non-diabetic participants and uncontrolled Type 2 diabetes patients. Figure 4(a) shows various lipids, carbohydrates, proteins, or nucleic acid components. Figures 4(b)-(g) determined several bands of importance. These bands include band region values of 1641 cm−1 (amide I), 1073 cm−1, and 2962 cm−1 (lipid CH3) (carbohydrates and glycosylated proteins), which are higher in type 2 diabetes patients with uncontrolled diabetes than in persons without the disease.



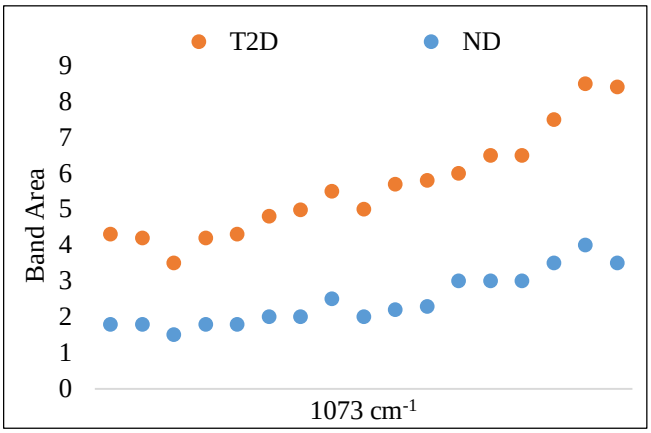
(a)



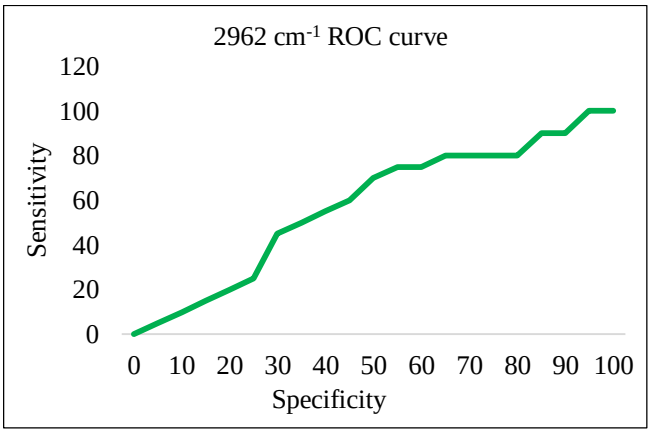
(b)



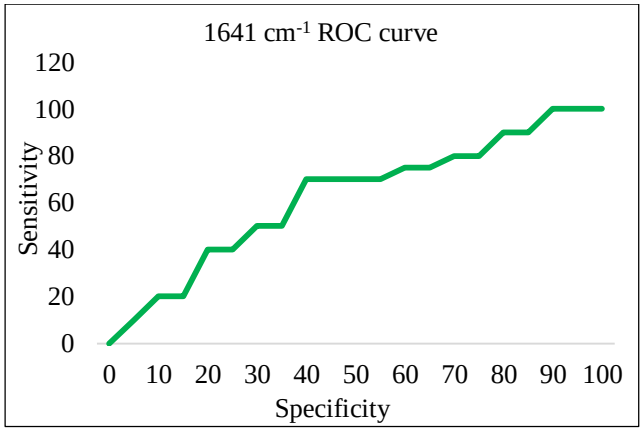
(c)



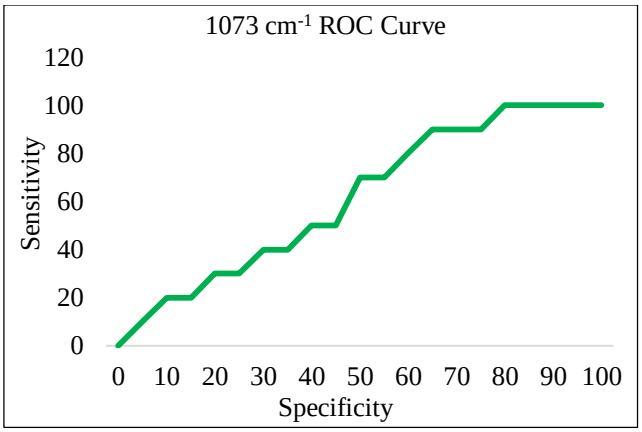
(d)



(e)



(f)



(g)

Fig. 4 Performance measures

The value of γ had an inordinate influence on the classification performance of the C owing to its impact on the enactment of the classification process. A minute value of γ causes under-fitting issues, while an excessive value leads to over-fitting problems (Pardo & Sberveglieri 2005). Both variables are selected in the range of {2−5, 2−4,..., 24, 25 }. This study adopts and C as 0.125 and 32, correspondingly (i.e., = 2 (0.125) and C =25 (32)).

The values of elements were designated using the error and trial method. Table 3 displays the details of the variable settings. The proposed optimization algorithm encounters perplexing hitches, including noise and outlier data, which can decrease the efficiency of the classification process. The outlier finding technique can be used in the re-processing step to find discrepancies in data; therefore, a good classification algorithm can be engendered for better decision-making.

Excluding the noise and outliers from the learning database will improve the predictive accuracy. In this work, ISSVM-MCSO is used to detect the outlier data in the given diabetes datasets. The objective of DBSCAN is to detect entities close to a given point to form dense areas, as shown in Table 4.

Table 3. Variables settings of experimentation

Algorithm	Parameters	Values
GWO	M	Linearly Decreased from 2-0
CSA	PA	0.12
	Flight Length	3
GA	Selection Mechanism	Route Wheel
	Cross over Ratio	0.3
	Mutation Ratio	0.12
PSO	Acceleration Constant	[3,3]
	Inertia	[0.8, 0.7]

Table 4. The result ISSVM-MCSO based outlier detection

Database	Instance (Org)	Instance (After Data Cleaning)	Outlier Data	Normal Data
PIDD	772	435	46	389
Diabetes Type	1098	1079	74	1008

Table 5. Results obtained by ISSVM-MCSO classifier on PIDD for different feature subsets

Fold	Selected Feature Subset	ACC	IoU	AUC	SEN	SPE	P-Value
1	{F1, F3, F5, F6}	0.918	0.937	0.925	0.892	0.952	0.035
2	{F2, F3, F5, F6, F8}	0.951	0.952	0.957	0.913	0.969	0.044
3	{F1, F2, F3, F7}	0.935	0.933	0.918	0.925	0.958	0.049
4	{F1, F2, F5, F6, F8}	0.947	0.931	0.925	0.965	0.957	0.040
5	{F2, F3, F8}	0.942	0.965	0.941	0.941	0.941	0.015
6	{F2, F2, F5}	0.942	0.928	0.897	0.872	0.952	0.073
7	{F1, F2, F6}	0.941	0.959	0.931	0.896	0.912	0.029
8	{F1, F2, F5, F7, F8}	0.935	0.966	0.925	0.907	0.958	0.091
9	{F1, F3, F5, F7}	0.922	0.930	0.898	0.872	0.947	0.024
10	{F2, F3, F5, F6, F7}	0.945	0.943	0.935	0.908	0.962	0.036
Mean	N/A	0.928	0.952	0.925	0.911	0.953	0.045
SD	N/A	0.008	0.015	0.019	0.022	0.015	0.023

The developed ISSVM-MCSO classifier is realized and its performance is evaluated under rigorous experimentation. Table 5 illustrates the complete results provided by the anticipated method. To improve the performance of the intended ISSVM-MCSO classifier, enacting the intended method is related to other prominent approaches found in the literature.

The basic SVM classifier is integrated with optimizers including CSA, PSO, GA, GWO, AGWO, EGWO, and ISSVM-MCSO, to facilitate improved classification performance. These approaches are executed and studied thoroughly using similar data. In the ISSVM-MCSO framework, PA managed the algorithm's diversification process. Therefore, this classification algorithm has produced rational solutions related to the basic SVM.

From Figure 5, it could be perceived that the standard deviation of the ISSVM-MCSO classifier is smaller than all other classifiers of the evaluation metrics. Consequently, the proposed system (ISSVM-MCSO) classifier displays more dependable solutions for detecting diabetes problems than the other classifiers.

More specifically, the ISSVM-MCSO classifier could not only improve the performance of the SVM classification algorithm but also allow better consequences for identifying DM. The comparative analysis carried out in this work determines that the proposed ISSVM-MCSO classifier offers an extremely effective method for DM identification. The mean value and standard deviation of evaluation metrics obtained from DTD by each classifier are demonstrated in Figures 5 and 6.

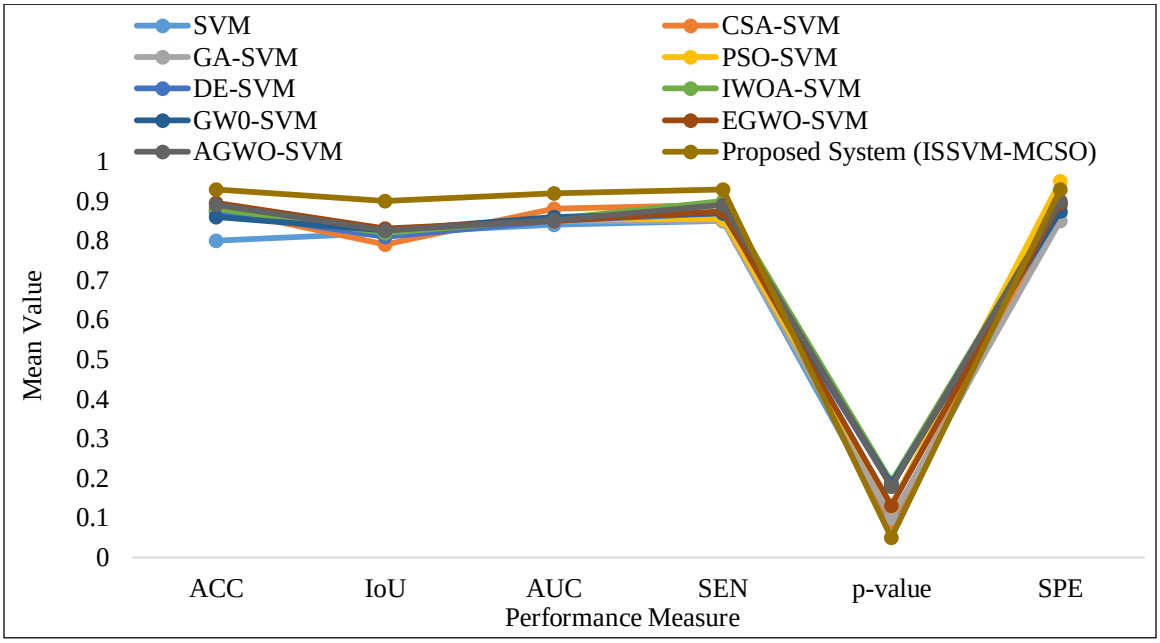


Fig. 5 Compares the DTD results according to the mean value

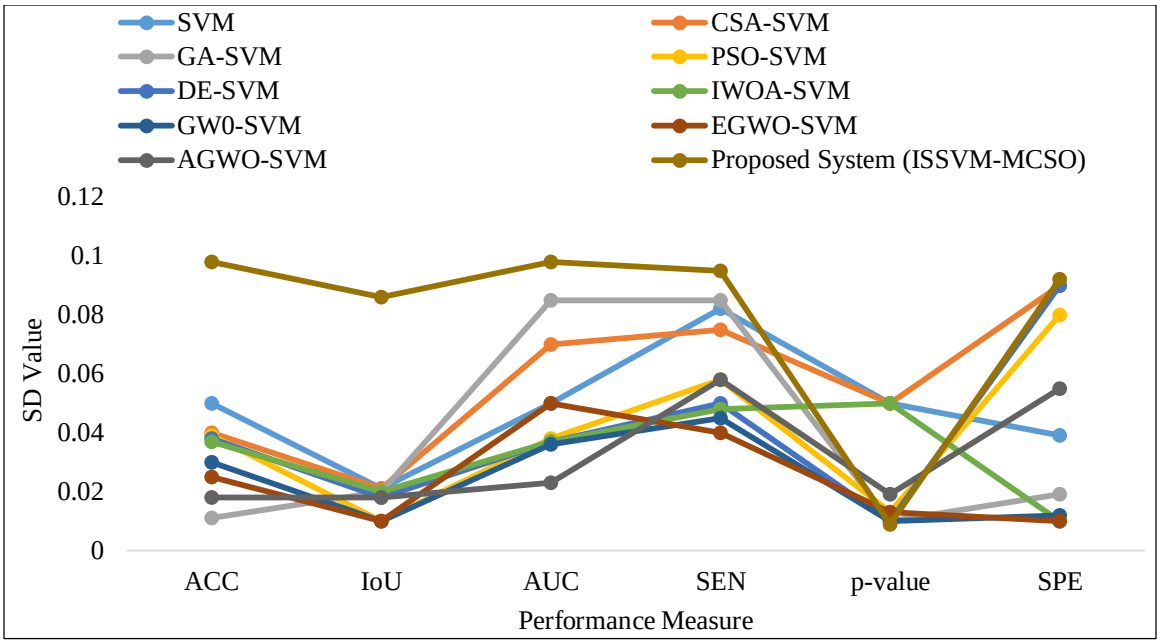


Fig. 6 Compares the DTD results according to the SD value

Specifically, a p-value was measured, and if it was smaller than 0.05 it describes that the solutions obtained to the proposed system ISSVM-MCSO classifier were significant deviancies from those of the other classifiers. The p-value> 0.05 signifies considerable differences among the fallouts of the proposed system ISSVM-MCSO classifier and the classifiers employed for assessment. Generally, one can simply recognize the p-values estimated by the Wilcoxon evaluation as being smaller than 0.05, which divulges the exceptional implementation of the proposed system ISSVM-MCSO algorithm.

5. Conclusion

Data classification in the DDS aids medical professionals and caregivers in studying unknown datasets by training a large volume of real-life databases. DM is a serious global health issue. Consequently, it is essential for the early detection, study, and control of hyperglycemia and its significance. This work develops an efficient model to identify DM employing a hybrid optimizer-based ISSVM classifier. The established optimization assimilates MCSO algorithms to utilize the entire capacity of ISSVM in the DDS. The proposed cohesive optimizer uses the strengths of the proposed system ISSVM-MCSO algorithms effectively to deliver favorable candidate approaches and achieve global optimum results successfully. In this study, an improved first population designated by CSA is employed to evaluate the positions of the exploratory person in the distinct exploratory space to gain optimal results with superior classification performance. The usefulness of this established cohesive optimizer-based ISSVM classifier is thoroughly assessed on the real-time datasets.

To determine the effectiveness of the proposed system ISSVM-MCSO algorithm, its enactment is compared with many state- ISSVM-based classifiers used to detect DM. The experimental research shows that the proposed ISSVM-MCSO can be considered an auspicious classification algorithm with exceptional performance metrics. In the first phase of the main experimentation, an ISSVM-MCSO-based classification model is implemented.

This system assimilates an ISSVM-MCSO to exploit the entire capacity of ISSVM to the DDS. The effectiveness of the proposed classifier was carefully examined using the real-world dataset, including PIDD and the DTD. To assess the performance of the recommended classifier, its enactment is contrasted to several state-of-the-art approaches using ISSVM in terms of classification IOU, sensitivity, accuracy, specificity, and AUC. The empirical outcomes reveal that ISSVM-MCSO could be a more effective classifier with exceptional classification performance to predict DM.

References

[1] Victor Chang et al., "Pima Indians Diabetes Mellitus Classification Based on Machine Learning (ML) Algorithms," *Neural Computing and Applications*, vol. 35, no. 22, pp. 16157-16173, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Salliah Shafi Bhat et al., "A Risk Assessment and Prediction Framework for Diabetes Mellitus Using Machine Learning Algorithms," *Healthcare Analytics*, vol. 4, pp. 1-12, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[3] Md. Jamal Uddin et al., "A Comparison of Machine Learning Techniques for the Detection of Type-2 Diabetes Mellitus: Experiences from Bangladesh," *Information*, vol. 14, no. 7, pp. 1-19, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[4] Amin Mansoori et al., "Prediction of Type 2 Diabetes Mellitus Using Hematological Factors Based on Machine Learning Approaches: A Cohort Study Analysis," *Scientific Reports*, vol. 13, pp. 1-11, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] Qiuhan Lu et al., "Longitudinal Metabolomics Integrated with Machine Learning Identifies Novel Biomarkers of Gestational Diabetes Mellitus," *Free Radical Biology and Medicine*, vol. 209, pp. 9-17, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Kirti Kangra, and Jaswinder Singh, "Comparative Analysis of Predictive Machine Learning Algorithms for Diabetes Mellitus," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 3, pp. 1728-1737, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[7] Chukwuebuka Joseph Ejiyi et al., "A Robust Predictive Diagnosis Model for Diabetes Mellitus Using Shapley-Incorporated Machine Learning Algorithms," *Healthcare Analytics*, vol. 3, pp. 1-11, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[8] Hind Alamro et al., "Type 2 Diabetes Mellitus and Its Comorbidity, Alzheimer's Disease: Identifying Critical microRNA Using Machine Learning," *Frontiers in Endocrinology*, vol. 13, pp. 1-13, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[9] Rashmi Ashtagi et al., "IoT-Based Hybrid Ensemble Machine Learning Model for Efficient Diabetes Mellitus Prediction," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 10s, pp. 714-726, 2023. [Google Scholar] [Publisher Link]

[10] Durga Parkhi et al., "Prediction of Postpartum Prediabetes by Machine Learning Methods in Women with Gestational Diabetes Mellitus," *Iscience*, vol. 26, no. 10, pp. 1-16, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] Ruth Sim et al., "Comparison of a Chronic Kidney Disease Predictive Model for Type 2 Diabetes Mellitus in Malaysia Using Cox Regression versus Machine Learning Approach," *Clinical Kidney Journal*, vol. 16, no. 3, pp. 549-559, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] Jun Zhang et al., "Machine Learning for Post-Acute Pancreatitis Diabetes Mellitus Prediction and Personalized Treatment Recommendations," *Scientific Reports*, vol. 13, pp. 1-10, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[13] Sven H. Loosen et al., "Prediction of New-Onset Diabetes Mellitus within 12 Months after Liver Transplantation-A Machine Learning Approach," *Journal of Clinical Medicine*, vol. 12, no. 14, pp. 1-12, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] Yan Li et al., "Machine Learning-Based Models to Predict One-Year Mortality among Chinese Older Patients with Coronary Artery Disease Combined with Impaired Glucose Tolerance or Diabetes Mellitus," *Cardiovascular Diabetology*, vol. 22, pp. 1-10, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] Mei-Yuan Liu et al., "Implementing A Novel Machine Learning System for Nutrition Education in Diabetes Mellitus Nutritional Clinic: Predicting 1-Year Blood Glucose Control," *Bioengineering*, vol. 10, no. 10, pp. 1-13, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[16] J. Jeba Sonia et al., "Machine-Learning-Based Diabetes Mellitus Risk Prediction Using Multi-Layer Neural Network No-Prop Algorithm," *Diagnostics*, vol. 13, no. 4, pp. 1-16, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[17] Kazuya Fujihara, and Hirohito Sone, "Machine Learning Approach to Drug Treatment Strategy for Diabetes Care," *Diabetes & Metabolism Journal*, vol. 47, no. 3, pp. 325-332, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[18] Chengyi Feng et al., "Machine Learning Models for Prediction of Invasion Klebsiella Pneumoniae Liver Abscess Syndrome in Diabetes Mellitus: A Singled Centered Retrospective Study," *BMC Infectious Diseases*, vol. 23, no. 1, pp. 1-12, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[19] Min Zhao et al., "A Machine Learning-Based Diagnosis Modeling of Type 2 Diabetes Mellitus with Environmental Metal Exposure," *Computer Methods and Programs in Biomedicine*, vol. 235, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] Andrya J. Durr et al., "Machine Learning for Spatial Stratification of Progressive Cardiovascular Dysfunction in a Murine Model of Type 2 Diabetes Mellitus," *Plos One*, vol. 18, no. 5, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[21] Ji-Yoon Lee, Doyeon Won, and Kiheon Lee, "Machine Learning-Based Identification and Related Features of Depression in Patients with Diabetes Mellitus Based on the Korea National Health and Nutrition Examination Survey: A Cross-Sectional Study," *Plos One*, vol. 18, no. 7, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[22] Khoula Al Sadi, and Wamadeva Balachandran, "Prediction Model of Type 2 Diabetes Mellitus for Oman Prediabetes Patients Using Artificial Neural Network and Six Machine Learning Classifiers," *Applied Sciences*, vol. 13, no. 4, pp. 1-22, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[23] Aidar Alimbayev et al., "Predicting 1-Year Mortality of Patients with Diabetes Mellitus in Kazakhstan Based on Administrative Health Data Using Machine Learning," *Scientific Reports*, vol. 13, pp. 1-12, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[24] Evangelos K. Oikonomou, and Rohan Khera, "Machine Learning in Precision Diabetes Care and Cardiovascular Risk Prediction," *Cardiovascular Diabetology*, vol. 22, pp. 1-16, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[25] Jinru Ding et al., "Machine Learning for the Prediction of Atherosclerotic Cardiovascular Disease during 3-Year Follow-Up in Chinese Type 2 Diabetes Mellitus Patients," *Journal of Diabetes Investigation*, vol. 14, no. 11, pp. 1289-1302, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[26] Sen Wang et al., "Identification of Ferroptosis-Related Genes in Type 2 Diabetes Mellitus Based on Machine Learning," *Immunity, Inflammation and Disease*, vol. 11, no. 10, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[27] Apoorva S. Chauhan et al., "Prediction of Diabetes Mellitus Progression Using Supervised Machine Learning," *Sensors*, vol. 23, no. 10, pp. 1-16, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[28] Yun-Nam Chan et al., "A Machine Learning Approach for Early Prediction of Gestational Diabetes Mellitus Using Elemental Contents in Fingernails," *Scientific Reports*, vol. 13, pp. 1-11, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[29] Xuelun Wu et al., "Development of Machine Learning Models for Predicting Osteoporosis in Patients with Type 2 Diabetes Mellitus-A Preliminary Study," *Diabetes, Metabolic Syndrome and Obesity*, vol. 16, pp. 1987-2003, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[30] Pooneh Khodabakhsh et al., "Prediction of In-Hospital Mortality Rate in COVID-19 Patients with Diabetes Mellitus Using Machine Learning Methods," *Journal of Diabetes & Metabolic Disorders*, vol. 22, pp. 1177-1190, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[31] Junli Zhang et al., "Prediction of the Risk of Bone Mineral Density Decrease in Type 2 Diabetes Mellitus Patients Based on Traditional Multivariate Logistic Regression and Machine Learning: A Preliminary Study," *Diabetes, Metabolic Syndrome and Obesity*, vol. 16, pp. 2885-2898, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[32] Yi-Ling Cheng et al., "Using Machine Learning for the Risk Factors Classification of Glycemic Control in Type 2 Diabetes Mellitus," *Healthcare*, vol. 11, no. 8, pp. 1-9, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[33] Xiaoqi Hu et al., "Prediction Model for Gestational Diabetes Mellitus Using the XG Boost Machine Learning Algorithm," *Frontiers in Endocrinology*, vol. 14, pp. 1-10, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[34] Qinpei Zhao et al., "Chinese Diabetes Datasets for Data-Driven Machine Learning," *Scientific Data*, vol. 10, pp. 1-8, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[35] Hong Pan et al., "A Risk Prediction Model for Type 2 Diabetes Mellitus Complicated with Retinopathy Based on Machine Learning and Its Application in Health Management," *Frontiers in Medicine*, vol. 10, pp. 1-15, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[36] Shubashini Rathina Velu, Vinayakumar Ravi, and Kayalvily Tabianan, "Machine Learning Implementation to Predict Type-2 Diabetes Mellitus Based on Lifestyle Behavior Patterns Using HBA1C Status," *Health and Technology*, vol. 13, pp. 437-447, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[37] Vansh Gupta et al., "Comparative Study of Machine Learning Models for Early Gestational Diabetes Mellitus," *2023 International Conference on Circuit Power and Computing Technologies (ICCPCT)*, pp. 1761-1766, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[38] Xinyu Liu et al., "Application of Machine Learning in Chinese Medicine Differentiation of Dampness-Heat Pattern in Patients with Type 2 Diabetes Mellitus," *Heliyon*, vol. 9, no. 2, pp. 1-11, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[39] Ayasha Malik et al., "Prognosis of Diabetes Mellitus Based on Machine Learning Algorithms," *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*, Delhi, India, pp. 1466-1472, 2023. [Google Scholar] [Publisher Link]

[40] Claudia C. Colmenares-Mejía et al., "Multivariable Prediction Model of Complications Derived from Diabetes Mellitus Using Machine Learning on Scarce Highly Unbalanced Data," *International Journal of Diabetes in Developing Countries*, pp. 1-11, 2023. [CrossRef] [Google Scholar] [Publisher Link]