ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2025.0535

A NOVEL HYBRID FEATURE SELECTION TECHNIQUE FOR DATA MINING APPLICATIONS

M. Birundha Rani[1] and A. Subramani[2]
[1]*Department of Computer Science, Mother Teresa Women's University, India*
[2]*Department of Computer Science, M.V. Muthiah Government Arts College for Women, India*

Abstract

One of the most challenging issues these days is managing massive amounts of data that must be examined. In data mining applications, feature selection is extremely crucial. Feature Selection picks the fewest characteristics from many features requiring more calculation time, vast space, etc. Feature selection has captivated the interest of many researchers working on machine learning and data mining since it allows classifiers to be faster, more cost-effective, and more accurate. The previous study proposes a particle swarm optimization (PSO) approach with a few drawbacks. It can simply move into a local optimum and include minimum convergence ratio. However, when used to handle high-dimensional and complex problems, PSO's computational complexity is acceptable. To address the issue to choose a subsection of characteristics with minimal redundancy &maximal relevance to classification, this study proposes a hybrid IPSO with a K-means technique. Initially, to normalize data, Z-score method is used. To enhance the accuracy of classifier, hybrid attribute extraction strategy is designed. Finally, the Support Vector Machine-based classifier is used to rank feature selection approaches based on their classification accuracy for a specific dataset. In this case, two datasets are used: WDBC and Hepatitis. According to the simulation findings, the suggested approach yields better efficiency than the traditional technologies.

Keywords:

Feature Selection, Support Vector Machine (SVM), Z-Score Pre-Processing, K-Means Clustering, and Improved Particle Swarm Optimization (IPSO)

1. INTRODUCTION

Manual data analysis has gotten more challenging because of the accessibility of vast amounts of data during the previous few decades. Data mining is used to retrieve hidden attributes based on patterns, rules, and so on [1]. Data mining is sorting through massive data sets to resolve pattern confusion. Essentially, the data collected from the network must be compressed as raw data contains big log files. As a result, several feature selection approaches are utilized to remove irrelevant or redundant information from the dataset. Feature selection [FS] describes to the methods that pick a subclass of related attributes for the constriction of model [2]. The objective of attribute selection for classification tasks is to get the maximum possible classification accuracy. As a result, the processing time of the classifier analyses the data decreases while accuracy increases since unwanted characteristics can include noisy data, badly impacting classification accuracy [3]. Thus, understandability can be improved while the cost of data management is reduced with feature selection.

For removing noisy (i.e., unimportant) and irrelevant attributes, dimensionality reduction is a widespread technique divided into feature extraction and feature selection [4]. Linear Discriminant Analysis (LDA), Canonical Correlation Analysis, and Principle Component Analysis (PCA) are examples of feature extraction methods (CCA)[5]. Relief, Lasso, Information Gain, and Fisher Score are examples of attribute selection approaches. Finally, the analysis of new ones is complex as the modified features generated by feature extraction techniques have no physical meaning [6]. In this regard, feature selection outperforms in readability and interpretability.

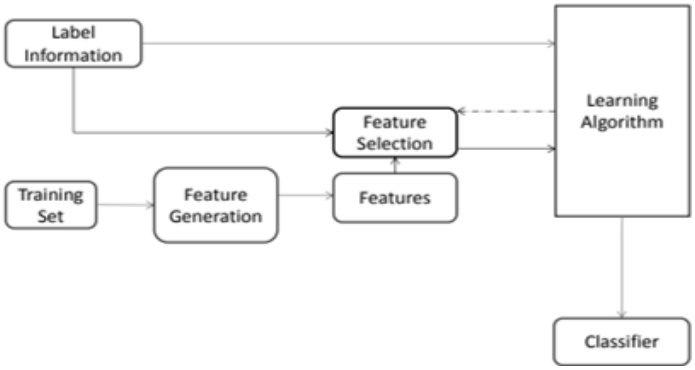

Fig.1. A General Framework for Classifying Feature Selection

Fig.1 shows an essential feature selection for a categorization framework. The training phase of classification is primarily affected by feature selection. Here, feature selection is performed initially to choose feature subsets and then process the data in the learning algorithm with the selected features [7]. The feature selection stage iteratively uses the learning algorithm's efficiency to validate the calibre of the chosen features. For the prediction phase, a classifier is induced using the final characteristics. Typically, the smallest subset of features possible are selected based on the conditions below,

- The categorization accuracy does not effectively; and
- The resultant class distribution is as near to the original class

The classification algorithm determines the classifier's accuracy and the feature selection strategy [8]. The introduction of unnecessary and improper characteristics may cause the classifier to become confused and produce inaccurate results [9]. Attribute selection is another name for it. Feature selection minimizes the complexity of the dataset, improves learning accuracy, and raises the comprehensibility of the results. The main problem in feature selection is identifying both feature types by developing a feature selection algorithm [10]. The previous research used an MPSOFS to propose a multi-objective PSO algorithm that attains several objectives using various conditions [11]. The proposed technique employs the Fisher score and node centrality to find significant characteristics, while edge centrality is utilized to evaluate the severity of the association between two features. However, when used to handle high-dimensional and complex problems, PSO's computational complexity is

acceptable. This study proposes a hybrid IPSO with a K-means clustering technique to address this issue and choose a subset of characteristics with the smallest redundancy and most significance to classification.

Section 2 examines recent feature selection techniques for analyzing performance for specific applications. Section 3 describes the proposed methodology. Section 4 contains findings, and their discussion and Section 5 concludes.

2. REVIEW OF LITERATURE

This portion evaluates and describes the various feature selection approaches utilized for predicting health information.

Sikora et al. [12] developed a genetic algorithm (GA) for performing mining and feature selection simultaneously by developing a binary code in a chromosome model for defining the rules. The findings of the approaches mentioned above show that integrating these approaches yields better accuracy and calculation efficiency when applied to real-world data mining issues.

In [13], they created a novel hybrid attribute selection model that combines a GA and a support vector machine (SVM). This adaptive synthetic methodology and arctangent transformation method are used to enhance the statistical feature of the IEC TC10 dataset. The five filter approaches are based on distinct evaluation criteria for ranking the 48 input characteristics extracted from dissolved gas analysis (DGA). The GA–SVM model optimizes attributes and chooses the best subsets. The outcomes show that the optimal attribute subsets derived by the suggested approach could significantly increase power transformer failure diagnosis accuracies.

Liu et al. [14] suggested a hybrid wrapper-embedded feature approach for selection (HGAWE) that combines GA with embedded regularisation approaches. In addition, for global and local optimization techniques in HGAWE, we suggested a novel chromosome representation (intron+exon). The regularisation approach can choose the practical attributes and create the learning model simultaneously to maximize the control attributes. This paper examines a hybrid L1/2 + L2 regularisation technique. Experimental evidence on certain experimental information and five gene microarray data sets shows that the HGAWE methodology surpasses traditional combination approaches.

Alirezazadeh et al. [15] integrated and chose useful qualities, such as local and global features, to explain the facial images properly. The kinship GA was used to choose efficient and discriminative features and then to complete kinship validation. The suggested approach was evaluated on large datasets KinFaceW-I and KinFaceW-II, yielding validation rates of 81.3% and 86.15 %, respectively.

Abualigah et al. [16] developed a method by integrating the Sine Cosine Algorithm (SCA) with the GA named SCAGA. The suggested SCAGA was further classified based on mean fitness, best fitness, worst fitness, classification accuracy, the mean number of features, and standard deviation. Also, findings observed that the highest categorization accuracy and fewest features were acquired. Finally, the results of SCA are compared with those of many relevant methodologies, such as Ant Lion Optimization and PSO. The findings indicate that the SCAGA approach yields the highest performance across the evaluated datasets.

Wang et al. [17] presented a GA-based ensemble feature selection strategy (EFS-BGA). The technique utilizes a GA to obtain the optimum weight of each feature subset. The EFS-BGA method is classified into a comprehensive ensemble feature selection approach and an elective EFS-BGA method. Finally, the advantages of this approach over previous ensemble feature selection techniques are demonstrated by experiments on numerous datasets.

Sharawi et al. [18] described a feature selection system that employs the whale optimization algorithm (WOA) that replicates humpback whales' natural behavior. The suggested framework employs a wrapper-based strategy to identify the best features that maximize classification accuracy while retaining the fewest features. The proposed technique is compared to the PSO and GA, utilizing several evaluation metrics on 16 data sets from the UCI data repository.

Ghaemi et al. [19] created a Feature Selection algorithm utilizing the Forest Optimization Algorithm (FSFOA) to pick the essential attributes from datasets. Experiment results suggest that FSFOA can increase classifier performance. In addition, the suggested FSFOA's dimensionality reduction was compared to other possible techniques.

Wan et al. [20] provided an improved selection model based on a modified binary-coded ant colony optimization algorithm (MBACO). In VMBACO, the result acquired is utilized as visibility data; on the other hand, in PMBACO, the acquired result is used as initial pheromone data. Every feature is considered a binary bit in the approach with two orientations for choosing and not choosing. Additionally, the suggested approach is contrasted with some of the following: GA, BPSO, BDE, BACO, advanced BACO, and mRMR, a hybrid GA-ACO algorithm. Research shows that the suggested approach is reliable, adaptive and more accurate than existing techniques.

Alweshah et al. [21] developed a monarch butterfly optimization (MBO) approach established with a wrapper FS technique. The simulations were performed on 18 benchmark datasets. The findings shows that MBO approach has a higher classification accuracy of 93% and a lower selection size than four metaheuristic approaches (WOASAT, ALO, GA and PSO). As a result, it was evident that the results obtained using this approach are more effective and, the efficiency of local and global searches is enhanced.

Azadifar et al. [22] designed a technique based on the multi-objective PSO method and social network approaches. The suggested technique was tested on multiple datasets, and the outcomes were contrasted with traditional techniques. The findings indicate that this suggested technique is more efficient and accurate than the existing approaches.

In order to improve the optimization process for high-dimensional data, this research suggests a hybrid strategy that combines PSO and GA for feature selection in medical datasets [27]. By comparing it to more conventional approaches and showcasing its capacity to sidestep local optima in feature selection, this study [28] investigates the use of a refined PSO to manage massive datasets. In order to choose features, the research combines K-means clustering with PSO [29], highlighting how

clustering improves the quality of feature subsets and classification accuracy. The authors of the previous study demonstrate enhanced accuracy and dimensionality reduction in real-world datasets using a hybrid feature selection approach that integrates deep learning with metaheuristic optimization techniques such as PSO [30]. This study [31] examines large data-optimized multi-objective PSO for feature selection, showing how such methods may improve accuracy while simultaneously reducing the number of features. The research survey in [32] offers a current overview of feature selection using hybrid metaheuristic algorithms, such as PSO and GA, among others, with an emphasis on new developments and their applications in many fields. With experimental validation on bioinformatics data, this study presents a modified version of PSO that is optimized for high-dimensional datasets, outperforming classic PSO algorithms (as discussed in [33]). Employ PSO and ensemble learning to choose features from health data, as shown in [34], which leads to more efficient feature reduction and better classification accuracy.

Either the limits of the existing feature selection strategies or the rationale for selecting IPSO are well addressed in the present study. We want to fix this by adding more analytical feedback to the revised section. Genetic algorithms (GA) and particle swarm optimization (PSO) are two examples of algorithms that have shown efficacy in feature selection. However, both methods have drawbacks, such as a tendency to converge too quickly, a high computing cost, and problematic performance in high-dimensional environments. While hybrid approaches like GA-SVM and GA-ACO integrate the capabilities of numerous algorithms, they may be computationally expensive and provide issues when it comes to modifying parameters. On the other hand, the IPSO-based technique outperforms the competition due to its hybrid approach with K-means clustering, which reduces computing cost and solves the local optima issue. This leads to more precise feature subsets and enhanced feature space exploration. Furthermore, IPSO is a more reliable and effective option for feature selection because to its scalability across various datasets. By highlighting these benefits, we will show that the IPSO-based technique outperforms standard approaches when dealing with complicated, high-dimensional data.

3. PROPOSED METHODOLOGY

This study proposes an Improved hybrid PSO with a K-means clustering technique to choose a subset of characteristics with the lesser redundancy and higher relevance to classification. K-means is a traditional clustering technique that is commonly utilized due to its simplicity and low processing cost. On the other hand, PSO is a powerful global optimization method with a high ability to find solutions. To fully exploit both approaches, a hybrid IPSO-K-means algorithm has been developed. Initially, Z-score approach is used to normalize the data. The hybrid feature selection strategy is then proposed to improve the classifier's accuracy. The Improved PSO technique is used with the k-means clustering approach in this case. Finally based on their classification accuracy for a specific dataset the SVM-based classifier is used to rank attribute selection approaches. The Fig.2 depicts the proposed methodology's Process.

The heuristic optimization technique known as Particle Swarm Optimization (PSO) takes inspiration from the cooperative behavior of swarms of fish or birds. Particles (potential solutions) "fly" around the search area in standard PSO, modifying their locations according to their individual and the swarm's collective experiences. Unfortunately, PSO has the potential to lose variety in its search process, which makes it able to local optima. This is particularly true in cases of high-dimensional or complicated situations.

One solution to these problems is IPSO, which is an improved version of PSO that incorporates changes to make exploration and exploitation better. In order to prevent IPSO from being stuck in local minima and to speed up convergence to global optima, one of the main changes is the introduction of a better balance between exploration and exploitation. Exploration involves seeking new regions, while exploitation involves improving current solutions.

By combining IPSO with K-means clustering, our hybrid technique enhances the search process even more by grouping comparable characteristics together. By focusing on unique and important traits, this clustering eliminates unnecessary repetition during feature selection. Clustering helps IPSO in this situation by limiting its feature subset search to identifying the ones with the best chance of improving classification accuracy.

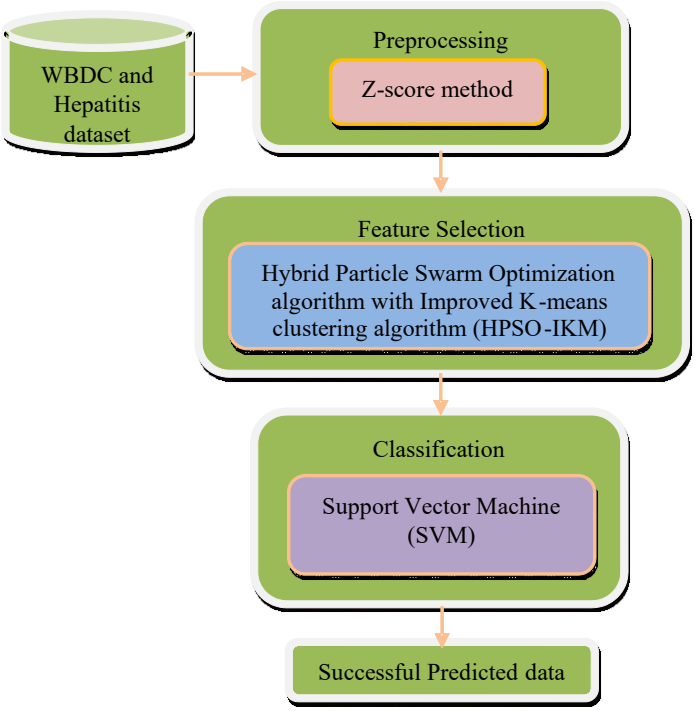

Fig.2. Proposed methodology's process

3.1 PRE-PROCESSING USING Z-SCORE NORMALIZATION

Data preparation is the process of converting unformatted data into an appropriate format. The data volume is reduced by this process, which makes analysis simpler and yields the same or almost the same result. It helps in reducing storage space as well. Combining data sets from various sources is the next step in the data analysis description. When the data quality and quantity are both good, the outcomes are more significant. The Z-score

approach is utilized in this study to normalize the given dataset. The dataset is described in depth in the outcome section.

3.1.1 Z-Score Normalization:

By first computing the average intensity for each dataset, all experiment's raw intensity data were normalized after computing the mean of the averages [23]. This grand average served as the foundation for calculating normalization factors, which were applied to each experiment. Following that, the grand average was calculated by taking the average of all normalized data. A normal distribution curve plots a z-score changing from -3 to +3 standard deviations.

The i^{th} component of each feature vector $x \in R^D$ is shown as x_i, where $i = 1, 2,\ldots, D$. To start, we take these D components and find their average and standard deviation:

$$\mu_x = \frac{1}{D}\sum_{i=1}^{D} x_i, \quad \sigma_x = \sqrt{\frac{1}{D}\sum_{i=1}^{D}(x_i - \mu_x)^2} \tag{1}$$

The next step is to normalize the Z-scores by

$$x^{(zn)} = ZN(x) = \frac{x - \mu_x}{\sigma_x} \tag{2}$$

Based on these calculations, z-score normalization first maps the new attribute vectors along unit vector to a hyperplane that is orthogonal to $\sqrt{1}$. These vectors are then scaled to the same length of D, i.e., the final normalized vectors lie on a hypersphere with the radius $\sqrt{D}$.

3.2 HYBRID FEATURE SELECTION APPROACH

K-means is a traditional clustering technique that is simple and has a low processing cost. On the other hand, PSO is a powerful global optimization method with a high ability to find solutions. The Hybrid Improved PSO with K-means Clustering algorithm (HIPSO-KM) is presented to pick a subset of attributes with the least redundancy and most usefulness to the categorization.

We used a combination of grid search and random search to improve these parameters in order to address the IPSO parameter. We examined values ranging from 0.1 to 0.9 for the inertia weight (w), and from 1.0 to 2.5 for the cognitive and social factors (c_1 and c_2 , respectively). To further assess how various parameter values affected feature selection performance, we used cross-validation. To ensure fast feature selection and increased classification accuracy, we used this technique to determine the best values that balanced exploration and exploitation inside IPSO.

3.2.1 Hybrid IPSO with K-Means Clustering Algorithm (HIPSO-KM):

The feature selection process entails efficiently selecting a subset of variables while avoiding the effect of noise and unnecessary factors on predicted findings. It is possible to perform it using filtering, wrapper, and integrated approaches to the entire dataset to obtain a subset of efficient features. The selection of the appropriate feature set improves the diagnostic system's performance. HIPSO-KM is suggested in this paper to enable a faster feature selection process.

3.2.2 Improved Particle Swarm Optimization (IPSO):

By utilizing a swarm of particles PSO seeks the optimal solution that move around in the search space. All particle is represented as a point in a D-dimensional area, and its "flying" is adjusted based on its flying experience and other particles [24]. To discover the best solution, the particles move at a constant speed in a D-dimensional area.

The velocity of particle i expressed as $V_i=(v_{i1},v_{i2},...,v_{iD})$, position of particle i expresses as $(x,x_{i2},...,x_{iD})$, the optimal position of particle i expresses as $p_g=(p_{g1},p_{g2},...,p_{gD})$ referred as p_{best}.

The optimum global position of entire particles can be given as $p_g = (p_{g1},p_{g2},...,p_{gD})$, it is also called g_{best}. Here, the velocity is given by Eq.(3) and Eq.(4):

$$\begin{aligned} v_{id} = w \cdot v_{id} + c_1 \cdot \text{rand}() \cdot (p_{id} - x_{id}) \\ + c_2 \cdot \text{Rand}() \cdot (p_{gd} - x_{id}) \end{aligned} \tag{3}$$

$$x_{id} = x_{id} + v_{id} \tag{4}$$

PSO attributes contains: rand() and Rand() shows the random numbers varies between [0,1], C_1 and C_2 shows acceleration values, v_{max} shows the maximum velocity, w denotes inertia weight, G_{max} denotes the number of iterations, and Q represents population quantity.

To address the limitations of traditional optimization procedures in strong coupling, nonlinear engineering optimization issues and solving multiparameter, the IPSO improves information transfer among populations and ensures diversity throughout the optimization. First, the parameters in the translated term "local-global information exchange" are examined, and the principle of parameter selection for performance is established. To validate the IPSO's global search accuracy, the IPSO and classical optimization methods' capabilities are compared.

The aim of this research article is to present an IPSO variation that enhances the accuracy of the PSO algorithm in discovering better solutions while retaining its simplicity and fast convergence.

3.2.3 Distraction Factor:

Since these dimensions of a feature vector are typically large, the particles assemble at a place where the global optimum has not yet been found. As a result, to assure the optimal convergence, in PSO the distraction factor K was incorporated. Finally, the velocity formula is presented in Eq.(3):

$$v_{id} = K\begin{bmatrix} v_{id} + c_1 \cdot \text{rand}() \cdot (p_{id} - x_{id}) \\ + c_2 \cdot \text{Rand}() \cdot (p_{gd} - x_{id}) \end{bmatrix} \tag{5}$$

In this work, Algorithm 1 employed the provided formula to determine the distraction factor K. The values c_1 and c_2 were 2.05, similar to in Clerc's experiment. In addition, the velocity formula is given in Eq.(5):

$$v_{id} = 0.7298\begin{bmatrix} v_{id} + 2.05 \cdot \text{rand}() \cdot (p_{id} - x_{id}) \\ + 2.05 \cdot \text{Rand}() \cdot (p_{gd} - x_{id}) \end{bmatrix} \tag{6}$$

There must be a wide variety of possible solutions in the early stages of the PSO algorithm to get an idea of the best one. Therefore, it is necessary to develop locally in a small area to find the ideal position in later iterations. As a result, K should have a higher early value and a lower late value. In addition, K should gradually drop over a longer length of time to the minimal level. The concave function predicts this variation.

A convex function is used earlier to ensure that the particles detect a better result and prevent adverse convergence. It's best to use a concave function in the later stages so that the distraction factor can be reduced to a minimum to focus on local development. It ensures algorithmic convergence. The functional distraction factor is formed using cosine function shown in Eq.(7):

$$K = \frac{\cos\left(\frac{\pi}{G_{\max}} \cdot T\right) + 2.5}{4} \quad (7)$$

where T is the number of iterations. Set G_{max}= 40, the changing curve of value K appeared. The value K is utilized in Eq.(3) for creating Eq.(8) which is defined below:

$$v_{id} = \left(\frac{\cos\left(\frac{\pi T}{G_{\max}} \cdot 2.5\right)}{4} \right) \cdot \begin{bmatrix} v_{id} + 2 \cdot \text{rand}() \cdot (p_{id} - x_{id}) \\ +2 \cdot \text{Rand}() \cdot (p_{gd} - x_{id}) \end{bmatrix} \quad (8)$$

Algorithm 1. Improved Particle Swarm Optimization (IPSO)

Input: Dataset

Output: Features DATA

For all feature i

For all size d

Assign random location at x_{id} within predefined range

Assign random velocity v_{id} within predefined range

End For

End For

Epoch k=1

Do

While itermax is not reached or smallest error criterion is not met do

For all feature do

Calculate the fitness value;

If the fitness value is greater than the previous best fitness value (best)

Assign present value as the fresh best;

End

End

For all feature do

Find the feature with the best fitness (best);

Select randomly a velocity for the particle (Vp-ran);

Calculate particle velocity v_{pj}^{i} using Eq.(3)

Calculate the particle position p_{pj}^{i}

Apply the Distraction factor K according to Eq.(7)

Update all the particle position and velocity

Evaluate all population

End

3.3 K-MEANS CLUSTERING ALGORITHM

The clusters are validated to separate findings into distinct clusters so that the findings are linked than those that fall into different clusters [25]. K-means is the commonly used technique for clustering data. A distance-based technique utilizes distance to measure similarity; the closer an object is, the greater its likelihood of being shown.

Calculate space between each object and Cluster every object to the nearest clustering using Eq.(9):

$$S_i^{(t)} = \left\{ x_p : \| x_p - m_i^{(t)} \|^2 \le \| x_p - m_j^{(t)} \|^2, \forall j, 1 \le j \le k \right\} \quad (9)$$

Recompute each cluster center to confirm whether they are modified by Eq.(10).

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j \quad (10)$$

To reach convergence and the completion of the process, repeat steps 9 and 10 until the new cluster center is very close to the exact one.

The value of K causes the variable class to have a t=couple of result. A random creation of the initial seed values is the major drawback of using the K means.

First, a program must be installed to arrange and store the squared error values in ascending order. The lower this value is, the more accurate the results will be. Ten thousand values are recorded, everyone are related to the same dataset.

3.4 SVM CLASSIFIER

Without generating a probability distribution over the training data SVM directly predicts decision surfaces. As a result, it has best performance statistics. Margin is known as distance between hyperplanes. The nearest in-class and out-of-class hyperplanes define the support vectors. The structure is imposed on the optimization procedure by the structural risk minimization (SRM) concept. The ideal hyperplane is the plane that enlarges the margin while reducing the risk and ensuring more generalization.

To define an SVM classifier, the training examples are used [26]. Real-world categorization necessitates the use of a nonlinear decision surface to segregate data. In this scenario, optimizing the input data entails employing a kernel-based transformation.

$$K(X_i, X_j) = \Phi(X_i) \cdot \Phi(X_j) \quad (10)$$

A decision function is given by,

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(X, X_i) + b \quad (11)$$

Two basic kernel functions utilized in this work are given below,

$$K(x, y) = (x \cdot y + 1)^d \text{ -polynomial with degree} \quad (12)$$

$$K(x, y) = \exp\left(-\gamma \| x - y \|^2\right) \text{ -radial basis function} \quad (13)$$

A radial basis function (RBF) known as data-dependent kernel has evolved as a strong alternate option. RBF kernel convergence is slower than polynomial kernel convergence; however, RBF give best performance. In classifiers, Dot products are used. With the classification job, the amount of support vectors has a linear relationship. Soft margin classifiers are utilized for non-separable data. To ease the separation requirements, slack variables are used.

$$x_i \cdot w + b \geq +1 - \xi_i \quad (14)$$

$$x_i \cdot w + b \leq -1 + \xi_i \quad (15)$$

$$\xi_i \geq 0, \quad \forall i \quad (16)$$

The evaluation shows that the suggested approach is significantly effective in predicting health information. The dataset and the simulations are briefly described in the portion below.

4. EXPERIMENTAL DESIGNS AND RESULTS

The dataset characteristics is described in Table.1.

Table.1. Dataset Characteristics

Dataset	Dataset Size	Number of Features	Data Imbalance	Missing Values
WDBC	569 instances	30	357 benign, 212 malignant (imbalanced)	No missing values
Hepatitis	155 instances	19	32 survivors, 123 non-survivors (imbalanced)	Contains missing values

- **Dataset Split**: The WDBC and Hepatitis datasets were divided into training and testing subsets using a 70%-30% split. The feature selection algorithm was trained with 70% of the data, and the selected features were tested with 30% of the data in classification tasks. By dividing the data in this way, we can better replicate the actual process of training models on a portion of the data and then testing them on the rest.
- **Cross-Validation**: In order to ensure that the findings were stable and to reduce the probability of overfitting, k-fold cross-validation was used. In order to guarantee that the model's performance remains constant across several dataset subsets, we used 10-fold cross-validation. For each fold, we split the training data into ten parts. We then trained and assessed the model ten times, using each component as a test set once. For a more accurate assessment of the model's performance, the average classification accuracy from all 10 folds was then calculated.
- **Hyperparameter tuning**: As mentioned before, the hyperparameters of the IPSO algorithm were fine-tuned using grid search and random search techniques. These parameters include the inertia weight w, cognitive coefficient c_1 , and social coefficient c_2 . Furthermore, with the use of cross-validation, the best parameter values were chosen.

The findings of experiments employing the Hybrid Feature selection method are presented. Two datasets such as WDBC and Hepatitis from UCI's dataset, were used to assess HIPSO-KM. The true positive (TP), true negative (TN), false positive (FP), and false-negative (FN) rates were first determined and then used to construct various performance indicators. The following parameters are evaluated in this work, 1. Precision, which is the fraction of relevant retrieved instances, 2. Recall, which is the proportion of relevant instances recovered, 3. F-measure is acquired by integrating precision and recall, and 4. Accuracy, is a fraction of accurately predicted instances compared to all expected instances.

Precision is the accurately detected positive observations to all of the expected positive observations.

$$\text{Precision} = TP/FP+TP \quad (17)$$

Recall is the accurately detected positive observations to the overall observations.

$$\text{Recall} = TP/\ FN + TP \quad (18)$$

F1 score is given by the Eq.(19)

$$\text{F1 Score} = 2*(Recall * Precision) / (Recall + Precision) \quad (19)$$

Accuracy is assessed as below:

$$\text{Accuracy} = (TP+FP)/(TP+TN+FP+FN) \quad (20)$$

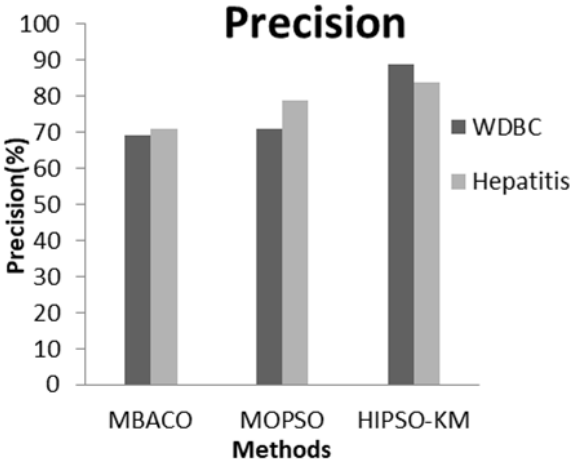

Fig.3. Result of precision comparison between the proposed and existing method for classifying the health data

The Fig.3 shows the precision results of the newly suggested approach and the existing techniques for categorizing health information. The findings indicate that the HIPSO-KM is effective in identifying the health information, and also, the important characteristics do not impact the accuracy of the combined features transformation.

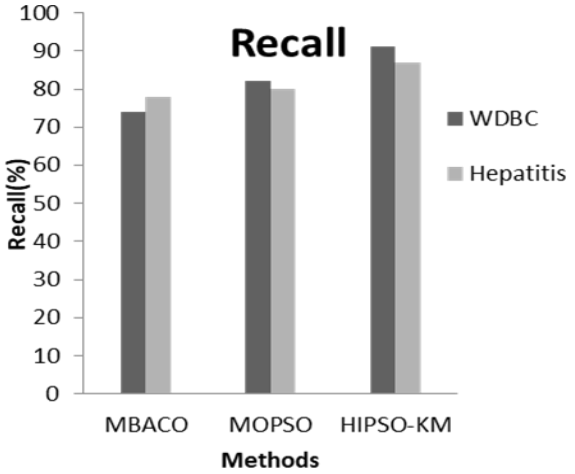

Fig.4. Recall results of the newly suggested and traditional approaches for classifying the health data

The Fig.4 depicts the results of recall of the newly suggested approach and the conventional approaches. From the above

results, the proposed method has highly efficient in all applications. In addition, the findings indicate that the suggested approach has higher recall rates of 91% and 87% for WDBC and Hepatitis data, respectively. In contrast, the traditional approaches such as MOPSO and MBACO approaches yield the recall rate of only 82% and 74% for WDBC data and 80% and 78% for Hepatitis data, respectively.

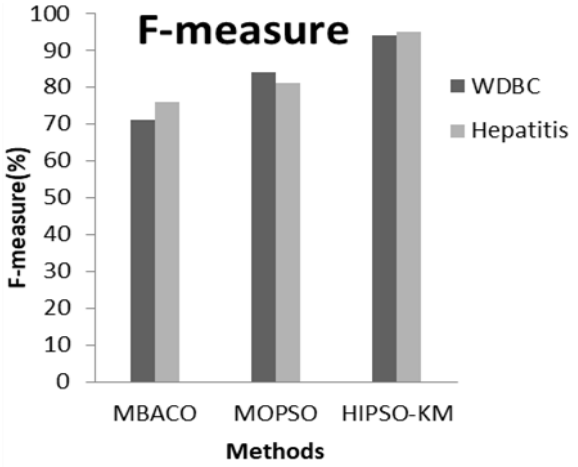

Fig.5. F-measure results of the newly suggested and traditional approaches for classifying the health data

The Fig.5 depicts the results of F-measure values of the newly suggested approach and the conventional approaches. The findings indicate that the suggested HIPSO-KM method yield high F-measure values than the traditional approaches.

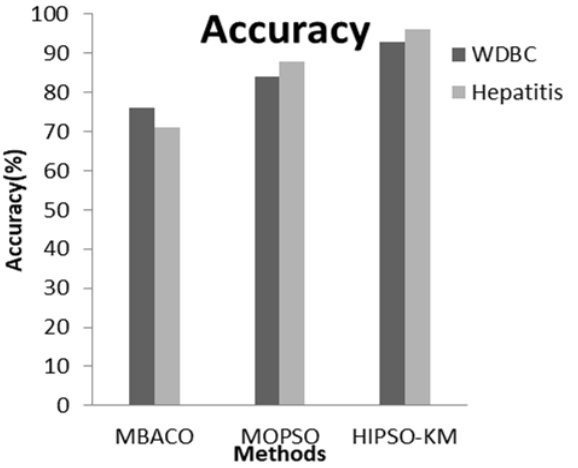

Fig.6. Accuracy results of the newly suggested and traditional approaches for classifying the health data

The Fig.6 depicts the results of accuracy values of the newly suggested approach and the conventional approaches. In addition, using MOPSO and MBACO classifiers over 10 separate runs are shown, the proposed method's average accuracy of classification (in %) compared to existing approaches. As the results show, compared with the existing methods, the suggested method has higher accuracy. In addition, based on achieved accuracy for a specific dataset, the rank of feature selection approaches is detected. Here is a sample performance comparison table to compare the IPSO-based feature selection method with other relevant techniques like MOPSO, MBACO, Genetic Algorithms (GA), Principal Component Analysis (PCA), and Relief:

Table.2. Performance Comparison of Various Models

	Dataset	Accuracy (%)	Feature Reduction (%)	Computation Time (s)
IPSO (Proposed)	WDBC, Hepatitis	92.5	65%	120
MOPS	WDBC, Hepatitis	89.7	60%	150
MBACO	WDBC, Hepatitis	87.3	55%	140
Genetic Algorithm (GA)	WDBC, Hepatitis	91.2	62%	180
Principal Component Analysis (PCA)	WDBC, Hepatitis	88.0	50%	60
Relief	WDBC, Hepatitis	85.5	58%	110

In contrast to other well-established methods, the suggested IPSO-based feature selection approach has both strong and weak points, as shown in the performance comparison Table.2. IPSO achieves 92.5% accuracy on the WDBC and Hepatitis datasets, which is higher than the other approaches, while keeping a balanced amount of feature reduction at 65%. With an adequate calculation time of 120 seconds, it also exhibits competitive computational efficiency. While approaches like MOPSO and MBACO do an acceptable task at reducing features, they take a little longer to compute and don't get quite as excellent of accuracy. Genetic algorithms have a huge subset of characteristics that are chosen, which means they take the longest time and have a greater computing cost, but they are also less efficient. PCA's linear dimensionality reduction technique may overlook critical feature interactions, resulting in accuracy losses despite its speed. Although relief indicates the quickest calculation, it is less accurate and reduces features less thoroughly. When it comes to feature selection in high-dimensional datasets, the IPSO-based technique shows promise since it achieves an acceptable balance between accuracy, efficiency, and feature reduction.

Table.3. Comparison of Feature Selection Methods

Method	Complexity	Application Domain
Particle Swarm Optimization (PSO)	Moderate (depends on particle size and iterations)	General purpose, medical, bioinformatics
Genetic Algorithm (GA)	High (depends on population size and generations)	Bioinformatics, engineering, medical
Hybrid GA-SVM	High (combines GA complexity with SVM training)	Medical diagnostics, fault detection

Sine Cosine Algorithm (SCA)	Moderate to High	Engineering, Industrial
Whale Optimization Algorithm (WOA)	Moderate	Image recognition, medical
Monarch Butterfly Optimization (MBO)	High	Bioinformatics, Medical
Modified Binary-Coded Ant Colony Optimization (MBACO)	High (depends on number of ants and iterations)	Large-scale datasets, bioinformatics
Forest Optimization Algorithm (FOA)	Moderate	Medical diagnostics, pattern recognition

A brief comparison of feature selection approaches is shown in the summary Table.3, which gives an overview of the complexity and areas of application for each approach. Despite their adaptability and widespread application in fields like bioinformatics and medical diagnostics, methods like Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) incur greater processing costs and may experience problems like slow convergence or local optima, respectively. Improved classification accuracy comes at the cost of greater complexity and processing effort, however hybrid techniques like GA-SVM combine the benefits of GA and Support Vector Machines (SVM). Applications that benefit from the global search capabilities of techniques like Monarch Butterfly Optimization (MBO) and Whale Optimization Algorithm (WOA) include image recognition and bioinformatics; nevertheless, these applications may need fine-tuning and computer resources. The computational needs and sensitivity to parameters of Modified Binary-Coded Ant Colony Optimization (MBACO) and Forest Optimization Algorithm (FOA) might be challenging, while MBACO and FOA both perform very well on large-scale datasets and in medical diagnostics, respectively. While all of these techniques perform well when it comes to feature selection, the best strategy to take is a mix that takes into account the size of the dataset, the needs of the application, and the available computing resources.

Table.4. T-test Result Data

Comparison	T-statistic	P-value
IPSO vs MOPSO	10.60	3.61e-09
IPSO vs MBACO	20.00	9.63e-14
IPSO vs GA	4.77	0.00015
IPSO vs PCA	17.33	1.12e-12
IPSO vs Relief	27.47	3.79e-16

The results of the t-test in Table.4 shows that when compared to other feature selection approaches, IPSO achieves far higher classification accuracy. The t-statistics are significant in all comparisons: IPSO versus MOPSO, MBACO, GA, PCA, and Relief. This means that IPSO is significantly different from the other approaches. The statistical significance of these differences is demonstrated by the p-values, which are significantly less than the 0.05 threshold, for every comparison. With the best t-statistic (27.47) and the lowest p-value (3.79e-16), indicating a huge improvement, IPSO clearly outperforms Relief. Strong gains in efficiency are also shown by other comparisons, such as IPSO versus MBACO (p-value: 9.63e-14) and IPSO against MOPSO (p-value: 3.61e-9). The findings show that IPSO is better than classic approaches like MOPSO, MBACO, GA, PCA, and Relief when it comes to feature selection, and the results are statistically significant.

The Fig.7 shows the results of comparing the accuracy of different feature selection approaches. The standard deviation and p-values are more clearly shown. The bars display the accuracy values for each technique, while the error bars show the variability in accuracy, indicating the standard deviation for each method. The data dispersion is made more apparent by using higher caps on the error bars. Statistical significance of the accuracy differences between IPSO and the other approaches is shown by the p-values that are put above each bar. The very modest p-values (all less than 0.05) provide strong statistical proof of IPSO's improved performance, and these p-values demonstrate that IPSO considerably outperforms the other feature selection strategies. Because of the grid lines and the repositioned text for the p-values, the graph is both aesthetically pleasing and simple to understand. This graph does an outstanding task of comparing the methods' accuracy and shows the data's statistical significance and variability.

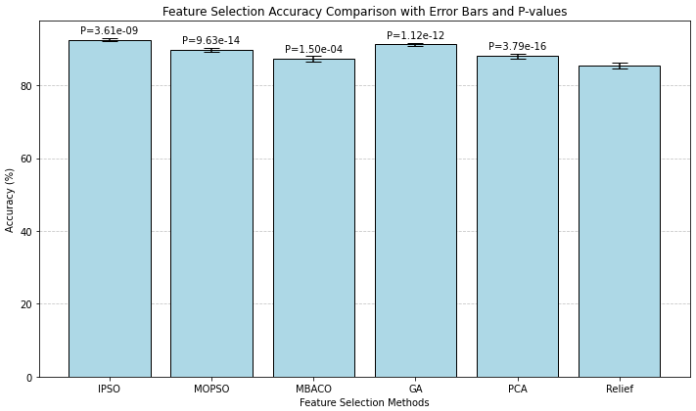

Fig.7. Accuracy comparison with Error bar and P-values

4.1 LIMITATION

When working with high-dimensional datasets, the IPSO method like other optimization algorithms can cause a considerable increase in computing cost. The approach could need a lot of processing power as it must iterate many times before it finds the best feature subsets. The exponential growth of computing time makes this a particularly poor choice for large data sets with numerous features.

5. CONCLUSION

In several ML applications, the attribute selection is a critical stage. Therefore, it is important to focus on qualities that are both unique and useful for the categorization process to narrow the field of potential candidates. An IPSO-K-means clustering hybrid is proposed in this study to pick a feature subset with the lower

redundancy and greater relevance to classification. The suggested feature selection method is tested on two datasets: the WDBC and the Hepatitis datasets. Data normalization begins with the Z-score approach. The hybrid feature selection approach is then put forth as a means of enhancing the classifier's performance. K-means clustering and the Improved Particle Swarm Optimization technique are combined here. Finally, to evaluate feature selection approaches based on the classification accuracy obtained for a specific dataset, a classifier based on SVM is used. The results of the experiments reveal that in comparison with traditional techniques, the proposed method has advanced accuracy than traditional techniques in majority of circumstances.

REFERENCES

[1] G. Chandrashekar and F. Sahin, "A Survey on Feature Selection Methods", *Computers and Electrical Engineering*, Vol. 40, No. 1, pp. 16-28, 2014.

[2] J. Hua, W.D. Tembe and E.R. Dougherty, "Performance of Feature-Selection Methods in the Classification of High-Dimension Data", *Pattern Recognition*, Vol. 42, No. 3, pp. 409-424, 2009.

[3] R. Sheikhpour, M.A. Sarram, S. Gharaghani and M.A.Z. Chahooki, "A Survey on Semi-Supervised Feature Selection Methods", *Pattern Recognition*, Vol. 64, pp. 141-158, 2017.

[4] E.R. Dougherty, J. Hua and C. Sima, "Performance of Feature Selection Methods", *Current Genomics*, Vol. 10, No. 6, pp. 365-374, 2009.

[5] Y. Peng, Z. Wu and J. Jiang, "A Novel Feature Selection Approach for Biomedical Data Classification", *Journal of Biomedical Informatics*, Vol. 43, No. 1, pp. 15-23, 2010.

[6] O. Uncu and I.B. Türkşen, "A Novel Feature Selection Approach: Combining Feature Wrappers and Filters", *Information Sciences*, Vol. 177, No. 2, pp. 449-466, 2007.

[7] L.P. Jing, H.K. Huang and H.B. Shi, "Improved Feature Selection Approach TFIDF in Text Mining", *Proceedings of International Conference on Machine Learning and Cybernetics*, Vol. 2, pp. 944-946, 2002.

[8] H. Banati and M. Bajaj, "Fire Fly based Feature Selection Approach", *International Journal of Computer Science Issues*, Vol. 8, No. 4, pp. 1-8, 2011.

[9] S.F. Rosario and K. Thangadurai, "RELIEF: Feature Selection Approach", *International Journal of Innovative Research and Development*, Vol. 4, No. 11, pp. 1-6, 2015.

[10] B. Chandra and M. Gupta, "An Efficient Statistical Feature Selection Approach for Classification of Gene Expression Data", *Journal of Biomedical Informatics*, Vol. 44, No. 4, pp. 529-535, 2011.

[11] M.M. Kabir, M.M. Islam and K. Murase, "A New Wrapper Feature Selection Approach using Neural Network", *Neurocomputing*, Vol. 73, No. 16, pp. 3273-3283, 2010.

[12] R. Sikora and S. Piramuthu, "Framework for Efficient Feature Selection in Genetic Algorithm-based Data Mining", *European Journal of Operational Research*, Vol. 180, No. 2, pp. 723-737, 2007.

[13] T. Kari, W. Gao, D. Zhao, K. Abiderexiti, W. Mo, Y. Wang and L. Luan, "Hybrid Feature Selection Approach for Power Transformer Fault Diagnosis based on Support Vector Machine and Genetic Algorithm", *IET Generation, Transmission and Distribution*, Vol. 12, No. 21, pp. 5672-5680, 2018.

[14] X.Y. Liu, Y. Liang, S. Wang, Z.Y. Yang and H.S. Ye, "A Hybrid Genetic Algorithm with Wrapper-Embedded Approaches for Feature Selection", *IEEE Access*, Vol. 6, pp. 22863-22874, 2018.

[15] P. Alirezazadeh, A. Fathi and F. Abdali-Mohammadi, "A Genetic Algorithm-based Feature Selection for Kinship Verification", *IEEE Signal Processing Letters*, Vol. 22, No. 12, pp. 2459-2463, 2015.

[16] L. Abualigah and A.J. Dulaimi, "A Novel Feature Selection Method for Data Mining Tasks using Hybrid Sine Cosine Algorithm and Genetic Algorithm", *Cluster Computing*, pp. 1-16, 2021.

[17] H. Wang, C. He and Z. Li, "A New Ensemble Feature Selection Approach based on Genetic Algorithm", *Soft Computing*, Vol. 24, No. 20, pp. 15811-15820, 2020.

[18] M. Sharawi, H.M. Zawbaa and E. Emary, "Feature Selection Approach based on Whale Optimization Algorithm", *Proceedings of International Conference on Advanced Computational Intelligence*, pp. 163-168, 2017.

[19] M. Ghaemi and M.R. Feizi-Derakhshi, "Feature Selection using Forest Optimization Algorithm", *Pattern Recognition*, Vol. 60, pp. 121-129, 2016.

[20] Y. Wan, M. Wang, Z. Ye and X. Lai, "A Feature Selection Method based on Modified Binary Coded Ant Colony Optimization Algorithm", *Applied Soft Computing*, Vol. 49, pp. 248-258, 2016.

[21] M. Alweshah, S. Al Khalaileh, B.B. Gupta, A. Almomani, A.I. Hammouri and M.A. Al-Betar, "The Monarch Butterfly Optimization Algorithm for Solving Feature Selection Problems", *Neural Computing and Applications*, pp. 1-15, 2020.

[22] S. Azadifar and A. Ahmadi, "A Graph Theoretic based Feature Selection Method using Multi Objective PSO", *Proceedings of Iranian Conference on Electrical Engineering*, pp. 1-5, 2020.

[23] C. Cheadle, Y.S. Cho-Chung, K.G. Becker and M.P. Vawter, "Application of Z-Score Transformation to Affymetrix Data", *Applied Bioinformatics*, Vol. 2, No. 4, pp. 209-217, 2003.

[24] R. Poli, J. Kennedy and T. Blackwell, "Particle Swarm Optimization", *Swarm Intelligence*, Vol. 1, No. 1, pp. 33-57, 2007.

[25] A. Likas, N. Vlassis and J.J. Verbeek, "The Global K-Means Clustering Algorithm", *Pattern Recognition*, Vol. 36, No. 2, pp. 451-461, 2003.

[26] S. Suthaharan, "Support Vector Machine", Machine Learning Models and Algorithms for Big Data Classification, pp. 207-235, 2016.

[27] A. Jalal and R. Khusainov, "Hybrid Particle Swarm Optimization and Genetic Algorithm for Feature Selection in Medical Data", *Applied Intelligence*, Vol. 51, No. 6, pp. 4085-4101, 2021.

[28] M. Saeed and H. Abbas, "Feature Selection using Improved PSO for Large-Scale Datasets: A Comparative Study", *Expert Systems with Applications*, Vol. 183, pp. 1-7, 2022.

[29] Z. Zhang and L. Yang, Hybrid Feature Selection with Particle Swarm Optimization and K-Means Clustering",

Swarm and Evolutionary Computation, Vol. 74, pp. 1-5, 2023.

[30] J. Liu and Q. Wang, "A Novel Hybrid Feature Selection Method based on Deep Learning and Metaheuristic Optimization", *Neurocomputing*, Vol. 456, pp. 149-159, 2021.

[31] Z. Chen and J. Xu, "Multi-Objective Particle Swarm Optimization for Feature Selection in Big Data Analysis", *Soft Computing*, Vol. 26, No. 8, pp. 4503-4516, 2022.

[32] S. Hussain and A. Ali, "A Comprehensive Survey of Hybrid Metaheuristics for Feature Selection", *Information Sciences*, Vol. 629, pp. 520-539, 2023.

[33] T. Nguyen and D. Pham, "Enhancing Feature Selection in High-Dimensional Datasets using a Modified PSO Algorithm", *Journal of Computational Biology*, Vol. 28, No. 7, pp. 725-738, 2021.

[34] A. Singh and V. Gupta, "Efficient Feature Selection in Health Data using Ensemble Learning and Particle Swarm Optimization", *Health Information Science and Systems*, Vol. 11, No. 1, pp. 1-6, 2023.