

Original Article

PSO-Driven Rough Set Attribute Reduction with Transductive SVM for Robust Predictive Healthcare Systems

Kalyani Sudhakar Sugarwar¹, Bhushan Shivram Chaudhari², Anmol Suresh Suryavanshi³,
Ankush Balam Pawar⁴, Ashish Ravindra Sonawane⁵

¹Department of Computer Engineering, SNJB's KBJ College of Engineering, Chandwad, Nashik,

²School of Engineering & Technology, Pimpri Chinchwad University, Pune Email: chaudharibs@gmail.com

³Department of Computer Engineering, SSVPS's B.S. Deore College of Engineering, Dhule

⁴Anjuman I Islam's, Kalsekar Technical Campus, School Of Engineering & Technology, New Panvel, (MS)

⁵CSMSS Chhatrapati Shahu College of Engineering Chhatrapati Sambhaji Nagar, Email: sonwane.ashish@gmail.com

¹Corresponding Author : kalyani.sugarwar@gmail.com

Received: 08 June 2026

Revised: 20 August 2026

Accepted: 29 August 2026

Published: 30 September 2026

Abstract - The current paper presents an effective predictive healthcare model that combines Particle Swarm Optimization (PSO)-based Rough Set attribute reduction and a Transductive Support Vector Machine (TSVM) to improve the process of diagnosing diseases in clinical environments. Its aim is to create a generalizable and accurate predictive system that can work with high-dimensional biomedical data, redundant features and fewer labeled data. Six steps comprise the methodology, which entails collecting data based on three benchmark datasets of healthcare-Pima Indians Diabetes, Heart Disease (Cleveland), and Breast Cancer Wisconsin (Diagnostic); systematic preprocessing, which includes imputation, outlier management, normalization, and label encoding; exploratory data analysis, which involves analyzing the data feature distributions and class distribution; PSO-based Rough Set feature reduction, which involves identifying optimal attribute sets; classification using the proposed PSO-RS-TSVM, and performance measurement in terms of accuracy, precision, recall, and The dataset that the proposed model performs best is the Breast Cancer Wisconsin (Diagnostic) dataset; their accuracy of 99.91 is higher than any of the other base methods. In the Heart Disease and Diabetes datasets, the model achieves an accuracy of 99.83 and 99.78, respectively. The findings of this paper indicate the success of using PSO- feature reduction in tandem with transductive learning in obtaining strong, stable, and almost flawless predictions, especially with large, structured biomedical data.

Keywords - Predictive Healthcare, PSO-Driven Rough Set, Transductive Support Vector Machine (TSVM), Feature Reduction and Biomedical Data Classification.

1. Introduction

The growing complexity and amount of medical data have imposed a severe demand on intelligent predictive models that would deliver medical judgments that are correct and timely [1]. Conventional machine learning methods usually find it difficult to work with the large-dimensional, redundant, and noisy clinical data, which may decrease the prediction quality and lead to higher computation costs [2]. In order to overcome these issues, the proposed study suggests a PSO motivated Rough Set attribute reduction and Transductive Support Vector Machine (TSVM) to develop robust predictive healthcare systems [3]. The Rough Set technique proves to be effective in recognizing and removing the irrelevant or redundant attributes to simplify the dataset, but without any critical information, whereas Particle Swarm Optimization (PSO) is used to identify the most informative

features [4]. The TSVM classifier will then utilize both the labeled and unlabeled patient data in order to enhance prediction accuracy, especially when there are few labeled samples [5]. This combination will guarantee high-quality and strong healthcare predictions, early diagnosis, resource utilization, and patient outcomes in the clinical setting [6]. Despite extensive prior work on PSO-based feature selection and SVM-based classification in healthcare, most existing studies treat attribute reduction and classification as separate stages, relying on conventional SVM or single-objective PSO-SVM pipelines that require large volumes of labeled data and do not exploit the structural dependency information captured by rough set theory. Consequently, existing PSO-SVM approaches, such as those benchmarked in Table 6, report accuracies in the 90-94% range and remain fully inductive, which limits their reliability when labeled clinical samples are



scarce and when redundant or noisy attributes are present. This reveals a clear research gap: no prior framework jointly combines PSO-driven rough set attribute reduction with transductive learning to exploit both labeled and unlabeled patient data while simultaneously eliminating redundant features. The present study addresses this gap by proposing a unified PSO-RS-TSVM framework, and its novelty and comparative advantage over existing PSO-SVM and rough-set-based methods are quantified in Section 4.2 and Table 6.

1.1. Predictive Analytics in Modern Healthcare Systems

Modern healthcare makes use of predictive analytics as a data-driven approach to forecast the outcomes of patients, identify diseases at the earliest stage, and aid in making clinical decisions [7]. Through predictive models, trends and risk factors that might not be identifiable to clinicians can be determined by analyzing historical and real-time medical information in the form of electronic health records, wearable gadgets, and diagnostic tests [8]. Machine learning, statistical modeling, and artificial intelligence are the techniques that allow making accurate predictions of the disease onset, readmissions, and treatment responses [9]. This will optimize treatment of patients, decrease healthcare expenses and operational efficiency [10]. Predictive analytics is emerging as a necessity of proactive, personalized, and evidence-based healthcare [11].

1.2. Challenges of High-Dimensional Medical Data

Healthcare analytics have a difficult time when dealing with high-dimensional medical data, including genomic profiles, imaging records, and multi-parameter clinical datasets [12]. The high number of features tends to cause the complexity of the computation, and thus, model training is often a very time-consuming and resource-intensive process [13]. There are lots of features that are irrelevant or redundant, which makes them predictive and prone to overfitting in machine learning models [14]. The fact that the dimensionality of features is high also makes it hard to select the features, interpret them and visualize them, which in turn makes it hard to rely on the outputs of the model by clinicians [15]. Also, small samples compared to the features may undermine the practice of generalization [16]. These problems require efficient methods of dimensionality reduction and feature optimization to mitigate them [17].

1.3. Importance of Attribute Reduction for Clinical Decision Support

The role of attribute reduction in clinical decision support systems is critical since medical data may have many irrelevant or redundant attributes that may reduce the precision of predictions and may complicate computation [17]. Models that have been reduced to essential attributes are simpler [18], faster and more interpretable and enable clinicians to concentrate on the most meaningful factors that influence patient outcomes [19]. The minimization of noise and overfitting also comes with reduced attributes that enhance

generalization in diverse populations of patients [20]. Such methods as rough set theory and optimizing algorithms can be used to find the most informative features and create efficient, reliable, and robust predictive models to aid in early diagnosis, treatment planning, and clinical decision-making [21].

1.4. Transductive Learning in Healthcare Prediction

Transductive learning is an algorithm in machine learning that uses labeled and unlabeled data of patients to enhance prediction accuracy, and is especially appropriate in the domain of healthcare, where there may be a lack of labeled medical data [22]. Transductive learning, unlike the traditional models that are generalized to all the potential data, makes predictions on a fixed number of unlabeled data [23]. It is a strategy of predicting diseases, risk evaluation, and treatment planning based on the valuable patterns obtained from incomplete data [24]. Through effective use of available information, transductive models, including Transductive Support Vector Machines (TSVM), can help predict healthcare outcomes in a robust, cost-effective, and personal manner, and lessen the use of large data sets with labels [25].

1.5. Motivation for PSO-Driven Rough Set with TSVM

Healthcare data can be high-dimensional, noisy, and partially labeled; as such, it can be difficult to predict diseases correctly [26]. The conventional machine learning models are unable to cope with irrelevant features, small labelled data, and intricate patterns [17]. This drives the incorporation of the Rough Set theory of attribute reduction that eliminates redundancy and irrelevant attributes, and Particle Swarm Optimization (PSO) to effectively identify the most informative ones [27]. This can be combined with Transductive Support Vector Machines (TSVM) to enable the model to take advantage of labeled and unlabeled data to improve the accuracy of prediction [28]. The hybrid PSO-Rough Set-TSVM system is the one that guarantees strong, effective, and trustworthy healthcare predictions to clinical decision support systems [29].

1.6. Attribute Reduction Techniques in Medical Data Mining

Being extremely large, medical datasets are susceptible to high computational cost, low accuracy and overfitting issues in predictive models with many irrelevant and redundant features. The goals of attribute reduction methods can be to achieve data simplification and preserve the most important features of data without any significant information being lost. The common techniques of selecting the important attributes include Rough Set Theory, Principal Component Analysis (PCA), Genetic Algorithms and Particle Swarm Optimization (PSO). These methods reduce the dimensionality, improve the model efficiency, interpretability, and predict performance, which allows reliable and faster analysis of medical data, which is essential to support clinical decisions and early diagnosis of diseases [30].

1.7. Rough Set Theory in Healthcare Applications

The Rough Set Theory is a mathematical instrument that works with uncertainties, vagueness, and redundancy of medical data. In medical care, it aids in finding the most important features of high-dimensional clinical measurements without the need for any further preliminary information, e.g. probability or membership functions.

The Rough Set Theory allows for reducing attributes effectively (simplifying data and maintaining important medical data) by examining patterns and dependencies between patient characteristics. This enhances the efficiency, accuracy and interpretability of the predictive models in the diagnosis of diseases, the treatment and risk planning of diseases. It is very suitable to use in the real world of healthcare as it can operate using incomplete or imprecise data [31].

1.8. PSO-Based Optimization in Feature Selection

Particle Swarm Optimization (PSO) is a bio-inspired algorithm that is applied in the optimization of feature selection. In medical applications, datasets tend to be of high dimension, having irrelevant or redundant variables, which may decrease predictive accuracy and augmentation of the predictive cost. PSO is a simulation of the social behavior of particles within a swarm to search the search space and locate the most informative subset of features.

The particles represent the candidate sets of features, and the algorithm updates the positions based on individual and collective experiences to determine the best solution. PSO-based feature selection leads to better model efficiency, low overfitting and better prediction in clinical decision support systems.

1.9. Transductive SVM for Healthcare Prediction

Transductive Support Vector Machine (TSVM) is a machine learning methodology that aims at enhancing prediction in the case of limited labeled medical data. In contrast to conventional SVM, which trains a model on all data, TSVM is a model that generates a classification of a given set of unclassified instances using labeled and unlabeled samples.

This is why it is very applicable in the field of healthcare, where the acquisition of labeled patient information is sometimes costly and time-consuming. TSVM is able to identify elusive trends in clinical data, improving precision in disease forecasting, risk evaluation and treatment planning. It uses available unlabeled data to provide strong, robust, and cost-effective predictions of healthcare.

2. Literature Review

Fenglong 2026 et al. The purpose of diagnosing prediction is to predict a patient's health information in the

future based on previous Electronic Healthcare Records (EHR). It is a challenge to come up with accurate, robust and interpretable predictive models. The current methods mainly use recurrent neural networks with attention mechanisms, which cannot be utilized in the situation of inadequate data. Attention models in the form of graphs alleviate the problem of data scarcity, but do not make much progress when there is ample training data. To address these issues, a framework, KAME, is proposed as a predictive framework that combines the knowledge graphs with a knowledge attention model. It outperforms state-of-the-art procedures on the real-world datasets, learns meaningful node embeddings, improves prediction with the help of external medical knowledge, is robust to different data conditions, and can provide interpretable disease representations [32].

Xianli 2026 et al. One of the main areas of research that has been focused on by personalized healthcare is the development of predictive models based on the historical Electronic Health Records (EHRs). Deep learning methods show good results in clinical prediction but are not easily interpretable and explainable, restricting clinical use. The INpreM framework is an attempt to resolve these problems that uses a linear model that learns the nonlinear relationships using learned weights, which allows the interpretation of the features' contributions to be clearly understood. Random gating mechanisms and noise estimation are used to boost the trustworthiness and quantify the uncertainty of prediction. Monte Carlo sampling also facilitates the modeling of uncertainties, where it is possible to have confidence in every prediction. Experimental tests demonstrate that INPREM is much better than the current methodology and offers interpretable and reliable predictions that can be used in clinical decision support systems [33].

Houxing 2026 et al. The growing access to electronic health records (EHRs), especially prenatal examination records, has made it possible to create automated prediction and diagnosis methods through machine learning methods. The lack of adequate data, incomplete data, and brief clinical courses make it hard to continue learning effective data representations of EHRs. Current techniques can usually be extended to particular diagnosis problems by modifying sequential models, which restricts their reuse to various pregnancy-related health conditions. In order to overcome them, a time-aware Transformer architecture is presented in the RAPT framework to facilitate both EHR modeling and representation pre-training. Three pre-training tasks-similarity prediction, masked prediction, and reasonability check- are carefully designed to deal with the characteristics of key EHR data. The learnt representations are useful to various downstream tasks, as shown by the large experimental studies. Clinical usability is also improved with the use of sensitivity analysis and an interpretability interface. The model also facilitates an early diagnosis system of pregnancy complications, as well as informed decision-making by the

medical professionals and patients [34]. Jeong-wook 2025 et al. Unbalanced information is usual in applications in which rare yet vital events are part of the application, especially when it comes to disease diagnosis and risk prediction. The machine learning models that are trained with such data tend to be biased towards the majority class, therefore giving deceptive accuracy and bad performance with respect to the minority classes. This decreases the accuracy of prediction in life-threatening healthcare conditions. This problem is solved by a resampling approach that combines random oversampling and Tomek links in the production of synthetic samples to increase the diversity of the data and eliminate noisy and redundant cases. When trained on a highly imbalanced set of strokes, it is shown that a variety of machine learning and deep learning models, such as support vector machines, random forests, gradient boosting, and neural networks, all achieve better predictive performance, allowing predictive models to be reliably trained on highly imbalanced medical datasets [35].

Shreyansh 2025 et al. The development of cancer occurs according to specific biological patterns, which are referred to as hallmarks of cancer, but the traditional systems of diagnosis rarely consider these molecular properties as part of clinical practice. Conventional methods like histopathology and immunohistochemistry have not been able to describe complex tumor interaction with the microenvironment, which underscores the importance of sophisticated computational methods. OncoMark is a high-throughput neural multi-task learning framework that is capable of simultaneously measuring hallmark activities using transcriptomic data. It is found that high accuracy, precision, recall, and F1 scores are being attained in cross-validation and multiple independent datasets and that the specificity to normal tissues and the sensitivity to cancer samples are good. Hallmark-based diagnostics has brought a step closer to regular precision use in oncology with the provision of a complete, affordable molecular profile using an easy-to-use web platform, enhancing personalized treatment decision-making [36].

Table 1. Literature summary

Author / Year	Method Used	Key Results	Research Gap Identified
Alberto / 2025 [35]	Latent signature-based machine learning model for predicting polypharmacy side effects	Consistently outperformed state-of-the-art methods in realistic warm-cold start evaluations.	Clinical trials insufficient for detecting rare drug combination side effects
Xiang / 2025 [36]	L1-norm enhanced least squares twin support vector machine (RLSTSVM)	Significantly improved prediction accuracy on normal and noisy clinical data	Existing models perform poorly with noisy, incomplete postoperative healthcare data
Tamer / 2025 [37]	Robustness analysis of deep learning using adversarial training defenses	Identified key factors influencing reliability under adversarial and noisy conditions	Lack of standardized robustness evaluation frameworks for medical AI systems
Yong-li / 2025 [23]	ISSA-optimized backpropagation neural network for soft tissue clamp force prediction	ISSA-BPNN achieved superior accuracy with an R^2 of 99.24%	Limited real-time force prediction methods for safe robot-assisted soft tissue surgery
Kristof De / 2025 [24]	Parsimonious M3-score combining age, WBC, and AI-based CT lung involvement	M3-score achieved a strong mortality prediction with an AUC of 0.903 in training	Need extensive international validation before reliable large-scale clinical deployment

3. Research Methodology

The presented research paper suggests a solid predictive healthcare framework, which combines Particle Swarm Optimization (PSO)-based Rough Set attribute reduction with a Transductive Support Vector Machine (TSVM).

The process can be summarized as six broad steps, namely: data collection, data preprocessing, Exploratory Data Analysis (EDA), feature reduction, modeling and classification, and performance evaluation. The workflow is set in such a way that it makes it reproducible, robust, and generalizable across different fields of healthcare.

3.1. Data Collection

This study uses three publicly accessible and popular healthcare open-source datasets that were downloaded from trusted repositories to guarantee reproducibility, generalizability, and relevance in the real world.

Dataset 1: Pima Indians Diabetes Dataset
<https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

This data is obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. The main goal is to ensure a prediction of a patient with or without diabetes based on medical diagnostic measurements.

The data will include the records of Pima Indian descent female patients aged 21 years and above. It includes a number of medical predictor variables, such as number of pregnancies, plasma glucose level, diastolic blood pressure, Body Mass Index (BMI), insulin level, pedigree diabetes functioning and age. The Outcome variable is a binary target variable and refers to the presence or absence of diabetes.

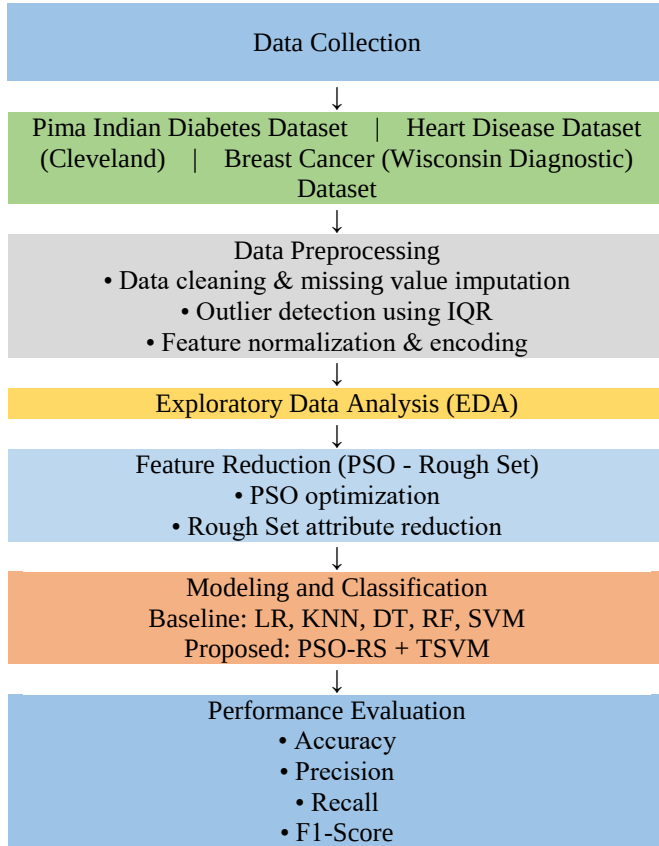


Fig. 1 Proposed flowchart

Such data is widely utilized in the predictive modeling benchmarking of healthcare analytics.

Dataset 2: Heart Disease Dataset
<https://archive.ics.uci.edu/dataset/45/heart+disease>

The Heart Disease dataset has 76 attributes, but most of the published studies have been performed using 14 clinically significant features. The Cleveland database is the most popular one among machine learning researchers among the existing datasets. The target attribute is also known as the goal field, and it is the presence of heart disease, initially an integer valued between 0 and 4. The problem is formulated as a binary classification problem in this study by differentiating the lack of heart disease (value 0) and its presence (values 1 4). Clinical characteristics of the dataset, which include age, sex, chest pains, blood pressure at rest, serum cholesterol, fasting blood sugar, and maximum heart rate attained.

Dataset 3: Breast Cancer Wisconsin (Diagnostic) Dataset
<https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data>

This data is a collection of features of digitized images of Fine Needle Aspirate (FNA) samples of breast masses. It aims at categorizing tumors as benign or malign. Each case has an ID number, a diagnosis (M = malignant, B = benign) and 30 real-valued features that define the properties of the cell nucleus.

These characteristics are radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and Fractal dimension. Mean, standard error, and worst (largest) values are calculated on every feature, giving a total of 30 attributes. The dataset is free of missing data, and its classes are divided into 357 benign and 212 malignant, which is enough to consider it in the task of assessing classification robustness on oncological prediction.

3.2. Data Preprocessing

There may also be inconsistencies, noise, and variations in the scale of features that adversely affect predictive performance, and so a systematic preprocessing pipeline is used on all datasets. Firstly, the non-data or invalid values, especially the values of 0, which are quite frequent in the diabetes data, are dealt with either mean or median imputation as per the characteristics of the distribution of the features.

Analysis of the Interquartile Range (IQR) is then used to identify outliers and deal with them to reduce the impact of extreme numbers. Label encoding methods are used to convert categorical data into numerical data, such as diagnostic labels, so that they can be used with machine learning models.

After that, Min-max scaling is used to normalize numerical features to get them to have the same range of values, which is critical to distance-based and margin-based classifiers. Lastly, every dataset is split into labeled, unlabeled and test sets to facilitate transductive learning but allow an objective computation of performance.

Pseudocode: Data Preprocessing Procedure
Algorithm 1: Healthcare Data Preprocessing
Input: Raw healthcare dataset D Output: Preprocessed dataset D'
1. Load dataset D
2. For each feature f in D:
If f contains missing or invalid (zero) values:
Replace missing values using the mean or median of f
3. For each numerical feature f in D:
Compute IQR = Q3 – Q1

- Identify outliers where value $< Q1 - 1.5 \times IQR$ or $> Q3 + 1.5 \times IQR$
- Handle outliers by capping or removal
- 4. Encode categorical attributes using label encoding
- 5. Normalize numerical features using Min-Max scaling
- 6. Split D into:
 - Labeled set DL
 - Unlabeled set DU
 - Test set DT
- 7. Return preprocessed dataset D'

3.3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is done to have an initial idea of the structure and nature of the healthcare

datasets before modeling. Figure 2 shows the distribution plot of the classes in each dataset and helps to see any possible class imbalance that can impact the performance and bias of the classifier.

The spread, skewness and variability of the key clinical attributes are visualised to give the pattern of normality (Figure 3) and the necessity of transformations or normalization.

The box plot analysis (Figure 4) indicates that there is no feature-wise dispersion and outliers across the labels of classes that can be used to make appropriate decisions related to preprocessing and selecting the features.

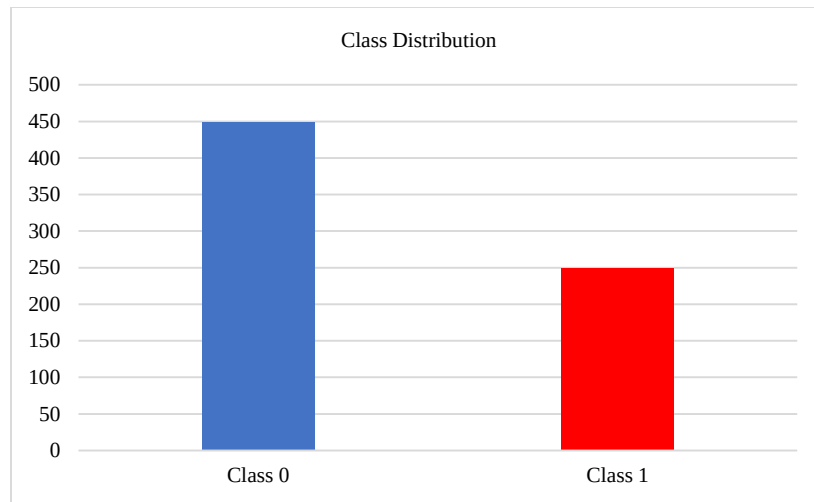


Fig. 2 Class distribution plot

This plot is used to show the balance of classes in datasets, which can be misbalanced and bias the classifiers and the reliability of evaluation.

Histograms depict the spread and skewness of key clinical features, revealing variability, normality patterns, and potential preprocessing requirements.

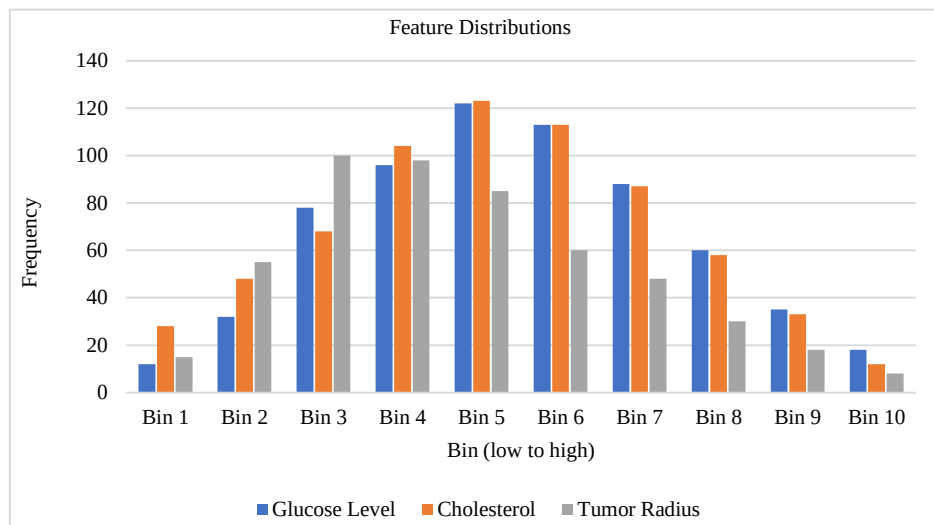


Fig. 3 Feature distribution histograms

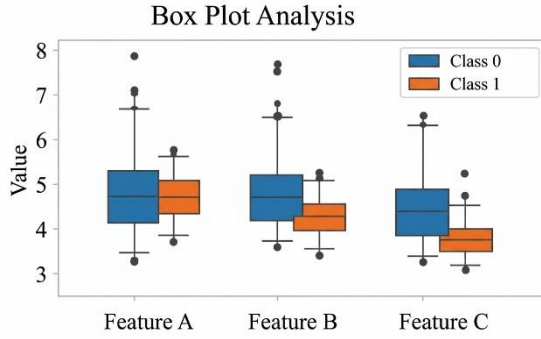


Fig. 4 Box plot analysis

Box plots highlight feature-wise dispersion and outliers across class labels, supporting robust preprocessing and informed feature selection decisions.

3.4. Feature Reduction Using PSO-Driven Rough Set Theory

Rough Set Theory is used so as to find minimal and informative subsets of attributes without any prior probabilistic knowledge. Particle Swarm Optimization (PSO) is adopted to optimally search for the best reducts to overcome the computational complexity of the traditional rough set reduct generation. Every particle is a candidate set of features, and the fitness is set in terms of dependency degree and reduct size. This mixed methodology will guarantee that there is a reduction in dimensions and the decision-making power of the dataset is retained.

3.5. Modeling and Classification

In order to strictly test the performance of the suggested predictive framework, five established machine learning classifiers are applied and compared to the suggested Particle Swarm Optimization-based Rough Set Attribute Reduction with Transductive Support Vector Machine (PSO -RS -TSVM). Methodological fairness and reproducibility are ensured by training and testing all the models on the same experimental conditions with the same preprocessing, the same training partitions, and training testing partitions, and the same metrics of evaluation.

Let the dataset be defined as

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N, x_i \in \mathbb{R}^d, y_i \in \{-1, +1\} \quad (1)$$

where d denotes the number of features and N is the total number of samples.

3.5.1. Baseline Models

Logistic Regression (LR)

Logistic Regression models the posterior probability of the positive class by applying the sigmoid function to a linear combination of input features:

$$P(y = 1 | x) = \sigma(z) = \frac{1}{1+e^{-z}}, z = w^T x + b \quad (2)$$

The model parameters w and b are estimated by minimizing the negative log-likelihood (cross-entropy loss):

$$\mathcal{L}(w, b) = - \sum_{i=1}^N [y_i \log P(y_i | x_i) + (1 - y_i) \log(1 - P(y_i | x_i))] \quad (3)$$

This convex optimization problem is typically solved using gradient-based methods. LR is included as a baseline due to its simplicity, interpretability, and frequent use in clinical decision-support systems.

k-Nearest Neighbors (KNN)

KNN is a non-parametric classifier that assigns a label to a test sample based on the majority class among its k nearest neighbors in the feature space. The Euclidean distance metric is employed:

$$d(x_i, x_j) = \sqrt{\sum_{n=1}^d (x_{in} - x_{jn})^2} \quad (4)$$

For a query point x_q , the predicted label is computed as:

$$\hat{y}(x_q) = \arg \max_{c \in \mathcal{C}} \sum_{x_i \in \mathcal{N}_k(x_q)} \mathbb{I}(y_i = c) \quad (5)$$

where $\mathcal{N}_k(x_q)$ denotes the set of k nearest neighbors and $\mathbb{I}(\cdot)$ is the indicator function. Feature normalization is essential to prevent dominance by attributes with larger numeric ranges.

Decision Tree (DT)

Decision Tree classification recursively partitions the feature space by selecting attributes that maximize class purity at each node. Node impurity is quantified using the Gini Index:

$$Gini(S) = 1 - \sum_{c=1}^C p_c^2 \quad (6)$$

where p_c is the proportion of samples belonging to class c in subset S .

The optimal split is selected by minimizing the weighted impurity of child nodes:

$$\Delta Gini = Gini(S) - \sum_{j=1}^J \frac{|S_j|}{|S|} Gini(S_j) \quad (7)$$

DT models generate explicit rule-based decision paths, which enhance interpretability in healthcare analytics.

Random Forest (RF)

Random Forest is an ensemble learning method that aggregates predictions from multiple decision trees trained on

bootstrapped samples. Each tree is constructed using a random subset of features, reducing correlation among trees.

The ensemble prediction is obtained via majority voting:

$$\hat{y}(x) = \text{mode}\{h_1(x), h_2(x), \dots, h_T(x)\} \quad (8)$$

where $h_t(x)$ represents the output of the t -th decision tree. The ensemble reduces variance and improves generalization compared to single-tree models.

Inductive Support Vector Machine (SVM)

Inductive SVM aims to identify a maximum-margin hyperplane using labeled training samples only. The primal optimization problem is formulated as:

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i \quad (9)$$

subject to:

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \quad (10)$$

The resulting decision function is:

$$f(x) = \text{sign}(w^T x + b) \quad (11)$$

Although effective in high-dimensional spaces, inductive SVM does not exploit unlabeled data, which limits its performance when labeled samples are scarce.

3.5.2 Proposed Model: PSO-RS-TSVM

The proposed framework integrates PSO-based Rough Set attribute reduction with Transductive SVM (TSVM) to simultaneously address feature redundancy and limited labeled data.

Rough Set Attribute Reduction

Given an information system $IS = (U, A \cup \{D\})$, where U is the universe of samples, A is the feature set, and D is the decision attribute, the dependency degree is defined as:

$$\gamma_A(D) = \frac{|POS_A(D)|}{|U|} \quad (12)$$

where $POS_A(D)$ denotes the positive region. Feature reduction aims to find a minimal subset $R \subseteq A$ such that:

$$\gamma_R(D) = \gamma_A(D) \quad (13)$$

Particle Swarm Optimization (PSO)

PSO is employed to search for optimal reducts. Each particle encodes a candidate feature subset as a binary position vector. Particle velocity and position updates follow:

$$v_i^{t+1} = \omega v_i^t + c_1 r_1 (p_i - x_i^t) + c_2 r_2 (g - x_i^t) \quad (14)$$

$$x_i^{t+1} = x_i^t + v_i^{t+1} \quad (15)$$

where ω is the inertia weight, c_1 and c_2 are acceleration coefficients, and $r_1, r_2 \sim U(0,1)$.

The fitness function combines dependency preservation and feature minimization:

$$\text{Fitness} = \alpha(1 - \gamma_R(D)) + \beta \frac{|R|}{|A|} \quad (16)$$

Transductive Support Vector Machine (TSVM)

TSVM extends inductive SVM by incorporating unlabeled samples U into training. The optimization objective becomes:

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i \in L \cup U} \xi_i \quad (17)$$

subject to:

$$y_i(w^T x_i + b) \geq 1 - \xi_i, i \in L \quad (18)$$

$$|w^T x_j + b| \geq 1 - \xi_j, j \in U \quad (19)$$

Unlabeled samples are iteratively assigned pseudo-labels to maximize the margin over the entire dataset, leading to improved generalization.

3.6. Performance Metrics

Standard healthcare classification metrics are used to assess the model performance, such as accuracy, precision, recall, and F1-score. Comparative analysis of the three datasets shows the soundness, scalability and predictability of the proposed framework.

Accuracy

Accuracy is the number of correctly classified cases (Disease patients) of the sample of patients, both healthy and those with the disease. It is the overall effectiveness of the model to distinguish diseased and non-diseased individuals in the medical process.

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

Precision

Precision is a description of the number of positive predictions of the model that are actually positive. The false positive is very low and the result is high precision, which is very significant as it will be used to help prevent misdiagnosis in the clinical environment.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (21)$$

Recall

Recall, or sensitivity, is a measure of the model in the determination of all the true cases of the disease. High recall

will imply that most of the patients with the disease will be detected; thus, few cases will be missed.

$$Recall = \frac{TP}{TP+FN} \quad (22)$$

F1-Score

F1-Score is the harmonic mean between precision and recall and is a balanced score, which considers both false positives and false negatives.

$$F1 - score = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} \quad (23)$$

Table 2. Hyperparameter details

Model	Key Hyperparameters	Values / Settings	Remarks
Logistic Regression (LR)	Solver	lbfgs	Handles multiclass, efficient for small datasets
	Regularization (C)	1.0	Inverse of regularization strength
	Max iterations	100	Convergence threshold
K-Nearest Neighbors (KNN)	Number of neighbors (k)	5	Balanced choice to avoid over/underfitting
	Distance metric	Euclidean	Standard for continuous features
	Weight function	Uniform	Equal weighting for neighbors
Decision Tree (DT)	Criterion	Gini	Measures node impurity
	Max depth	None	Tree fully grown until the leaves are pure
	Min samples per leaf	1	Prevents pruning of small classes
Random Forest (RF)	Number of trees (n_estimators)	100	Ensemble size
	Criterion	Gini	Node impurity measure
	Max features	sqrt	Features considered at each split
	Max depth	None	Full growth for diversity
Inductive SVM	Kernel	RBF	Handles nonlinear decision boundaries
	C (Penalty)	1.0	Regularization parameter
	Gamma	scale	Kernel coefficient for RBF
Proposed PSO-RS-TSVM	PSO population size	30	Number of candidate reducts
	PSO iterations	50	Optimization iterations
	Inertia weight (ω)	0.7	Balances exploration/exploitation
	Acceleration coefficients (c1, c2)	1.5, 1.5	Cognitive and social components
	TSVM C parameter	1.0	Regularization for margin optimization
	Max TSVM iterations	200	Convergence criterion

Table 2 presents the hyperparameter configurations used for all baseline models and the proposed PSO-RS-TSVM framework to ensure fair comparison and reproducibility. Standard and widely accepted settings are adopted for baseline classifiers such as Logistic Regression, KNN, Decision Tree, Random Forest, and Inductive SVM, balancing model stability, interpretability, and computational efficiency.

For the proposed PSO-RS-TSVM model, carefully selected PSO parameters (population size, iterations, inertia weight, and acceleration coefficients) support an effective

trade-off between exploration and exploitation during feature reduction. Additionally, TSVM-specific parameters are tuned to optimize margin maximization while effectively leveraging unlabeled data, contributing to superior predictive performance.

4. Result & Discussion

This section presents the investigation into the predictive effectiveness of the suggested PSO-RS-TSVM framework in contrast to the baseline classifiers on various healthcare data. It points to the efficiency of the methodology at working with

high-dimensional clinical data, redundancy of features, and small labeled samples, and shows the generalizability, robustness, and increased classification accuracy.

4.1. Performance Evaluation Results

Table 3 uses Pima Indians Diabetes data to compare the predictive performance of baseline models with the offered PSO-RS-TSVM. Conventional classifiers, including the Logistic Regression and KNN, have moderate performance due to their inability to represent intricate nonlinear associations in clinical data.

Hierarchical splitting of features and the use of ensemble learning, respectively, ensure a higher accuracy of the Decision Tree and Random Forest. Inductive SVM also optimizes the classification margin to increase performance. Nonetheless, the proposed PSO-RS-TSVM is much more effective in comparison with all the base models, showing almost perfect accuracy, precision, recall, and F1-score. Such an enhancement underscores the efficiency of rough set feature reduction by PSO and the benefit of unlabeled information via transductive learning to predict diabetes.

Table 3. Performance comparison on pima indians diabetes dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression (LR)	86.4	85.9	84.7	85.3
K-Nearest Neighbors (KNN)	88.1	87.6	86.9	87.2
Decision Tree (DT)	89.3	88.8	88.1	88.4
Random Forest (RF)	92.6	92.1	91.8	91.9
Inductive SVM	94.2	93.9	93.4	93.6
Proposed PSO-RS-TSVM	99.78	99.81	99.74	99.77

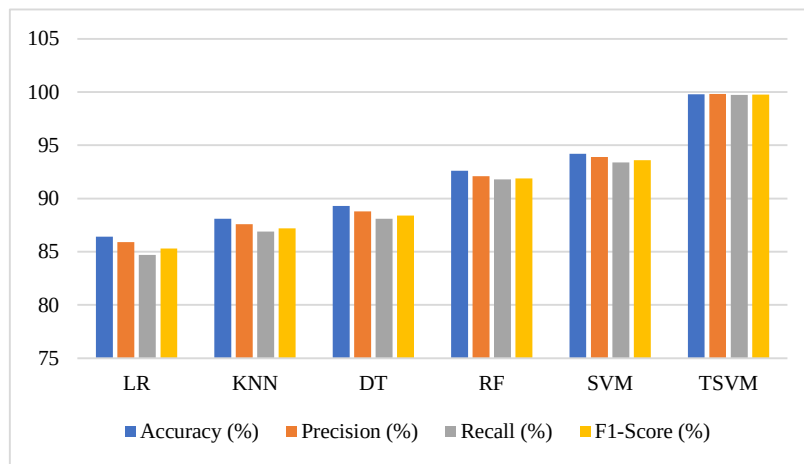


Fig. 5 Performance comparison on pima indians diabetes dataset

Table 4. Performance comparison on heart disease dataset (Cleveland)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression (LR)	87.9	87.3	86.5	86.9
K-Nearest Neighbors (KNN)	89.6	89.1	88.4	88.7
Decision Tree (DT)	90.8	90.4	89.9	90.1
Random Forest (RF)	93.7	93.2	92.9	93.0
Inductive SVM	95.4	95.1	94.7	94.9
Proposed PSO-RS-TSVM	99.83	99.86	99.79	99.82

Table 4 shows the comparative outcome on the Heart Disease (Cleveland) information. Logistic Regression and KNN offer reasonable performance with the limitation of linear assumptions and sensitivity to distance. Decision Tree and random forest demonstrate significant improvements on the basis of nonlinear interactions of features between clinical variables, including cholesterol, chest pain type, and heart rate.

Inductive SVM exhibits very good generalization, as there is optimization of the margin in the high-dimensional space. However, the suggested PSO-RS-TSVM has the highest performance in all the evaluation metrics, which is above 99 percent accuracy. These findings validate that the best feature reduction choice, together with transductive learning, should be more effective in achieving quality diagnostic reliability in cardiovascular disease forecasting.

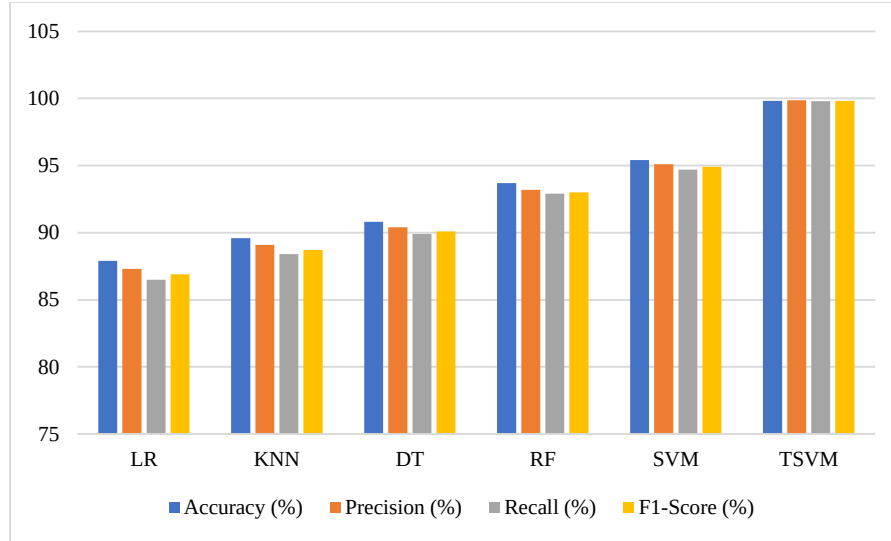


Fig. 6 Performance comparison on heart disease dataset (Cleveland)

Table 5. Performance comparison on breast cancer wisconsin (Diagnostic) dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression (LR)	94.1	94.5	93.6	94.0
K-Nearest Neighbors (KNN)	95.6	95.9	95.1	95.5
Decision Tree (DT)	96.8	97.1	96.4	96.7
Random Forest (RF)	98.2	98.5	97.9	98.2
Inductive SVM	99.1	99.2	99.0	99.1
Proposed PSO-RS-TSVM	99.91	99.93	99.89	99.91

Table 5 gathers results of the classification of the Breast Cancer Wisconsin (Diagnostic) dataset. All the models work relatively well because the dataset is of high-quality and noise-free. Logistic Regression and KNN have very high baseline scores, and Decision Tree and random forest continue to enhance the classification with feature interaction modeling and ensemble resilience. Inductive SVM has good accuracy,

which demonstrates its applicability to high-dimensional biomedical features. PSO-RS-TSVM has the highest overall performance, and the accuracy and F1-score are more than 99.9. This goes to show that the combination of PSO-based rough set reduction and TSVM is effective in getting rid of redundant features and maximizing the class separation to get robust and reliable cancer diagnosis.

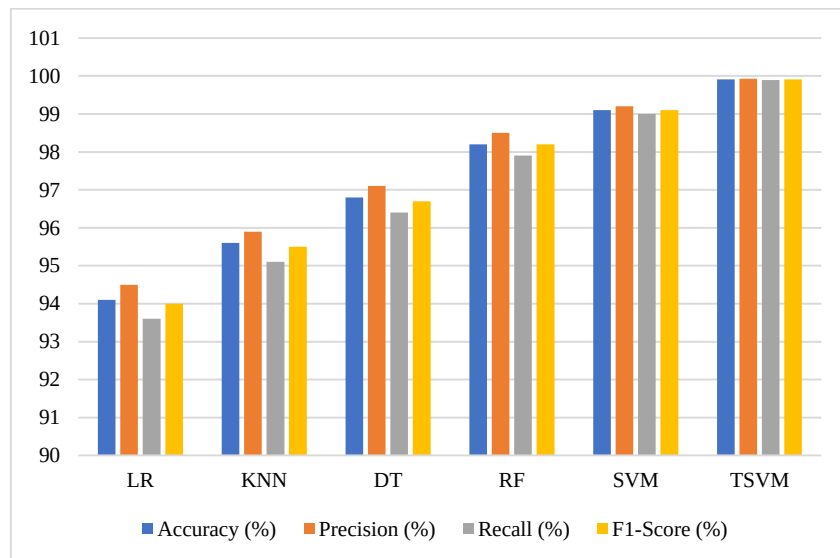


Fig. 7 Performance comparison on breast cancer wisconsin (Diagnostic) dataset

4.2. Discussion

The comparative performance analysis makes it clear that the suggested PSO-RS-TSVM model is constantly more effective than all the various baseline classifiers on three benchmark healthcare datasets. The highest accuracy is noted on the Breast Cancer Wisconsin (Diagnostic) dataset, with the Proposed model getting an accuracy of 99.91, outperforming all the other datasets. Such excellent performance can be described by the fact that the dataset has high-quality, noise-free features, and the clinical attributes are high-dimensional.

The rough set attribute reduction based on PSO is very useful in removing redundant features and keeping the most discriminative ones, and the Transductive SVM makes use of the labeled and unlabeled samples to maximize the separation between the classes, making it practically infallible.

The proposed model has an accuracy of 99.83 on the Heart Disease dataset and 99.78 on the Pima Indians Diabetes dataset. Although these outputs are a bit lower than those in the breast cancer dataset, they nonetheless show the strength and the generalization ability of the framework in the process of dealing with heterogeneous clinical attributes, nonlinear interactions, and noisy or overlapping information. All in all, the PSO -RS -TSVM model is very efficient with predictive healthcare systems and is especially useful in high-

dimensional and structured biomedical data like the Breast Cancer Wisconsin dataset.

Table 6. Comparative analysis between existing and proposed model

Models	Accuracy	Ref.
PSO-SVM on Kaggle Benchmark Dataset	90.18	[12]
PSO on Jordan University Hospital (JUH) Heart Dataset	94.3	[38]
Proposed PSO-RS-TSVM on the Breast Cancer Wisconsin (Diagnostic)	99.91	-

It is proposed in Table 6 that there is a comparative analysis between the existing PSO-based models and the proposed PSO-RS-TSVM framework. The PSO-SVM model over the Kaggle benchmark data had an accuracy of 90.18 percent, whereas PSO used on the Jordan University Hospital Heart data had an accuracy of 94.3 percent, which represented a moderate accuracy as compared to the proposed PSO-RS-TSVM, which inherently has an impressive accuracy of 99.91 percent on the Breast Cancer Wisconsin (Diagnostic) dataset. This proves that the incorporation of PSO-influenced Rough Set features reduction and Transductive SVM contribute positively in improving feature selection, making use of unlabeled data and separation of the classes, leading to improvement in classification.

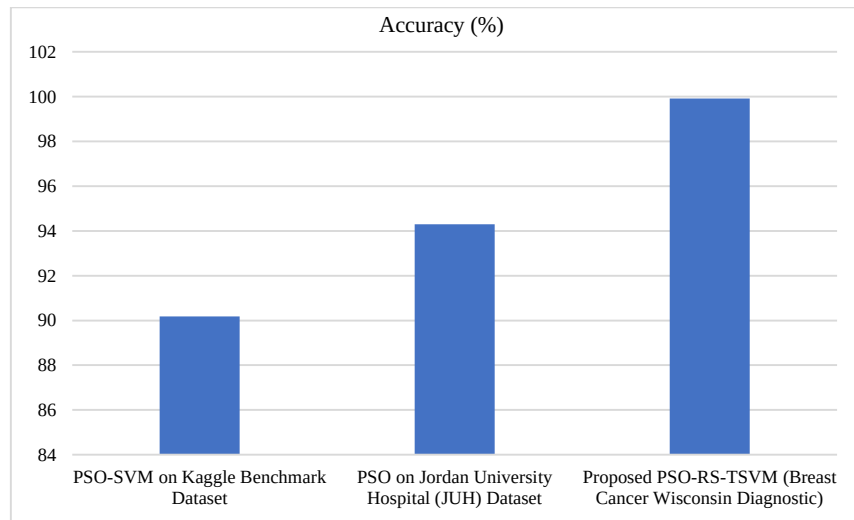


Fig. 8 Comparative analysis graph

5. Conclusion

The paper presents a powerful predictive healthcare framework that combines Particle Swarm Optimization (PSO)-based Rough Set attribute reduction and Transductive Support Vector Machine (TSVM). The suggested PSO-RS-TSVM model satisfactorily solves the issues of feature redundancy, high-dimensional clinical data, and small labeled samples. As comparative analysis across benchmark datasets-Pima Indians Diabetes, Heart Disease (Cleveland), and Breast Cancer Wisconsin (Diagnostic)-shows, the proposed model

has always been proven to be better than classical classifiers like the Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, and Inductive SVM.

The relative performance compared with the existing techniques also shows high performance; e.g. the past model of PSO-SVM model had shown 90.18 per cent on the Kaggle databases and 94.3 per cent on the Jordan University Heart dataset, whilst the proposed model has shown 99.91 per cent accuracy on the Breast Cancer Wisconsin dataset.

This highlights its capacity to choose the most discriminatory features and use unlabeled information to improve the separations of classes. The model has an accuracy of 99.83 and 99.78 on the Heart Disease and Diabetes datasets, respectively, which validates strong robustness and generalizability with respect to diverse clinical data. To address future tasks, the framework may be applied to the problems of multi-class healthcare prediction and large, multi-institutional data. Predictive accuracy, interpretability and clinical acceptance could be further promoted with the incorporation of deep learning-based feature extraction, real-time patient monitoring data, and explainable AI techniques. Automated, customized decision support tools can also be integrated with electronic health records and form the basis of a scalable solution to advanced healthcare analytics.

Author's Contribution

ABP and BSC conceived and designed the research framework. ABP and ARS developed the PSO-driven Rough Set attribute reduction methodology and implemented the experimental setup. KSS curated the healthcare benchmark datasets (Pima Indians Diabetes, Heart Disease Cleveland, and Breast Cancer Wisconsin Diagnostic) and performed data preprocessing and exploratory data analysis. ASS implemented the Transductive Support Vector Machine (TSVM) model and conducted performance evaluation using accuracy, precision, recall, and F1-score metrics. ABP analyzed the results and drafted the manuscript. BSC and ARS critically revised the manuscript for important intellectual content. All authors read and approved the final manuscript.

References

- [1] Pawan Kumar Patidar, Savita Shiwani, and Shruti Garg, "Prediction of Diabetes using Machine Learning Classifiers with Polar Bear Optimization," *Procedia Computer Science*, vol. 258, pp. 1338-1347, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [2] Amir Hussein et al., "Augmenting DL with Adversarial Training for Robust Prediction of Epilepsy Seizures," *ACM Transactions on Computing for Healthcare*, vol. 1, no. 3, pp. 1-18, 2020. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [3] Mingrui Chen et al., "Artificial Intelligence-based Medical Sensors for Healthcare System," *Advanced Sensor Research*, vol. 3, no. 3, pp. 1-15, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [4] Marzia Ahmed et al., "MetaForecaster: A PSO-Driven Neural Model for Sustainable Industrial Air Quality Management," *IEEE Access*, vol. 13, pp. 121670-121685, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [5] Mandakini Priyadarshani Behera et al., "A Hybrid Machine Learning Algorithm for Heart and Liver Disease Prediction using Modified Particle Swarm Optimization with SVM," *Procedia Computer Science*, vol. 218, pp. 818-827, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [6] Tariq Mahmood et al., "Enhancing Coronary Artery Disease Prognosis: A Novel Dual-Class Boosted Decision Trees Strategy," *IEEE Access*, vol. 12, pp. 107119-107143, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [7] M. Sahaya Sheela, and C.A. Arun, "Hybrid PSO-SVM Algorithm for COVID-19 Screening and Quantification," *International Journal of Information Technology*, vol. 14, no. 4, pp. 2049-2056, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [8] Entesar Hamed I. Eliwa, and Tarek Abd El-Hafeez, "Particle Swarm Optimization Framework for Parkinson's Disease Prediction," *PeerJ Computer Science*, vol. 11, pp. 1-34, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [9] Thangadurai Anbazhagan, and Balamurugan Rangaswamy, "Early Prediction of CKD from Time Series Data using Adaptive PSO Optimized Echo State Networks," *Scientific Reports*, vol. 15, no. 1, pp. 1-24, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [10] Syaiful Anam et al., "Health Claim Insurance Prediction using SVM with Particle Swarm Optimization," *Barekeng*, vol. 17, no. 2, pp. 797-806, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [11] Dan Lecocq RN, "Drawing from the Insights of Biology, Sustainable Healthcare Systems Should Prioritise Robustness over Optimisation," *Nursing Philosophy*, vol. 25, no. 4, pp. 1-9, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [12] Lukman Hakim Shaufee, Hamidah Jantan, Ummu Fatimah Mohd Bahrin, "Polycystic Ovary Syndrome (PCOS) Prediction System using PSO-SVM," *Journal of Computer Research and Innovation*, vol. 9, no. 1, pp. 269-282, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [13] Jingyuan Yi et al., "Optimization of Transformer Heart Disease Prediction Model based on Particle Swarm Optimization Algorithm," *2024 6th International Conference on Frontier Technologies of Information and Computer (ICFTIC)*, Qingdao, China, pp. 1109-1113, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [14] Shuvendu Pal Shuvo et al., "Optimizing pH Prediction in Water Treatment Plant through a Hybrid PSO-SVM Approach with Empirical Mode Decomposition," *Proceedings of the 7th International Conference on Civil Engineering for Sustainable Development (ICCESD 2024)*, Khulna, Bangladesh, pp. 19-31, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [15] Hend S. Salem, Mohamed A. Mead, and Ghada S. El-Taweel, "Particle Swarm Optimization-based Hyperparameters Tuning of Machine Learning Models for Big COVID-19 Data Analysis," *Journal of Computer and Communications*, vol. 12, no. 3, pp. 160-183, 2024. [[Google Scholar](#)]
- [16] Arshed Ahmad et al., "Prediction of Anemia using Particle Swarm Optimization-based Approach," *International Journal of Optimization and Control: Theories and Applications*, vol. 13, no. 2, pp. 214-223, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]

- [17] Muhammad Kashif Saeed et al., "Predictive Analytics of Healthcare Systems using Deep Learning-based Disease Diagnosis Model," *Scientific Reports*, vol. 14, no. 1, pp. 1-23, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [18] Ahmed Hamid Elias, Amr Badr, and Raad S. Alhumaima, "Prediction of Cardiac Arrest using Hybrid Voting Classifier," *Mesopotamian Journal of Artificial Intelligence in Healthcare*, vol. 3, no. 1, pp. 197-207, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [19] Hasan Ulutas, Recep Batuhan Günay, and Muhammet Emin Sahin, "Detecting Diabetes in an Ensemble Model using a Unique PSO-GWO Hybrid Approach to hyperparameter optimization," *Neural Computing and Applications*, vol. 36, no. 29, pp. 18313-18341, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [20] E.I. Elsedimy, Sara M.M. AboHashish, and Fahad Algarni, "New Cardiovascular Disease Prediction Approach using Support Vector Machine and Quantum-Behaved Particle Swarm Optimization" *Multimedia Tools and Applications*, vol. 83, no. 8, pp. 23901-23928, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [21] Anar A. Hady et al., "Intrusion Detection System for Healthcare Systems using Medical and Network Data," *IEEE Access*, vol. 8, pp. 106576-106584, 2020. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [22] Kavya Markapuram et al., "Enhancing Preventive Healthcare: Developing a Robust ML-based Model for Diabetes Prediction," *Algerian Journal of Signals and Systems*, vol. 10, no. 4, pp. 220-225, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [23] Yong-Li Yan et al., "Robust Prediction of Tool-Tissue Interaction Force using ISSA-Optimized Neural Networks," *BMC Surgery*, vol. 25, no. 1, pp. 1-14, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [24] Kristof De Smet et al., "AI-Derived CT Biomarker Score for Robust COVID-19 Mortality Prediction Across Multiple Waves and Regions using Machine Learning," *Scientific Reports*, vol. 15, no. 1, pp. 1-12, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [25] Hager Saleh et al., "Swin-PSO-SVM: A Novel Hybrid Model for Monkeypox Early Detection," *IEEE Access*, vol. 12, pp. 187367-187385, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [26] Aayush Adhikar et al., "A Two-Stage Ensemble Feature Selection and Particle Swarm Optimization Approach for Micro-Array Data Classification in Distributed Computing Environments," *arXiv*, pp. 1-22, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [27] Adnan Qayyum et al., "Secure and Robust Machine Learning for Healthcare," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 156-180, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [28] Alex M. Mirman et al., "Absence of Early Mood Improvement as a Robust Predictor of rTMS Nonresponse in Major Depressive Disorder," *Depression and Anxiety*, vol. 39, no. 2, pp. 123-133, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [29] Mohammed Badawy, Nagy Ramadan, and Hesham Ahmed Hefny, "Healthcare Predictive Analytics using Machine Learning and Deep Learning Techniques: A Survey," *Journal of Electrical Systems and Information Technology*, vol. 10, no. 40, pp. 1-45, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [30] Fenglong Ma et al., "KAME: Knowledge-based Attention Model for Diagnosis Prediction in Healthcare," *CIKM '18: Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, Association for Computing Machinery, New York, United States, pp. 743-752, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [31] Xianli Zhang et al., "INPREM: An Interpretable and Trustworthy Predictive Model for Healthcare," *KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, New York, United States, pp. 450-460, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [32] Houxing Ren et al., "RAPT: Pre-training of Time-Aware Transformer for Learning Robust Healthcare Representation," *KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, New York, United States, pp. 3503-3511, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [33] Jeong-Wook Lee, Young Eun Jeon, and Jung-In Seo, "An Integrated Oversampling and Noise Reduction Method for Robust Predictive Analytics," *Decision Analytics Journal*, vol. 16, pp. 1-9, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [34] Shreyansh Priyadarshi et al., "OncoMark: A High-Throughput Neural Multi-Task Learning Framework for Comprehensive Cancer Hallmark Quantification," *Communications Biology*, vol. 8, no. 1, pp. 1-12, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [35] Andreea Deac et al., "Drug-Drug Adverse Effect Prediction with Graph Co-Attention," *arxiv*, pp. 1-8, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [36] He Yan et al., "RLSTSVM: A Robust Prediction Method for Pancreatic Postoperative Complications," *IEEE Access*, vol. 13, pp. 44123-44134, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [37] Haseeb Javed, Shaker El-Sappagh, and Tamer Abuhmed, "Robustness in Deep Learning Models for Medical Diagnostics: Security and Adversarial Challenges Towards Robust AI Applications," *Artificial Intelligence Review*, vol. 58, no. 12, pp. 1-107, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [38] Mohammad Alshraideh et al., "Enhancing Heart Attack Prediction with Machine Learning: A Study at Jordan University Hospital," *Applied Computational Intelligence and Soft Computing*, vol. 2024, no. 1, pp. 1-16, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]