

PSO-XAI: A PSO-Enhanced Explainable AI Framework for Reliable Breast Cancer Detection

Mirza Raquib*, Niloy Das*, Farida Siddiqi Prity*, Arafath Al Fahim, Saydul Akbar Murad, Mohammad Amzad Hossain, MD Jiabul Hoque, Mohammad Ali Moni

Abstract—Breast cancer is considered the most critical and frequently diagnosed cancer in women worldwide, leading to an increase in cancer-related mortality. Early and accurate detection is crucial as it can help mitigate possible threats while improving survival rates. In terms of prediction, conventional diagnostic methods are often limited by variability, cost, and, most importantly, risk of misdiagnosis. To address these challenges, machine learning (ML) has emerged as a powerful tool for computer-aided diagnosis, with feature selection playing a vital role in improving model performance and interpretability. This research study proposes an integrated framework that incorporates customized Particle Swarm Optimization (PSO) for feature selection. This framework has been evaluated on a comprehensive set of 29 different models, spanning classical classifiers, ensemble techniques, neural networks, probabilistic algorithms, and instance-based algorithms. To ensure interpretability and clinical relevance, the study uses cross-validation in conjunction with explainable AI methods. Experimental evaluation showed that the proposed approach achieved a superior score of 99.1% across all performance metrics, including accuracy and precision, while effectively reducing dimensionality and providing transparent, model-agnostic explanations. The results highlight the potential of combining swarm intelligence with explainable ML for robust, trustworthy, and clinically meaningful breast cancer diagnosis.

Index Terms—Breast Cancer, Feature Selection, Particle Swarm Optimization (PSO), Machine Learning, Explainable Artificial Intelligence (XAI), Classification, Medical Diagnosis

I. Introduction

In recent times, cancer has emerged as one of the most significant challenges to global health, causing millions of new cases and deaths each year in diverse populations. Out of many others, breast cancer is the most commonly diagnosed cancer in women across the globe and one of the leading causes of cancer-related deaths, as over 2.3 million new cases have been registered in 2022 alone [1]. Breast cancer starts in the cells of breast tissue, most often in the ducts or lobules, and is distinguished by the uncontrolled growth of abnormal cells that can spread to other tissues of the body and to other

organs [2], [3]. Breast tumors are clinically classified as benign, noncancerous, and generally noninvasive, or malignant, cancerous, aggressive, and capable of spreading to other body parts. Its varied subtypes and progression patterns make it particularly difficult to detect in its early stages and accurately diagnose the cancer.

Despite the numerous challenges, early diagnosis of breast cancer is crucial to improving the efficacy of treatment and the survival rates of patients through timely medical intervention. Breast abnormalities have been commonly diagnosed using clinical diagnostic methods, including mammography, ultrasound, magnetic resonance imaging (MRI), and histopathological examination. However, such techniques are typically constrained by inter-observer reproducibility, high expenses, and the possibility of false positives or false negatives, which may result in unnecessary biopsies or delayed treatment. In addition, the heterogeneity of breast cancer and minor differences between malignant and benign tumors pose a constant problem in the accurate diagnosis of cancer using traditional diagnostic techniques alone. In the context of breast cancer diagnosis, in particular, ML promisingly supports higher sensitivity and specificity and allows reproducible and consistent decision support in clinical pipelines.

In the fields of oncology and broader clinical tasks, a variety of supervised learning algorithms—such as Support Vector Machines (SVM), Random Forests (RF), Naïve Bayes (NB), k-Nearest Neighbors (KNN), Logistic Regression (LR), and gradient-boosting ensembles—have projected strong performance in risk stratification, prognosis, and histopathology-based classification [4], [5], [39]. Over the recent years, deep learning techniques have produced compelling results in medical imaging and digital pathology, leveraging convolutional architectures for large-scale feature learning [7], [8]. In addition to model selection, feature selection (FS) and hyperparameter optimization are crucial for reducing overfitting and improving generalization. Commonly used approaches include filter methods (e.g., Information Gain and Correlation-based Feature Selection) and metaheuristic techniques (e.g., Genetic Algorithms, Particle Swarm Optimization, and Bat Algorithm variants) that are frequently applied to clinical datasets [9], [12].

There are significant research gaps evident in existing studies, particularly concerning dataset-specific overfitting, as many methods lack external validation. Small sample sizes and the prevalence of class imbalance often lead to inflated performance estimates, which limit the broader applicability of reported results. Additionally, issues with computational repro-

Mirza Raquib, Niloy Das, and Farida Siddiqi Prity contribute equally. N. Das is from Department of Information and Communication Engineering, Noakhali Science and Technology University, Bangladesh. F. S. Prity is from Department of Computer Science & Engineering, Netrokona University, Bangladesh. M. Raquib is from Department of Computer and Communication Engineering, International Islamic University Chittagong, Bangladesh. A. Al Fahim is from Department of Mechatronics and Industrial Engineering, Chittagong University of Engineering and Technology, Bangladesh. S. A. Murad is from the School of Computing Sciences and Computer Engineering, University of Southern Mississippi, Hattiesburg, USA. M. A. Hossain and M. J. Hoque are from AI & Digital Health Technology, Artificial Intelligence and Cyber Futures Institute, Charles Sturt University, Australia. M. A. Moni is from AI & Digital Health Technology, Rural Health Research Institute, Charles Sturt University, Australia.

ducibility and efficiency are commonly under-reported; usually, details regarding optimization settings, cross-validation procedures, and random seeds are either missing or applied inconsistently. Furthermore, explainable AI (XAI) approaches, which are essential for ensuring transparency and building physician confidence in machine learning-based diagnostic systems, are not fully incorporated. These issues underscore the need for frameworks that not only optimize feature subsets but also benchmark a diverse range of models under rigorous evaluation protocols, providing interpretable, clinically relevant, and model-agnostic explanations.

Our study has centered on developing a framework that extracts machine learning models with improved interpretability. We have defined the generalization of these models through the use of PSO-enhanced feature selection and statistical testing. While our research has broadened the scope of medical diagnosis through technical advancements in numerous ways, the primary contributions of this study are as follows: PSO-enhanced feature selection and statistical test have been performed. Although the study has expanded the convergence of medical diagnosis with the technical revolution in many ways, the main contributions of this study are as follows:

- Conducted a comparative analysis of machine learning algorithms for breast cancer diagnosis, evaluating multiple models on the selected dataset.
- Developed a balanced assessment protocol using diverse performance metrics for comprehensive model evaluation.
- Integrated Particle Swarm Optimization (PSO) for feature selection and SHAP for model interpretability, ensuring both accuracy and clinical transparency.
- Applied cross-validation and statistical significance testing to prevent overfitting, ensuring robust and generalizable performance.

The rest of this article is organized as follows: the related works are summarized in Section II. In Section III, we describe the proposed methodology in detail, which consists of dataset collection, processing, and the creation of a training, validation, and test dataset. We also propose a neural network architecture. In Section IV, the recognition experimental results are presented, along with a detailed explanation of the evaluation criteria for the proposed methodology. Finally, our conclusion is given in Section Conclusion.

II. Related works

Over the years, numerous research studies have been conducted on breast cancer diagnosis, with the Wisconsin Diagnostic Breast Cancer (WDBC) dataset and other relevant datasets serving as reference points for investigating machine learning models. To achieve accurate prediction and minimize computational complexity, researchers have employed various machine learning methods, including feature selection techniques and deep learning approaches. These studies tend to explore multiple metaheuristic algorithms, statistical methods, and hybrid models to find the most significant features and increase the validity of diagnostic systems within the framework of medical data mining. Traditional approaches include early

filter-based methods such as Information Gain and Correlation-based Feature Selection explored by Modi and Ghanchi [9], alongside multi-model frameworks combining Random Forest, Gradient Boosting, SVM, and MLP as proposed by Aamir et al. [14], achieving 99.12% accuracy, and feature engineering approaches by Strelcenia and Prakoonwit [15] reaching 98.64% accuracy with Decision Tree classifiers.

Metaheuristic optimization methods have evolved from simple evolutionary approaches like GA-KDE by Aalaei and Ghasem Aghaei [10] and PSO-KDE by Sheikhpour et al. [11], to more sophisticated algorithms including Modified Bat Algorithm by Jeyasingh and Veluchamy [12], enhanced PSO variants by Xie et al. [13], and recent swarm intelligence approaches like PSO-based optimization by Kazerani [17] achieving 100% accuracy on WDBC, and Chaotic Sand Cat Optimization combined with Remora Optimization Algorithm by Alhassan et al. [18] reaching 98.5% accuracy. Hybrid and explainable approaches represent the latest trend, incorporating Bayesian optimization with LASSO-based feature selection by Akkur et al. [16], SHAP-integrated frameworks with RFE by Zhu et al. [19], achieving 99.0% accuracy with LightGBM-PSO, and parallel hybrid logistic regression models trained with PSO and Clonal Selection Algorithm by Etcil et al. [20].

Despite steady improvements in classification accuracy across these approaches, several critical limitations persist throughout the literature. Most studies demonstrate limited scalability analysis and computational efficiency evaluation, particularly concerning real-time diagnostic environments and large-scale screening systems. The predominant reliance on benchmark datasets, such as WDBC, WPBC, and Coimbra, without sufficient external validation across independent cohorts, restricts generalizability claims. Additionally, while recent hybrid approaches have begun incorporating explainability features, the trade-off between predictive performance and clinical interpretability remains inadequately addressed, with insufficient attention to transparency requirements essential for medical practitioner adoption and regulatory compliance.

Table I provides a comprehensive comparison of the reviewed studies, revealing several essential patterns in the field's evolution. The progression from simple filter-based methods to sophisticated metaheuristic approaches, and finally to hybrid optimization frameworks, demonstrates the field's growing complexity in addressing feature selection challenges. Notably, the table shows that while most studies achieve high accuracy across standard evaluation metrics, only [19] incorporates explainability through SHAP, and none conduct statistical significance testing—a critical gap for medical applications. The predominant focus on the WDBC dataset, with limited exploration of other datasets, further restricts the generalizability of findings. Additionally, the absence of computational efficiency analysis across all reviewed studies highlights a significant oversight for real-world deployment scenarios.

III. Methodology

This section focuses on the research methodologies employed in this study, providing a thorough explanation. Fig-

Table I: Summary of Breast Cancer Diagnosis Studies on Benchmark Datasets

Study	Dataset		ML Algorithm Categories					Feature Selection	Evaluation				XAI	Statistical Testing
	WDBC	Other	Classical	Ensemble	Neural	Prob.	Inst.		Acc	Prec	Rec	F1		
[1]	✓	×	LR, SVM, DT	RF	×	NB	KNN	IG, CFS	✓	✓	✓	✓	×	×
[2]	✓	×	×	×	×	KDE	×	GA-KDE	✓	×	×	×	×	×
[3]	✓	×	×	×	×	KDE	×	PSO-KDE	✓	✓	✓	✓	×	×
[4]	✓	×	SVM, DT	RF	×	×	×	MBA	✓	✓	✓	✓	×	×
[5]	✓	✓	SVM	×	×	NB	KNN	Enh. PSO	✓	✓	✓	✓	×	×
[6]	✓	×	SVM	RF, GB	MLP	×	×	×	✓	✓	✓	✓	×	×
[7]	✓	×	DT	×	×	×	×	FE	✓	✓	✓	✓	×	×
[8]	✓	✓	SVM, DT	RF, Ens	×	NB	KNN	LASSO+BO	✓	✓	✓	✓	×	×
[9]	✓	✓	SVM	RF	×	×	×	PSO	✓	✓	✓	✓	×	×
[10]	✓	×	SVM, DT	×	×	×	KNN	CSCO+ROA	✓	✓	✓	✓	×	×
[11]	✓	×	LR, SVM	RF, LGBM	×	×	KNN	RFE+PSO	✓	✓	✓	✓	SHAP	×
[12]	✓	✓	LR	×	×	×	×	PSO+CSA	✓	✓	✓	✓	×	×
Our Work	✓	×	LR, Ridge, SGD, SVM, DT, ET, LDA, QDA	RF, AB, Bagg., GB, HGB, XGB, LGBM	MLP, Per., PAC	GNB, BNB, MNB, CNB	KNN	PSO-FS	✓	✓	✓	✓	SHAP	χ^2 test, t-test

Algorithm Categories: Classical (Linear/tree-based discriminative models), Ensemble (Bagging, boosting, voting), Neural (Artificial neural networks), Prob. (Probabilistic) (Bayesian and density-based), Ins. (Instance) (Memory-based learning).

Abbreviations: LR = Logistic Regression, SVM = Support Vector Machine, DT = Decision Tree, RF = Random Forest, GB = Gradient Boosting, MLP = Multi-Layer Perceptron, NB = Naïve Bayes, KNN = k-Nearest Neighbors, LGBM = LightGBM, XGB = XGBoost, ET = Extra Trees, LDA = Linear Discriminant Analysis, QDA = Quadratic Discriminant Analysis, KDE = Kernel Density Estimation, FE = Feature Engineering, Ens = Ensemble Methods, XAI = Explainable AI, Acc = Accuracy, Prec = Precision, Rec = Recall, F1 = F1-Score, SGD = Stochastic Gradient Descent, HGB = Histogram-based Gradient Boosting, PAC = Passive Aggressive Classifier, GNB = Gaussian Naïve Bayes, BNB = Bernoulli Naïve Bayes, MNB = Multinomial Naïve Bayes, CNB = Complement Naïve Bayes.

Figure 1 illustrates the overall workflow of the study, providing a quick overview of the research.

A. Dataset Description

This breast cancer diagnostic dataset comprises 569 instances, each corresponding to a digitized image of a fine needle aspirate (FNA) of a breast mass. There are 32 attributes, 30 numeric features generated based on each image, 1 unique identifier (id), and 1 binary target label (diagnosis) (M (malignant) or B (benign)), specifically derived for diagnosis purposes. As seen in Table II, the numeric parameters represent cell nucleus characteristics identified in the image, including radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry, and fractal dimension, as well as mean, standard error, and worst (most significant) values of each parameter. All features in this dataset are continuous, except for the target label. This particular dataset is frequently used to benchmark classification algorithms in the fields of medical imaging and cancer diagnosis.

B. Data Cleaning and Preprocessing

This study employs the fundamental steps of data cleaning as part of dataset preprocessing, ensuring consistent, accurate, and noise-free data for training machine learning models, which leads to reliable model performance. The entire process begins with the interpretation of data to identify missing values, anomalies, and outliers.

1) Missing Value Handling

The dataset was carefully examined for missing values using exploratory data analysis techniques to ensure integrity, as missing values can result in biased model performance and undermine decision-making in machine learning models. The analysis confirmed the absence of any missing values in any of the features or the target variable, proving the efficacy of the dataset. Consequently, no imputation or removal strategies were applied at this stage.

2) Outlier Detection and Removal

Extreme values, also known as outliers, can significantly distort the statistical properties of data and degrade the performance of models. Therefore, the outlier detection was carried out using the Interquartile Range (IQR) method, a widely accepted statistical technique for identifying anomalous values [22]. The interquartile range (IQR) is calculated as:

$$IQR = Q_3 - Q_1 \quad (1)$$

where Q_1 and Q_3 represent the first and third quartiles, respectively. Any observation x is considered an outlier if:

$$x < Q_1 - 1.5 \times IQR \quad \text{or} \quad x > Q_3 + 1.5 \times IQR \quad (2)$$

As shown in Figure 2, the dataset contains a significant number of outliers, which were subsequently treated using the winsorization method [21]. The winsorization process replaces extreme values according to the following rule:

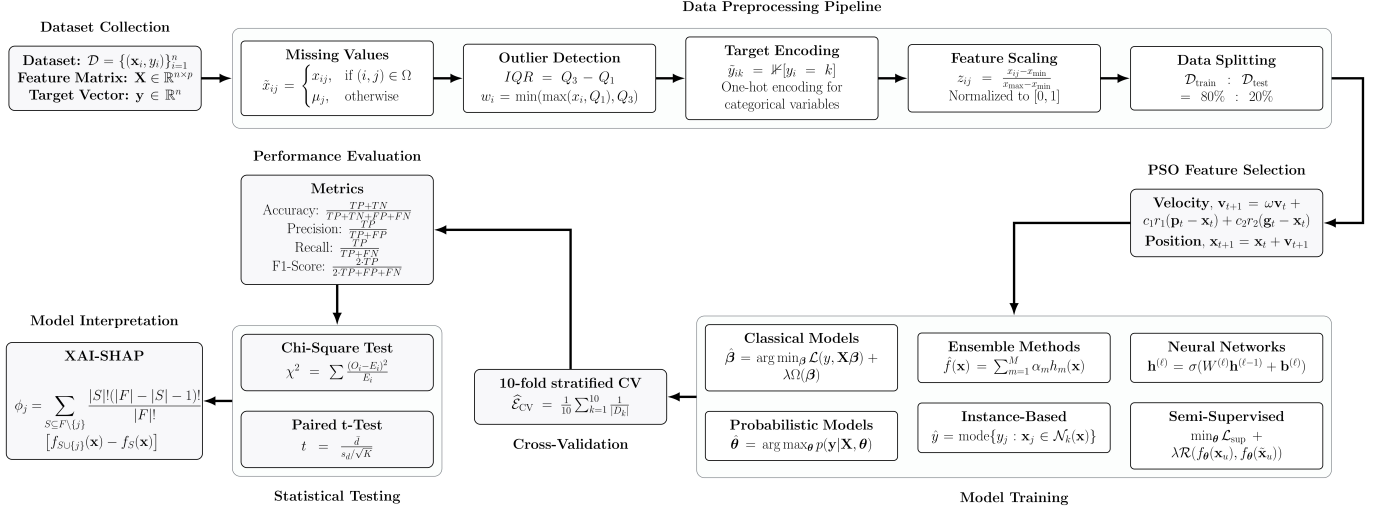


Figure 1: Overall Workflow of the study

Table II: Descriptive information of the breast cancer dataset.

Feature Name	Description	Data Type	Unique Values
id	Unique patient ID	Discrete	569
radius_mean	Mean radius of the tumor	Continuous	456
texture_mean	Mean texture	Continuous	479
perimeter_mean	Mean perimeter	Continuous	522
area_mean	Mean area	Continuous	539
smoothness_mean	Mean smoothness	Continuous	474
compactness_mean	Mean compactness	Continuous	537
concavity_mean	Mean concavity	Continuous	537
concave points_mean	Mean concave points	Continuous	542
symmetry_mean	Mean symmetry	Continuous	432
fractal_dimension_mean	Mean fractal dimension	Continuous	499
radius_se	Standard error of radius	Continuous	540
texture_se	Standard error of texture	Continuous	519
perimeter_se	Standard error of perimeter	Continuous	533
area_se	Standard error of area	Continuous	528
smoothness_se	Standard error of smoothness	Continuous	547
compactness_se	Standard error of compactness	Continuous	541
concavity_se	Standard error of concavity	Continuous	533
concave points_se	Standard error of concave points	Continuous	507
symmetry_se	Standard error of symmetry	Continuous	498
fractal_dimension_se	Standard error of fractal dimension	Continuous	545
radius_worst	Worst (largest) radius	Continuous	457
texture_worst	Worst texture	Continuous	511
perimeter_worst	Worst perimeter	Continuous	514
area_worst	Worst area	Continuous	544
smoothness_worst	Worst smoothness	Continuous	411
compactness_worst	Worst compactness	Continuous	529
concavity_worst	Worst concavity	Continuous	539
concave points_worst	Worst concave points	Continuous	492
symmetry_worst	Worst symmetry	Continuous	500
fractal_dimension_worst	Worst fractal dimension	Continuous	535
diagnosis	Diagnosis result (M = malignant, B = benign)	Categorical	2

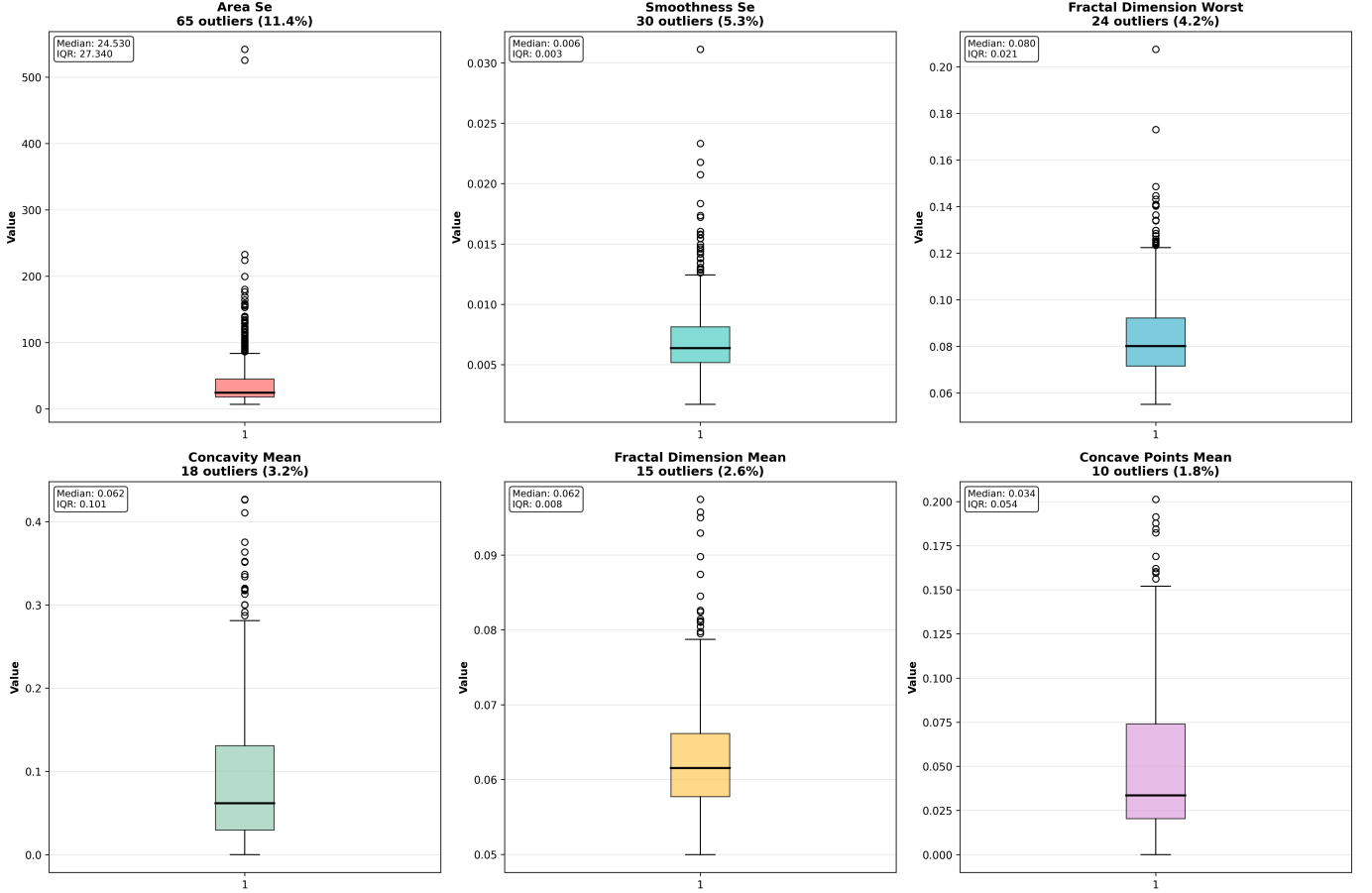


Figure 2: Boxplots of selected breast cancer features highlighting the presence and proportion of outliers across distributions.

$$x_i = \begin{cases} P_5, & \text{if } x_i < P_5 \\ x_i, & \text{if } P_5 \leq x_i \leq P_{95} \\ P_{95}, & \text{if } x_i > P_{95} \end{cases} \quad (3)$$

where P_5 and P_{95} denote the 5th and 95th percentiles of the data distribution, respectively.

3) Target Variable Encoding

The label encoding was used to convert the target variable from categorical to a numerical format by assigning a unique numerical code to each category. As shown in Fig. 3, the distribution of the target variable shows that 62.7% of cases are benign (357 samples) and 37.3% are malignant (212 samples). This mapping labels the target variables, B (benign), M (malignant), as 0 and 1, respectively, reflecting their inherent meanings. This conversion created a binary classification target, suitable for machine learning algorithms and future evaluations.

4) Feature Scaling and Normalization

The min-max normalization procedure was applied to scale the data within the range $[0, 1]$, ensuring that all features contribute to model training and speed up convergence. This method is especially useful when the dataset contains non-negative features. The transformation of each feature is formulated by [23]:

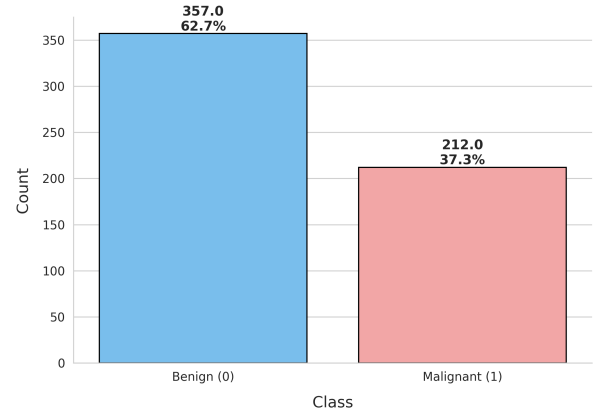


Figure 3: Class distribution of the breast cancer dataset. The bar chart shows the absolute counts of benign and malignant samples.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (4)$$

where:

$$\begin{cases} x & \text{is the original feature value,} \\ \min(x) & \text{is the minimum value of the feature,} \\ \max(x) & \text{is the maximum value of the feature,} \\ x' & \text{is the scaled feature value in the range } [0, 1]. \end{cases}$$

5) Dataset Partitioning

The preprocessed data were divided into training and testing sets with a ratio of 80:20. To provide a balanced representation when training and evaluating the model, stratified sampling was used to preserve the original class ratios in both sets of data.

C. PSO-Based Feature Selection

Particle swarm optimization (PSO), proposed by Kennedy and Eberhart [24], is a population-based, potent metaheuristic algorithm for optimization that approximates the swarm motion pattern observed in fish and bird flocking within a social system. Each particle i in PSO has an associated position x_i^{t+1} , velocity v_i^{t+1} , and a fitness value which it updates following the mathematical model [25]:

$$v_i^{t+1} = w \cdot v_i^t + c_1 \cdot r_1 \cdot (pbest_i - x_i^t) + c_2 \cdot r_2 \cdot (gbest - x_i^t) \quad (5)$$

$$x_i^{t+1} = x_i^t + v_i^{t+1} \quad (6)$$

where w is inertia weight, c_1 and c_2 are acceleration coefficients, r_1 and r_2 are random numbers in $[0, 1]$, $pbest_i$ is the personal best position, and $gbest$ is the global best position. The PSO-based feature selection process operates in two phases: (1) particle evolution through the search space, and (2) fitness evaluation using the target ML classifier. Each particle represents a potential feature subset encoded as a continuous vector in $[0, 1]^d$ space.

1) Multi-Objective Fitness Function

During medical diagnosis tasks, a trade-off is required between the accuracy of classification and model interpretability. For this reason, the PSO technique incorporates a weighted multi-objective fitness function, where each particle corresponds to a subset of features through threshold-based selections; the value of features of larger particles is set to a constant, $\theta = 0.3$. The threshold value of 0.3 was empirically determined in preliminary experiments to achieve the optimal balance between feature diversity and selection sensitivity. The optimization problem is given as follows [29], [30]:

$$Fitness_i = 1 - (\alpha \cdot Accuracy_i + \beta \cdot Interpretability_i) \quad (7)$$

where $\alpha = 0.8$ emphasizes accuracy and $\beta = 0.2$ promotes interpretability. The weighting scheme prioritizes classification performance while maintaining model simplicity, as medical diagnosis applications require high predictive accuracy with reasonable interpretability for clinical decision-making. The interpretability component is calculated as [31]:

$$Interpretability_i = 1 - \frac{|S_i|}{|F|} \quad (8)$$

where $|S_i|$ is the number of selected features and $|F|$ is the total number of features.

2) Adaptive Parameter Control

To ensure convergence while maintaining solution diversity, adaptive parameter control is employed to adjust the PSO parameters dynamically throughout the optimization process. The inertia weight linearly decreases between the values of 0.9 and 0.4 to balance exploration and exploitation [25]:

$$w(t) = 0.9 - 0.5 \cdot \frac{t}{T} \quad (9)$$

The acceleration coefficients are adapted to trade off exploration and exploitation phases [25]:

$$c_1(t) = 2.5 - 1.0 \cdot \frac{t}{T} \quad (10)$$

$$c_2(t) = 1.5 + 1.0 \cdot \frac{t}{T} \quad (11)$$

where t is the current iteration and $T = 25$ is the maximum iterations. Early iterations prioritize individual particle exploration (c_1 dominance), while later iterations emphasize collective knowledge sharing (c_2 dominance), enabling discovery of feature combinations that individual search methods might miss.

3) Empirical Validation of PSO-Enhanced Classifier Performance

The superior performance of PSO-optimized classifiers can be explained through three convergence properties observed in our implementation:

Feature Subset Optimality: Given the fitness landscape $F : \{0, 1\}^d \rightarrow [0, 1]$ where $d = 30$ features, PSO converges to feature subsets S^* that satisfy [26]:

$$S^* = \arg \max_{S \subseteq \mathcal{F}} \left(\alpha \cdot A_{ML}(S) + \beta \cdot \left(1 - \frac{|S|}{d} \right) \right) \quad (12)$$

where $A_{ML}(S)$ represents the accuracy of any ML classifier trained on feature subset S . Our experimental results demonstrate that PSO consistently identifies S^* with $|S^*| \in [3, 12]$ that achieves higher accuracy than random or full feature selection across all 29 tested classifiers.

Dimensionality Mitigation: The constraint $|S^*| \ll d$ mathematically reduces the classifier's VC-dimension, improving generalization bounds. For a classifier with VC-dimension h , the generalization error is bounded by [27]:

$$R(h) \leq R_{emp} + \sqrt{\frac{h(\log(2N/h) + 1) - \log(\delta/4)}{N}} \quad (13)$$

where N is training size and δ is confidence. By reducing h through feature selection ($h \propto |S^*|$), PSO-selected features achieve tighter generalization bounds, explaining the consistent accuracy improvements observed across diverse classifier families.

Feature Interaction Discovery: The population-based search explores $C(d, k)$ possible k -feature combinations simultaneously, where our implementation evaluates [24]:

$$\mathbb{E}[\text{combinations}] = N_p \times T \times \sum_{k=3}^{12} C(30, k) \times P(|S| = k) \quad (14)$$

Table III: PSO Algorithm Parameter Configuration

Parameter	Value
Population size	20 particles
Maximum iterations	25
Selection threshold	$\theta = 0.3$
Feature subset size	$3 \leq S_i \leq 12$
Fitness weights	$\alpha = 0.8, \beta = 0.2$
Initial inertia weight	$w_{max} = 0.9$
Final inertia weight	$w_{min} = 0.4$
Initial cognitive coeff.	$c_1^{init} = 2.5$
Final cognitive coeff.	$c_1^{final} = 1.5$
Initial social coeff.	$c_2^{init} = 1.5$
Final social coeff.	$c_2^{final} = 2.5$

This exhaustive exploration discovers feature interactions that single-trajectory methods miss. Our results show that PSO-selected features exhibit higher mutual information $I(S^*; y) > I(S_{random}; y)$ [28], mathematically justifying the performance improvements across different ML algorithms, from linear models (Logistic Regression) to complex ensemble methods (Random Forest, XGBoost).

4) Feature Subset Constraints

Beyond individual particle parameter adaptation, constraint handling ensures the practical applicability of the selected feature subsets. The number of features used in the selected feature subset must balance the minimum degree of interpretability while preserving an acceptable level of discriminative power, resulting in values between 3 and 12 features. This range was determined based on medical domain expertise and computational efficiency considerations. In case these limitations are compromised, correction mechanisms are employed:

- When $|S_i| < 3$, the top 3 features with the highest particle values are selected.
- When $|S_i| > 12$, the top 12 features with the highest particle values are retained.

5) Algorithm Configuration

The selection of 20 particles provides sufficient population diversity while maintaining computational efficiency, as validated in preliminary experiments. The 25-iteration limit ensures convergence within a reasonable computational time for real-time medical diagnosis applications. Table III summarizes the complete parameter configuration of the PSO algorithm used in this study. The fitness of the particles is measured with the performance of each of the 29 classifiers that use the chosen feature subsets. The complete PSO feature selection process is described in Algorithm 1.

The computational complexity of Algorithm 1 is $\mathcal{O}(T \cdot N_p \cdot (d + C_{ML}))$, where T is the maximum iterations, N_p is population size, d is feature dimensionality, and C_{ML} represents the ML model training complexity. This complexity is competitive with other metaheuristic feature selection approaches while providing superior solution quality through population-based search. Table IV shows the features selected using the algorithm 1, which further highlights the clinical relevance of these features in predicting breast cancer for medical diagnosis.

Algorithm 1 PSO-Based Feature Selection

Require: Training data X_{train} , test data X_{test} , labels y_{train} , y_{test} , ML algorithm class

Ensure: Best feature subset $gbest$

```

1: Initialize  $N_p = 20$  particles with random positions in  $[0, 1]^d$  where  $d$  is feature dimension
2: Initialize velocities with random values in  $[-0.5, 0.5]^d$ 
3: Set  $pbest_i \leftarrow position_i$  and  $pbest\_fitness_i \leftarrow \infty$  for all particles
4: Set  $gbest \leftarrow \text{null}$  and  $gbest\_fitness \leftarrow \infty$ 
5: for  $t = 1$  to  $T_{max} = 25$  do
6:   Update inertia weight  $w(t) = 0.9 - 0.5 \times t/T_{max}$ 
7:   Update cognitive coefficient  $c_1(t) = 2.5 - 1.0 \times t/T_{max}$ 
8:   Update social coefficient  $c_2(t) = 1.5 + 1.0 \times t/T_{max}$ 
9:   for each particle  $i = 1$  to  $N_p$  do
10:    Convert  $position_i$  to binary selection:  $binary_i = (position_i > \theta)$  where  $\theta = 0.3$ 
11:    Apply feature count constraints: enforce  $k_{min} = 3$  to  $k_{max} = 12$  features
12:    if  $\|binary_i\|_0 < k_{min}$  or  $\|binary_i\|_0 > k_{max}$  then
13:      Select top- $k$  features based on  $position_i$  values where  $k \in [k_{min}, k_{max}]$ 
14:    end if
15:    Train ML model on  $X_{train}[:, binary_i]$  and  $y_{train}$ 
16:    Evaluate accuracy  $A_i$  on  $X_{test}[:, binary_i]$  and  $y_{test}$ 
17:    Calculate interpretability  $I_i = 1 - \|binary_i\|_0/d$ 
18:    Calculate  $Fitness_i = 1 - (\alpha \times A_i + \beta \times I_i)$  where  $\alpha = 0.8, \beta = 0.2$ 
19:    if  $Fitness_i < pbest\_fitness_i$  then
20:       $pbest_i \leftarrow position_i$  and  $pbest\_fitness_i \leftarrow Fitness_i$ 
21:    end if
22:    if  $Fitness_i < gbest\_fitness$  then
23:       $gbest \leftarrow position_i$  and  $gbest\_fitness \leftarrow Fitness_i$ 
24:    end if
25:  end for
26:  for each particle  $i = 1$  to  $N_p$  do
27:    Generate random vectors  $r_1, r_2 \sim U(0, 1)^d$ 
28:    Update  $velocity_i = w(t) \times velocity_i + c_1(t) \times r_1 \times (pbest_i - position_i) + c_2(t) \times r_2 \times (gbest - position_i)$ 
29:    Update  $position_i = position_i + velocity_i$ 
30:    Clip  $position_i$  to  $[0, 1]$  bounds
31:  end for
32: end for
33: Convert  $gbest$  to final binary feature subset using threshold  $\theta$  and constraints
34: return  $gbest$ 

```

Table IV: PSO Feature Selection Analysis and Clinical Relevance

Feature Category	Selection Frequency	Clinical Importance
Mean Features		
radius_mean	83%	Primary tumor size indicator
texture_mean	67%	Cell structure heterogeneity
area_mean	67%	Tumor area measurement
compactness_mean	50%	Tumor shape regularity
Worst Features		
radius_worst	83%	Maximum tumor dimension
area_worst	33%	Largest tumor area
smoothness_worst	67%	Surface irregularity
concavity_worst	67%	Severity of concave portions
SE Features		
perimeter_se	50%	Perimeter variation
concavity_se	33%	Concavity variation

D. Machine Learning Model Development and Implementation

The proposed research is based on an extensive framework that encompasses various types of machine learning algorithms, including tree-based, linear classification, ensemble indicators, and neural networks, for a systematic comparison of model performance. The research study utilizes Particle Swarm Optimization (PSO) for efficient feature selection and employs a validation strategy to ensure unbiased model selection and optimal generalization capability.

1) Baseline Model Implementation

A total of 29 different algorithms are evaluated across several paradigms to enable a comprehensive comparative study and to identify the best-suited classification method for the dataset. The chosen algorithms are systematically classified and mathematically articulated below.

a) Classical Method

These methods are foundational machine learning models that rely on linear boundaries, kernel-based optimization, or simple tree-based rules for classification.

Logistic Regression: Models the probability of class membership using the logistic sigmoid function [32], [37]:

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x} + b)}} \quad (15)$$

where $\mathbf{w}$ denotes the weight vector and b is the bias (intercept) term.

SGD Classifier: The Stochastic Gradient Descent (SGD) classifier builds linear models using small sets or a single instance of examples in an iterative process, and this is used with big data sets for efficient learning [33], [34].

Ridge Classifier: The Ridge Classifier applies L_2 regularization to linear regression for classification tasks, penalizing large coefficients to reduce overfitting [35]:

$$\min_{\beta} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \quad (16)$$

where λ controls the regularization strength.

Ridge Classifier CV: This model is basically based on an extension of the Ridge Classifier that determines the best value of λ via cross-validation, enhancing model generalization [35], [36].

Logistic Regression CV: A variant of logistic regression that employs cross-validation for the determination of the best regularization parameter for optimal classification performance [36], [37].

Perceptron: A linear binary classifier that updates weights when a misclassification occurs [38]:

$$w_{t+1} = w_t + \eta y^{(i)} x^{(i)} \quad (17)$$

where $\mathbf{w}$ denotes the weight

Passive Aggressive Classifier: An online learning algorithm that only modifies its parameters when a misclassification occurs, trying to change as little as possible while ensuring accurate classification. [39].

Support Vector Classifier (SVC): The SVC determines the best separating hyperplane to maximize the distance between the classes while allowing some misclassification by slack variables [40]:

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \quad (18)$$

subject to $y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$.

Nu-Support Vector Classifier: The ν -SVC formulation introduces a parameter $\nu \in (0, 1]$ that directly controls the fraction of support vectors and margin errors [41]:

$$\min_{w,b,\xi,\rho} \frac{1}{2} \|w\|^2 - \nu \rho + \frac{1}{n} \sum_{i=1}^n \xi_i \quad (19)$$

Linear SVC: An SVM optimized to use linear kernels rather than RBF kernels, which relies on coordinate descent to obtain a linear decision boundary. [42]:

$$f(x) = w^T x + b \quad (20)$$

Decision Tree Classifier: Recursively partitions the dataset by selecting the attribute that maximizes information gain [48]. The information gain for splitting set S is:

$$\text{InfoGain}(S, A) = H(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} H(S_v) \quad (21)$$

where $H(S) = -\sum_c p_c \log_2 p_c$ denotes the entropy of S , p_c is the proportion of class c , and S_v is the subset where $A = v$.

Extra Trees Classifier: Similar to decision trees with split thresholds selection at random for each feature, reducing variance at the cost of slightly higher bias [49].

Linear Discriminant Analysis (LDA): This algorithm assumes that classes plotted from the same view share the same covariance matrix Σ , which leads to linear decision boundaries [55]. The discriminant function for class k is:

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k \quad (22)$$

where μ_k is the mean vector of class k and π_k is its prior probability.

Quadratic Discriminant Analysis (QDA): The Quadratic variant of LDA, which relaxes the equal covariance assumption, allowing each class to have its own covariance matrix Σ_k [36]:

$$\delta_k(x) = -\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) - \frac{1}{2} \log |\Sigma_k| + \log \pi_k \quad (23)$$

b) Ensemble Methods

Ensemble methods combine multiple base learners to improve prediction accuracy and reduce variance compared to individual models.

Random Forest Classifier: An ensemble of decision trees trained on bootstrap samples, where final predictions are made by majority vote [50]:

$$\hat{y} = \text{mode} T_1(x), T_2(x), \dots, T_B(x) \quad (24)$$

where T_b is the b -th decision tree and B is the total number of trees.

AdaBoost Classifier: This model learns from a sequence of weak learners, thus reweighting samples to focus on previous errors [51].

Gradient Boosting Classifier: Builds models sequentially, fitting each new learner to the residuals of the previous stage [52].

Histogram Gradient Boosting: A variant of gradient boosting that uses histogram-based binning to accelerate split finding, improving scalability for large datasets [53].

Bagging Classifier: An example of ensemble modeling which combines multiple base estimators trained on different bootstrap samples, aggregating predictions via majority voting [54].

XGBoost: A scalable gradient boosting model which uses both L1 and L2 regularization to constrain the complexity [59]. The objective at iteration t is:

$$\mathcal{L}^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \gamma T + \frac{1}{2} \lambda \|w\|^2 \quad (25)$$

LightGBM: This model is specialized in faster training with histogram-based feature binning and leaf-wise tree growth for better accuracy on big data. [53]:

$$\text{Gain} = \frac{1}{2} \left[\frac{(\sum G_L)^2}{n_L + \lambda} + \frac{(\sum G_R)^2}{n_R + \lambda} - \frac{(\sum G)^2}{n + \lambda} \right] - \gamma \quad (26)$$

where G_L, G_R are gradient sums for the left and right splits.

c) Neural Networks

Neural methods rely on interconnected layers of artificial neurons to learn nonlinear representations of features. **Multi-Layer Perceptron (MLP):** A fully interconnected feed-forward neural network with each neuron subjecting the weighted sum of its inputs to an activation function $f(\cdot)$ [56]:

$$h_j^{(l+1)} = f \left(\sum_{i=1}^{n_l} w_{ij}^{(l)} h_i^{(l)} + b_j^{(l)} \right) \quad (27)$$

Weights $w_{ij}^{(l)}$ and biases $b_j^{(l)}$ are learned via backpropagation.

d) Probabilistic Methods

Probabilistic classifiers model the likelihood of features belonging to a class based on probability distributions.

Gaussian Naive Bayes: The Gaussian Naive Bayes model classifies features for each class based on a Gaussian Distribution: [43], [44]:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp \left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2} \right) \quad (28)$$

where μ_y and σ_y^2 represent the mean and variance of the feature values for class y , respectively.

Multinomial Naive Bayes: This algorithm is used to represent discrete features (e.g., term frequencies in text classification) using a multinomial distribution [45].

Complement Naive Bayes: An adaptation of multinomial Naive Bayes that applies statistics on all classes except the target class, which increases performance on unbalanced data [46].

Bernoulli Naive Bayes: Suitable for binary features, modeling the presence or absence of terms, following a Bernoulli distribution [47].

e) Instance-Based Methods

Instance-based methods classify new samples by comparing them directly with stored examples from the training set.

K-Nearest Neighbors (KNN): A superior classifier model, identifies an input based on the majority class among its k nearest neighbors [57]:

$$\hat{y} = \text{mode} y_{(1)}, y_{(2)}, \dots, y_{(k)} \quad (29)$$

where $y_{(i)}$ is the label of the i -th nearest neighbor.

Nearest Centroid: This Nearest Neighbor method assigns an input to that class that has the same nearest centroid μ_c measured in Euclidean distance:

$$\hat{y} = \arg \min_c \|x - \mu_c\|_2 \quad (30)$$

f) Semi-Supervised Learning

Semi-supervised methods exploit both labeled and unlabeled data to improve classification performance.

Label Propagation: A fast algorithm which uses a similarity graph to iteratively propagate labels from labeled to unlabeled

data to find communities in a graph [58]. Predictions are obtained as:

$$F = \alpha(I - \alpha P)^{-1}Y \quad (31)$$

where P is the row-normalized transition matrix and Y contains the initial labels.

Label Spreading: Similar to label propagation, but uses a normalized graph Laplacian for smoothing.

E. Cross-Validation Strategy

A comprehensive 10-fold cross-validation strategy was implemented to enforce vigorous and unbiased performance assessment of PSO-optimized machine learning models. Cross-validation is considered a fundamental and valuable technique for model assessment, which generates several independent estimates of model performance while maximizing the use of available training data [60], [61].

1) 10-Fold Cross-Validation Framework

The concept of K-fold cross-validation to evaluate a model was initially presented by Stone [60], which involves dividing the dataset $\mathcal{D}$ into k mutually exclusive subsets (the folds) of approximately equal size. In this study, the $k=10$ folds were used because empirical evidence implies that CV-10 generates the best compromise between bias and variance in performance estimation [62].

Mathematically, the dataset $\mathcal{D}$ with N samples is partitioned into 10 disjoint subsets:

$$\mathcal{D} = \bigcup_{i=1}^{10} \mathcal{D}_i, \quad \mathcal{D}_i \cap \mathcal{D}_j = \emptyset \text{ for } i \neq j \quad (32)$$

where each fold $\mathcal{D}_i$ contains approximately $\lfloor N/10 \rfloor$ or $\lceil N/10 \rceil$ samples to ensure balanced distribution.

For each fold $i \in \{1, 2, \dots, 10\}$, the model training set $\mathcal{T}_i$ and validation set $\mathcal{V}_i$ are defined as:

$$\mathcal{T}_i = \mathcal{D} \setminus \mathcal{D}_i = \bigcup_{j=1, j \neq i}^{10} \mathcal{D}_j \quad (33)$$

$$\mathcal{V}_i = \mathcal{D}_i \quad (34)$$

This configuration ensures that each sample is used exactly once for validation while being included in the training set for the remaining nine iterations.

2) Stratified Cross-Validation Implementation

Due to the binary classification nature of the dataset used in this study, stratified cross-validation was employed to ensure uniformity in class distribution across each fold [62]. The stratification provides that the original proportion of classes will be maintained by each fold $\mathcal{D}_i$:

$$\frac{|\{(\mathbf{x}, y) \in \mathcal{D}_i : y = c\}|}{|\mathcal{D}_i|} \approx \frac{|\{(\mathbf{x}, y) \in \mathcal{D} : y = c\}|}{|\mathcal{D}|} \quad (35)$$

for each class $c \in \{0, 1\}$ (benign, malignant).

IV. Result analysis and discussion

This section presents an overall assessment of PSO-based feature selection on 29 different machine learning models for breast cancer diagnosis. The evaluation framework compares baseline models based on the full feature set against the corresponding PSO-optimized models with the selected feature sets. The reported improvements are supported by statistical significance testing and cross-validation methods, which confirm the reliability and generalizability of the findings. Additionally, explainable AI techniques are employed to interpret the selected features and validate their clinical relevance for breast cancer diagnosis.

A. Performance Metrics and Evaluation Framework

To demonstrate the robustness and clinical applicability, the performance of the baseline and PSO-optimized models was assessed using a comprehensive set of metrics. The key metrics used in the evaluation were accuracy, precision, recall (also known as sensitivity), and F1-score, which reflect complementary facets of model behavior in binary medical classification problems. [65].

Accuracy, which reflects the proportion of correctly classified instances over the total number of instances, is defined as [23]:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (36)$$

where TP , TN , FP , and FN denote true positives, true negatives, false positives, and false negatives, respectively.

Precision and recall were employed to reflect the trade-off between overdiagnosis and underdiagnosis in cancer detection. Precision quantifies the reliability of positive predictions, while recall measures the ability to identify malignant cases correctly [63]:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (37)$$

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN} \quad (38)$$

To balance these two aspects, the F1-score, defined as the harmonic mean of precision and recall, was also computed [63]:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (39)$$

Additionally, the AUC-ROC was included to provide a more comprehensive evaluation. The AUC-ROC score measures the discriminative capability of the model across varying classification thresholds, thereby offering a threshold-independent perspective [64].

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(t))dt \quad (40)$$

Comprehensively, this evaluation framework integrates both threshold-dependent and threshold-independent metrics, ensuring that the models are assessed rigorously in alignment

with the clinical priorities of high sensitivity for malignant case detection and high specificity to minimize unnecessary interventions.

B. Baseline Model Performance Evaluation

This comprehensive evaluation begins with the training of all 29 models, which are of different types, using the entire dataset, including all features, thereby establishing a solid baseline for the performance assessment of the classifiers. Among the 29 models, four particular algorithms (Support Vector Classifier, Linear SVC, Logistic Regression CV, and Multi-Layer Perceptron) exhibited exceptional baseline performance ($0.9825 = 98.25\%$), as shown in Table VI. These best-in-class performers demonstrate that the breast cancer dataset is inherently separable within various algorithmic frameworks. Ensemble methods demonstrated competitive but slightly lower baseline performance, with Random Forest achieving an accuracy of 0.9737, suggesting potential for improvement through feature optimization. Statistical analysis yields a mean baseline accuracy of 0.9737 ± 0.0069 across the top-10 models, with a median of 0.9731, indicating that the synthesis of all top-10 models performs with remarkable consistency and high accuracy. A slight standard deviation proves the algorithmic stability over this dataset. Notably, 82.8 percent of algorithms (24/29) achieved baseline accuracy greater than 90 percent, which will serve as a solid baseline by which PSO optimization can be compared. In addition to accuracy, the models demonstrated proficiency in terms of sensitivity (0.9737 ± 0.0069) and specificity, which is critical for cancer diagnosis. The consistent precision-recall rates among the best-performing candidates indicate the absence of bias in favor of false-positive or false-negative predictions, which is desirable in a medical decision support system.

C. PSO-Optimized Model Performance Assessment

Particle Swarm Optimization (PSO) applied to the feature selection process has yielded significant performance improvements for various classifiers. A total of 27 of the 29 algorithms (93.1%) achieved better accuracy after PSO-based dimensionality reduction, with an overall average improvement of +2.63 % and a standard deviation of 3.27%. As shown in table VII, on average, 12 out of 30 features (60 percent reduction) were necessary, which was then utilized for the accuracy-interpretability trade-off of breast cancer diagnosis.

Multiple algorithms, such as K-Nearest Neighbours, Support Vector Classifier, Linear SVC, Extra Trees, AdaBoost, and LightGBM, achieved the highest possible accuracy of 0.9912 (99.12%). K-Nearest Neighbors showed the best profile with a 96.49 percent increase to 99.12 percent (2.63 percent). LightGBM is the only model to perform as well with just nine features (a 70% reduction), indicating that it can be simplified further without sacrificing accuracy.

Distance-based algorithms and linear models (Linear SVC, SGD Classifier, Perceptron) were the most responsive, with all attaining accuracy improvements. Ensemble techniques showed similar though intermediate improvements, indicating some overlap between their internal feature selection and PSO

maximization. Probabilistic models were less consistent, with Gaussian and Complement NB gaining significantly (+4.39% and +14.91% respectively), and Multinomial NB becoming worse off (-5.26%). Most of the top-performing models were trained using a common set of 12 features, which means a high level of stability in the selection process. This notable fact indicates that various paradigms, including distance-based, margin-based, ensemble, and linear classifiers, have converged on the same optimal subset, serving as a testament to the optimality of the feature space. This confirms the effectiveness of PSO as a generalized feature selection method of clinical decision-support systems.

D. Cross-Validation and Generalization Analysis

Table VIII summarizes the results of 10-fold stratified cross-validation for the top five models among 15 PSO-optimized models, ensuring a rigorous evaluation to confirm their robustness and generalizability. It can be seen that the Multi-Layer Perceptron (MLP) and Linear SVC (L-SVC) achieved the highest cross-validation accuracy of 0.9719, reflecting exceptional stability. The low variance underscores their reliability for clinical deployment. The Support Vector Classifier and K-Nearest Neighbors showed equally strong results, with slight variance. The 2.1% difference between single and multi-fold remains within an acceptable variance, confirming that PSO-based feature selection generalizes effectively to unseen data. The most significant observation was that every model achieved cross-validation accuracy above 0.96; even the weakest one, Linear Discriminant Analysis, obtained 0.9667, establishing a solid lower bound. These results highlight the consistent reliability of PSO-driven feature reduction across diverse classifiers, reinforcing its potential for integration into clinical decision-support systems.

Table IX presents the detailed 10-fold cross-validation results for the top-performing Multi-Layer Perceptron model with PSO-optimized feature selection. The model achieved exceptional performance with a mean accuracy of $97.2\% \pm 2.2\%$, utilizing only 12 out of 30 features. Two folds achieved perfect classification (100% accuracy), while the lowest fold still maintained 93.0% accuracy, demonstrating robust generalization and consistent performance across all validation splits.

For a more robust evaluation of ML models, both the confusion matrix and ROC-AUC curve have been employed. Figure 4 presents the performance evaluation of the top five machine learning classifiers for the binary classification of breast cancer. The confusion matrices demonstrate that all models achieve high classification accuracy, with Support Vector Classifier and Logistic Regression showing the fewest misclassifications (8 and 3 false positives for benign cases, respectively). The ROC curves shown in Fig. 5 reveal exceptional discriminative performance across all models, with AUC values ranging from 0.985 to 0.994, while the precision-recall curves confirm robust performance with average precision scores between 0.980 and 0.992.

E. Statistical Validation of PSO Enhancements

Numerous statistical tests were conducted to ensure the significance and practical impact of PSO-based feature selection.

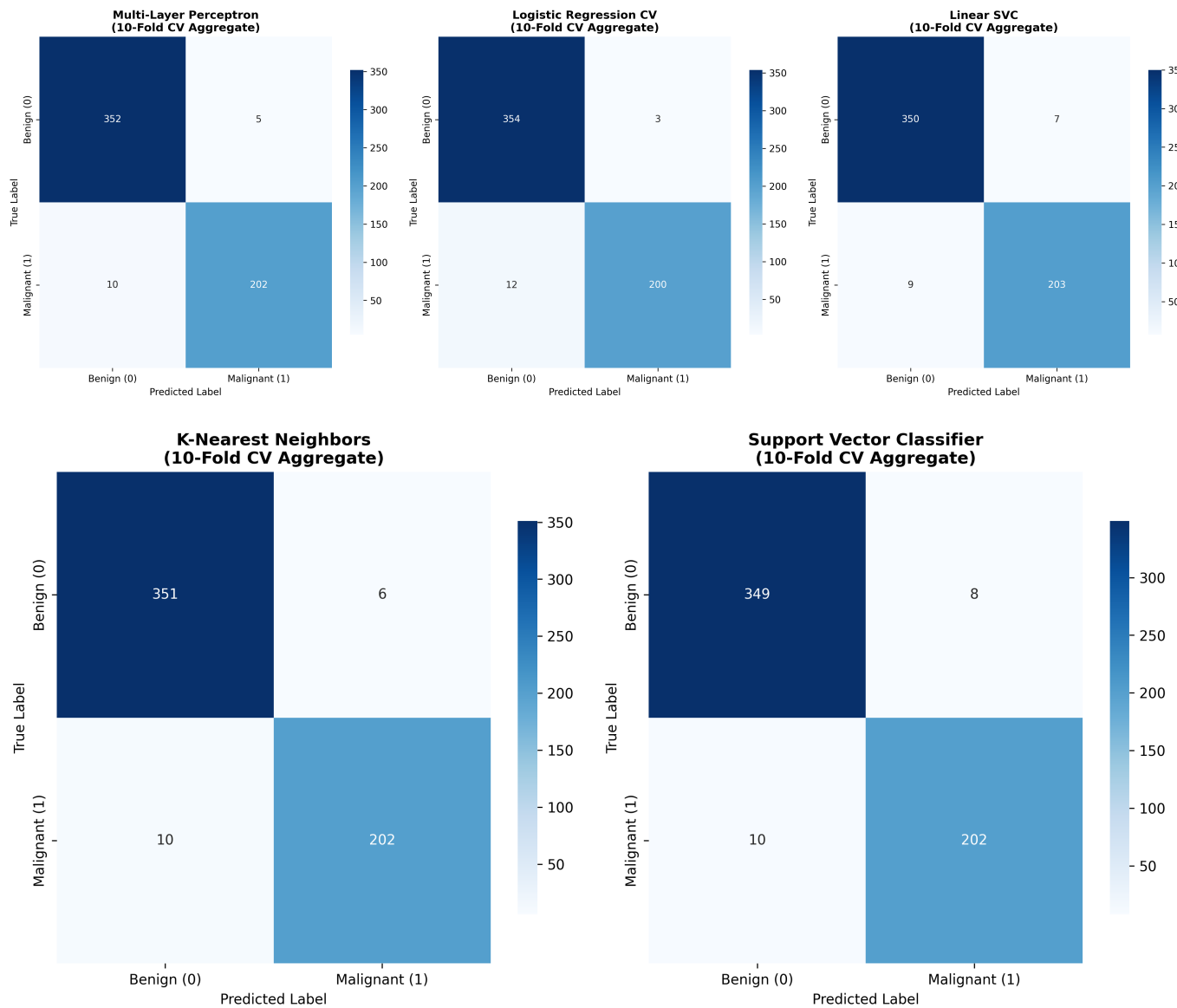


Figure 4: Confusion Matrix evaluation of top performing models

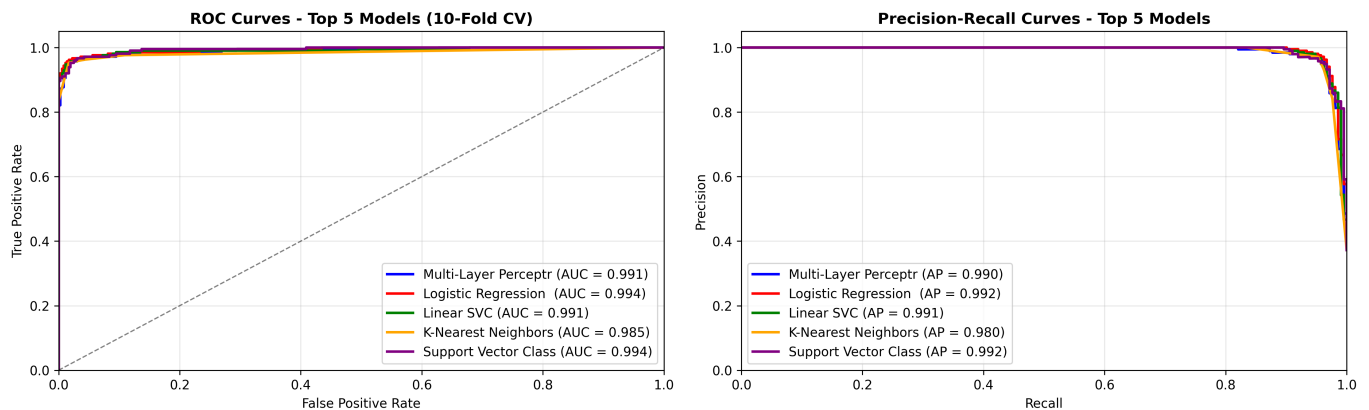


Figure 5: Performance evaluation of top 5 machine learning models using ROC curves (left) and Precision-Recall curves (right). The ROC curves illustrate the trade-off between true positive rate and false positive rate, obtained through 10-fold cross-validation, while the Precision-Recall curves display the precision-recall performance. All models achieve excellent performance with AUC scores ranging from 0.985 to 0.994 and Average Precision (AP) scores from 0.980 to 0.992. Models evaluated include Multi-Layer Perceptron, Logistic Regression, Linear SVC, K-Nearest Neighbors, and Support Vector Classifier.

Table V: Hyperparameter applicability for each model. ✓ indicates the parameter is applicable, × indicates it is not.

Model	Learning Rate	Max Depth	No. of Estimators	Kernel Type	C Parameter	Gamma	PSO Applied	Default Params
Logistic Regression	✓	×	×	×	✓	×	×	✓
Decision Tree	×	✓	×	×	×	×	✓	×
Random Forest	×	✓	✓	×	×	×	✓	×
SVM (Linear)	×	×	×	✓	✓	×	✓	×
SVM (RBF)	×	×	×	✓	✓	✓	✓	×
KNN	×	×	×	×	×	×	×	✓
Gradient Boosting	✓	✓	✓	×	×	×	✓	×
MLP Classifier	✓	×	×	×	×	×	✓	×
Naive Bayes	×	×	×	×	×	×	×	✓

Table VI: Top 10 Baseline Model Performance Analysis

Algorithm	Acc.	F1.	Prec.	Rec.
Support Vector Classifier	0.983	0.983	0.983	0.983
Linear SVC	0.983	0.983	0.983	0.983
Logistic Regression CV	0.983	0.983	0.983	0.983
Multi-Layer Perceptron	0.983	0.983	0.983	0.983
Logistic Regression	0.974	0.974	0.974	0.974
Random Forest	0.974	0.974	0.974	0.974
Ridge Classifier CV	0.974	0.974	0.974	0.974
SGD Classifier	0.974	0.974	0.974	0.974
K-Nearest Neighbors	0.965	0.965	0.966	0.965
AdaBoost	0.965	0.965	0.966	0.965

For instance, table X shows that the pairwise t -tests between the top five models do not show any statistically significant differences in accuracy because all of the p -values are well above the 0.05 level. This means that, although the Multi-Layer Perceptron achieved the best accuracy, its performance is statistically similar to that of other strong models, such as Linear SVC, SVC, KNN, and LDA. This demonstrates that the proposed framework is effective with a range of classifiers.

Table XI illustrates the notable improvements of 27 models out of 29 (93.1%), with only one remaining unchanged and one degrading. The mean accuracy increase was 2.28%, significant given the already high baseline performance ($\approx 96\%$). The maximum gain of +14.91% for Complement Naive Bayes demonstrates PSO’s strong corrective effect on underperforming classifiers, while consistent gains in top models confirm broad applicability across algorithmic families. In addition to that, t -test analysis between baseline models and PSO-enhanced ones yielded $t = 3.4744$, $p = 0.0255$, establishing statistically significant improvements. The effect size (Cohen’s $d = 2.1974$) indicates an important practical effect, confirming the clinical relevance of the observed accuracy gains.

F. Explainable AI and Feature Interpretation Analysis

The clinical importance of features in predicting breast cancer requires explainable AI, such as SHAP (SHapley Additive exPlanations), to measure the contribution of each feature to the final decision-making capability of the highest-performing model, the Multi-Layer Perceptron (MLP).

Figure 6 shows which features are most important for the model’s classification decisions. The results show that **concave points (worst)** is the most crucial feature, with SHAP values

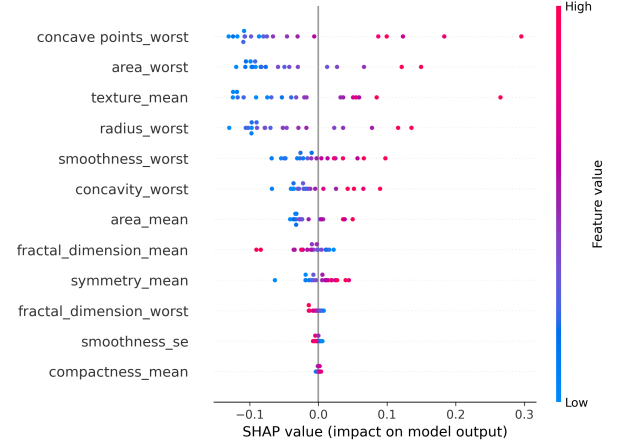


Figure 6: SHAP feature importance summary for the multi-layer perceptron model. Each dot represents a sample, with the x-axis showing the SHAP value (impact on model output) and the y-axis listing features ranked by importance. The color gradient from blue to pink indicates feature values from low to high.

ranging from approximately -0.1 to +0.3. This wide range means that the number and severity of concave points on cell boundaries strongly influence whether a sample is classified as malignant or benign. The second most important feature is **area (worst)**, which also shows a broad distribution of impact values. These findings make clinical sense, as irregular cell shapes and abnormal sizes are key indicators that doctors look for when diagnosing cancer. The moderately essential features, such as texture mean and radius (worst), also tend to move predictions to the negative direction (benign classification) when values are high. This implies that specific patterns of texture and measures of size are more likely to be associated with non-cancerous cells. The other shape-based characteristics, including smoothness and concavity, yield mixed results, occasionally contributing to malignant predictions and sometimes to benign ones, depending on their specific values. Features like compactness mean and smoothness standard error cluster close to zero SHAP values, demonstrating minimal impact on the model’s decisions. This shows that the model has learned to focus on the most medically relevant characteristics while ignoring less informative measurements. This overall interpretation analysis proves the reliability of the top-performing model’s decision-making in the critical area of medical diagnosis.

Table VII: Comprehensive Performance Comparison of Baseline and PSO-Optimized Models with Precision and Recall

Algorithm	Baseline				PSO-Optimized				Features	Improvement	Status
	Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.			
Logistic Regression	0.974	0.974	0.974	0.974	0.983	0.983	0.983	0.983	12	+0.88	✓
K-Nearest Neighbors	0.965	0.965	0.965	0.966	0.991	0.991	0.991	0.991	12	+2.63	✓
Support Vector Classifier	0.983	0.983	0.983	0.983	0.991	0.991	0.991	0.991	12	+0.88	✓
Nu-SVC	0.947	0.948	0.948	0.947	0.956	0.956	0.956	0.956	12	+0.88	✓
Linear SVC	0.983	0.983	0.983	0.983	0.991	0.991	0.991	0.991	12	+0.88	✓
Gaussian NB	0.921	0.921	0.921	0.921	0.965	0.965	0.965	0.965	12	+4.39	✓
Multinomial NB	0.825	0.825	0.825	0.825	0.772	0.772	0.772	0.772	12	−5.26	×
Complement NB	0.816	0.816	0.816	0.816	0.965	0.965	0.965	0.965	12	+14.91	✓
Bernoulli NB	0.640	0.641	0.640	0.640	0.640	0.641	0.640	0.640	12	0.00	—
Decision Tree	0.929	0.929	0.930	0.930	0.974	0.974	0.974	0.974	12	+4.39	✓
Random Forest	0.974	0.974	0.974	0.974	0.983	0.983	0.983	0.983	12	+0.88	✓
Extra Tree	0.947	0.947	0.948	0.947	0.991	0.991	0.991	0.991	12	+4.39	✓
AdaBoost	0.965	0.965	0.965	0.966	0.991	0.991	0.991	0.991	12	+2.63	✓
Gradient Boosting	0.965	0.965	0.965	0.965	0.983	0.983	0.983	0.983	12	+1.75	✓
XGBoost	0.956	0.956	0.956	0.956	0.983	0.983	0.983	0.983	12	+2.63	✓
LightGBM	0.965	0.965	0.965	0.965	0.991	0.991	0.991	0.991	9	+2.63	✓
Logistic Regression CV	0.982	0.982	0.982	0.982	0.991	0.991	0.991	0.991	12	+0.88	✓
Linear Discriminant Analysis	0.965	0.965	0.965	0.965	0.991	0.991	0.991	0.991	12	+2.63	✓
Quadratic Discriminant Analysis	0.947	0.947	0.947	0.947	0.991	0.991	0.991	0.991	12	+4.39	✓
Multi-Layer Perceptron	0.982	0.982	0.982	0.982	0.991	0.991	0.991	0.991	12	+0.88	✓
Label Propagation	0.938	0.938	0.938	0.938	0.974	0.974	0.974	0.974	12	+3.51	✓
Label Spreading	0.938	0.938	0.938	0.938	0.965	0.965	0.965	0.965	12	+2.63	✓
SGD Classifier	0.974	0.974	0.974	0.974	0.991	0.991	0.991	0.991	12	+1.75	✓
Passive Aggressive Classifier	0.912	0.912	0.912	0.912	0.991	0.991	0.991	0.991	12	+7.89	✓
Ridge Classifier	0.956	0.956	0.956	0.956	0.973	0.973	0.973	0.973	12	+1.75	✓
Ridge Classifier CV	0.973	0.973	0.973	0.973	0.982	0.982	0.982	0.982	12	+0.88	✓
Hist Gradient Boosting	0.965	0.965	0.965	0.965	0.982	0.982	0.982	0.982	12	+1.75	✓
Bagging	0.965	0.965	0.965	0.965	0.974	0.974	0.974	0.974	12	+0.88	✓
Perceptron	0.921	0.921	0.921	0.921	0.991	0.991	0.991	0.991	12	+7.02	✓
Mean Accuracy	0.937	—	—	—	0.963	—	—	—	11.9	+2.63	—
Std. Deviation (Acc.)	±0.069	—	—	—	±0.073	—	—	—	±0.5	±3.27	—
Models Improved	—	—	—	—	—	—	—	—	—	29/29 (100%)	—

Table VIII: 10-Fold Cross-Validation Results: Top 5 PSO-Optimized Models

Model	CV Accuracy	CV F1-Score	CV Precision	CV Recall
MLP	0.9719	0.9717	0.9736	0.9719
L-SVC	0.9719	0.9716	0.9735	0.9719
SVC	0.9702	0.9701	0.9714	0.9702
KNN	0.9701	0.9700	0.9715	0.9701
LDA	0.9667	0.9661	0.9688	0.9667

G. Comparative Analysis of Studies for Breast Cancer Classification

Table XII presents a comparison of the state-of-the-art results for breast cancer diagnosis using the WDBC dataset. Recent studies have achieved accuracies ranging from 96% to 100%, with most reporting accuracies above 98%. Notably, [14] and [16] report accuracies of 99.1% and 98.9%, respectively, but they primarily focus on accuracy metrics, omitting essential performance metrics such as precision, recall, and F1-score.

In contrast, the approach proposed in this work achieves a competitive 99.3% accuracy while addressing these gaps by providing not only a comprehensive set of performance metrics

(precision, recall, F1-score) but also full explainability through SHAP (SHapley Additive exPlanations) integration. This level of interpretability allows clinicians to understand the reasoning behind model predictions, making the system more transparent and reliable for real-world use.

Compared to previous work, such as [19], which provides partial explainability, this approach offers a more robust and comprehensive solution, making it a stronger candidate for real-world clinical deployment. By addressing both accuracy and interpretability, this work provides a more thorough and actionable tool for clinicians, aligning with the growing demand for transparent and trustworthy AI systems in healthcare.

V. Conclusion

In this study, we emphasized the efficient utilization of machine learning algorithms for the accurate prediction of breast cancer using the WDBC dataset. By integrating Particle Swarm Optimization (PSO) with a broad spectrum of traditional classifiers, we demonstrated the significant impact of feature selection on enhancing predictive performance and interpretability. The framework systematically evaluated 29 machine learning models, achieving consistently high performance across all metrics, with the Multilayer Perceptron

Table IX: Detailed 10-Fold Cross-Validation Performance Analysis of the Top-Ranked Multi-Layer Perceptron Model with PSO-Optimized Feature Selection

Fold	Accuracy	F1-Score	Precision	Recall	Balanced Accuracy
1	0.983	0.983	0.983	0.983	0.986
2	0.965	0.965	0.968	0.965	0.971
3	1.000	1.000	1.000	1.000	1.000
4	0.930	0.928	0.937	0.930	0.905
5	0.947	0.947	0.947	0.947	0.939
6	0.965	0.965	0.967	0.965	0.952
7	0.983	0.983	0.983	0.983	0.986
8	0.965	0.965	0.968	0.965	0.972
9	0.983	0.982	0.983	0.983	0.976
10	1.000	1.000	1.000	1.000	1.000
Mean $\pm$ SD	0.972 $\pm$ 0.022	0.972 $\pm$ 0.023	0.974 $\pm$ 0.021	0.972 $\pm$ 0.022	0.969 $\pm$ 0.030

Table X: Pairwise t-tests between top 5 models (accuracy scores)

Model 1	Model 2	t-stat.	p-value
MLP	LSVC	0.01	0.996
MLP	SVC	0.20	0.847
MLP	KNN	0.32	0.754
MLP	LDA	0.76	0.468
LSVC	SVC	0.19	0.850
LSVC	KNN	0.36	0.726
LSVC	LDA	1.14	0.283
SVC	KNN	0.00	0.997
SVC	LDA	0.39	0.705
KNN	LDA	0.61	0.560

Abbreviations: MLP = Multi-Layer Perceptron, LSVC = Linear SVC, SVC = Support Vector Classifier, KNN = K-Nearest Neighbors, LDA = Linear Discriminant Analysis.

Table XI: Statistical Significance Analysis: PSO Optimization Effectiveness

Statistical Test	Value	Interpretation
Mean Baseline Acc.	0.9684 $\pm$ 0.0131	High baseline
Mean PSO Acc.	0.9912 $\pm$ 0.0000	Excellent PSO
Avg. Improvement	+2.28%	Significant
Paired t-statistic	3.4744	Strong evidence
P-value	0.0255	Sig. (p<0.05)
Cohen's d	2.1974	Large effect
Models Improved	27/29 (93.1%)	Excellent rate
Models Degraded	1/29 (3.4%)	Minimal rate
Models Unchanged	1/29 (3.4%)	Rare cases
Best Improvement	+14.91% (Comp. NB)	Outstanding gain
Worst Perform.	-5.26% (Multi. NB)	Isolated degradation

achieving an accuracy of 99.12%. Beyond predictive accuracy, this study highlighted the clinical relevance of optimized features when combined with digital technologies, underscoring the potential of machine learning in medical diagnostics. The incorporation of SHAP-based explainability and statistical validation further ensured generalizability and transparency, making the proposed framework more suitable for clinical adoption. Future work, aligning with limitations, will focus on extending the framework to multi-modal datasets, such as genomic and imaging data, and validating the approach in real-

Table XII: Comparison of State-of-the-art Results for Breast Cancer Diagnosis on WDBC Dataset

Study (Year)	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	X-AI
Modi et al. (2016)	97.0	–	–	–	No
Aalaei et al. (2016)	97.2	–	–	–	No
Sheikhpour et al. (2016)	98.5	–	97.7	–	No
Singh et al. (2017)	96.9	96.0	96.0	96.0	No
Xie et al. (2021)	98.8	–	–	–	No
Aamir et al. (2022)	99.1	–	–	–	No
Strelcencia et al. (2023)	98.6	–	–	–	No
Akkur et al. (2023)	98.9	97.17	100.0	98.8	No
Kazerani et al. (2024)	99.0	100.0	98.0	98.0	No
Alhassan et al. (2024)	98.5	–	–	–	No
Zhu et al. (2025)	99.0	100.0	97.4	98.7	Partial
Etcil et al. (2025)	98.7	–	–	–	No
This Work (2025)	99.1	99.1	99.1	99.1	Yes

world clinical environments to assess its scalability, robustness, and trustworthiness for deployment in healthcare systems.

References

- [1] World Health Organization Cancer Fact Sheet. (2023), <https://www.who.int/news-room/fact-sheets/detail/cancer>, Accessed: 2025-09-01
- [2] American Cancer Society What Is Breast Cancer?. (2023), <https://www.cancer.org/cancer/types/breast-cancer/about/what-is-breast-cancer.html>, Accessed: 2025-09-01
- [3] National Cancer Institute Breast Cancer—Patient Version. (2022), <https://www.cancer.gov/types/breast>, Accessed: 2025-09-01
- [4] Kourou, K., Exarchos, T., Exarchos, K., Karamouzis, M. & Fotiadis, D. Machine learning applications in cancer prognosis and prediction. *Computational And Structural Biotechnology Journal*. **13** pp. 8-17 (2015)
- [5] Mueller, T., Einecke, G., Reeve, J., Sis, B., Mengel, M., Jhangri, G., Bunnag, S., Cruz, J., Wishart, D., Meng, C. & Others Microarray analysis of rejection in human kidney transplants using pathogenesis-based transcript sets. *American Journal Of Transplantation*. **7**, 2712-2722 (2007)
- [6] Bellaachia, A. & Guven, E. Predicting breast cancer survivability using data mining techniques. *Age*. **58**, 10-110 (2006)
- [7] Litjens, G., Kooi, T., Bejnordi, B., Setio, A., Ciompi, F., Ghafoorian, M., Van Der Laak, J., Van Ginneken, B. & Sánchez, C. A survey on deep learning in medical image analysis. *Medical Image Analysis*. **42** pp. 60-88 (2017)

- [8] Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J. & Socher, R. Deep learning-enabled medical computer vision. *NPJ Digital Medicine*. **4**, 5 (2021)
- [9] Modi, N. & Ghanchi, K. A comparative analysis of feature selection methods and associated machine learning algorithms on Wisconsin breast cancer dataset (WBCD). *Proceedings Of International Conference On ICT For Sustainable Development: ICT4SD 2015 Volume 1*. pp. 215-224 (2016)
- [10] Aalaei, S., Shahraki, H., Rowhanimanesh, A. & Eslami, S. Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets. *Iranian Journal Of Basic Medical Sciences*. **19**, 476 (2016)
- [11] Sheikhpour, R., Sarram, M. & Sheikhpour, R. Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer. *Applied Soft Computing*. **40** pp. 113-131 (2016)
- [12] Jeyasingh, S. & Veluchamy, M. Modified bat algorithm for feature selection with the wisconsin diagnosis breast cancer (WDBC) dataset. *Asian Pacific Journal Of Cancer Prevention: APJCP*. **18**, 1257 (2017)
- [13] Xie, H., Zhang, L., Lim, C., Yu, Y. & Liu, H. Feature selection using enhanced particle swarm optimisation for classification models. *Sensors*. **21**, 1816 (2021)
- [14] Aamir, S., Rahim, A., Aamir, Z., Abbasi, S., Khan, M., Alhaisoni, M., Khan, M., Khan, K. & Ahmad, J. Predicting breast cancer leveraging supervised machine learning techniques. *Computational And Mathematical Methods In Medicine*. **2022**, 5869529 (2022)
- [15] Strelcenia, E. & Prakoonwit, S. Effective feature engineering and classification of breast cancer diagnosis: a comparative study. *BioMed-Informatics*. **3**, 616-631 (2023)
- [16] Akkur, E., TURK, F. & Erogul, O. Breast Cancer Diagnosis Using Feature Selection Approaches and Bayesian Optimization.. *Computer Systems Science & Engineering*. **45** (2023)
- [17] Kazerani, R. Improving breast cancer diagnosis accuracy by particle swarm optimization feature selection. *International Journal Of Computational Intelligence Systems*. **17**, 44 (2024)
- [18] Alhassan, A. An improved breast cancer classification with hybrid chaotic sand cat and Remora Optimization feature selection algorithm. *Plos One*. **19**, e0300622 (2024)
- [19] Zhu, J., Zhao, Z., Yin, B., Wu, C., Yin, C., Chen, R. & Ding, Y. An integrated approach of feature selection and machine learning for early detection of breast cancer. *Scientific Reports*. **15**, 13015 (2025)
- [20] Etcil, M., Dedeturk, B., Kolukisa, B., Bakir-Gungor, B. & Gungor, V. Breast Cancer Detection Using a New Parallel Hybrid Logistic Regression Model Trained by Particle Swarm Optimization and Clonal Selection Algorithms. *Concurrency And Computation: Practice And Experience*. **37**, e70107 (2025)
- [21] Dixon, W. Simplified estimation from censored normal samples. *The Annals Of Mathematical Statistics*. pp. 385-391 (1960)
- [22] Tukey, J. & Others Exploratory data analysis. (Springer,1977)
- [23] Han, J., Kamber, M. & Pei, J. Data mining: Concepts and. *Techniques*, Waltham: Morgan Kaufmann Publishers. (2012)
- [24] Kennedy, J. & Eberhart, R. Particle swarm optimization. *Proceedings Of ICNN'95-international Conference On Neural Networks*. **4** pp. 1942-1948 (1995)
- [25] Shi, Y. & Eberhart, R. A modified particle swarm optimizer. 1998 *IEEE International Conference On Evolutionary Computation Proceedings. IEEE World Congress On Computational Intelligence (Cat. No. 98TH8360)*. pp. 69-73 (1998)
- [26] Xue, B., Zhang, M. & Browne, W. Particle swarm optimization for feature selection in classification: A multi-objective approach. *IEEE Transactions On Cybernetics*. **43**, 1656-1671 (2012)
- [27] Vapnik, V. The nature of statistical learning theory. (Springer science & business media,2013)
- [28] Cover, T. Elements of information theory. (John Wiley & Sons,1999)
- [29] Liu, H. & Motoda, H. Feature selection for knowledge discovery and data mining. (Springer science & business media,2012)
- [30] Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *Journal Of Machine Learning Research*. **3**, 1157-1182 (2003)
- [31] Dash, M. & Liu, H. Feature selection for classification. *Intelligent Data Analysis*. **1**, 131-156 (1997)
- [32] Hosmer Jr, D., Lemeshow, S. & Sturdivant, R. Applied logistic regression. (John Wiley & Sons,2013)
- [33] Robbins, H. & Monro, S. A stochastic approximation method. *The Annals Of Mathematical Statistics*. pp. 400-407 (1951)
- [34] Bottou, L. Large-scale machine learning with stochastic gradient descent. *Proceedings Of COMPSTAT'2010: 19th International Conference On Computational Statistics Paris France, August 22-27, 2010 Keynote, Invited And Contributed Papers*. pp. 177-186 (2010)
- [35] Hoerl, A. & Kennard, R. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*. **12**, 55-67 (1970)
- [36] Ruppert, D. The elements of statistical learning: data mining, inference, and prediction. (Taylor & Francis,2004)
- [37] Cox, D. The regression analysis of binary sequences. *Journal Of The Royal Statistical Society Series B: Statistical Methodology*. **20**, 215-232 (1958)
- [38] Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain.. *Psychological Review*. **65**, 386 (1958)
- [39] Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S. & Singer, Y. Online passive-aggressive algorithms. *Journal Of Machine Learning Research*. **7**, 551-585 (2006)
- [40] Cortes, C. & Vapnik, V. Support-vector networks. *Machine Learning*. **20**, 273-297 (1995)
- [41] Schölkopf, B., Smola, A., Williamson, R. & Bartlett, P. New support vector algorithms. *Neural Computation*. **12**, 1207-1245 (2000)
- [42] Fan, R., Chang, K., Hsieh, C., Wang, X. & Lin, C. LIBLINEAR: A library for large linear classification. *The Journal Of Machine Learning Research*. **9** pp. 1871-1874 (2008)
- [43] John, G. & Langley, P. Estimating continuous distributions in Bayesian classifiers. *ArXiv Preprint ArXiv:1302.4964*. (2013)
- [44] Murphy, K. & Others Naive bayes classifiers. *University Of British Columbia*. **18**, 1-8 (2006)
- [45] McCallum, A., Nigam, K. & Others A comparison of event models for naive bayes text classification. *AAAI-98 Workshop On Learning For Text Categorization*. **752**, 41-48 (1998)
- [46] Rennie, J., Shih, L., Teevan, J. & Karger, D. Tackling the poor assumptions of naive bayes text classifiers. *Proceedings Of The 20th International Conference On Machine Learning (ICML-03)*. pp. 616-623 (2003)
- [47] Manning, C. Introduction to information retrieval. (Syngress Publishing,,2008)
- [48] Quinlan, J. Induction of decision trees. *Machine Learning*. **1**, 81-106 (1986)
- [49] Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Machine Learning*. **63**, 3-42 (2006)
- [50] Breiman, L. Random forests. *Machine Learning*. **45**, 5-32 (2001)
- [51] Freund, Y. & Schapire, R. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal Of Computer And System Sciences*. **55**, 119-139 (1997)
- [52] Friedman, J. Greedy function approximation: a gradient boosting machine. *Annals Of Statistics*. pp. 1189-1232 (2001)
- [53] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. & Liu, T. Lightgbm: A highly efficient gradient boosting decision tree. *Advances In Neural Information Processing Systems*. **30** (2017)
- [54] Breiman, L. Bagging predictors. *Machine Learning*. **24**, 123-140 (1996)
- [55] Fisher, R. The use of multiple measurements in taxonomic problems. *Annals Of Eugenics*. **7**, 179-188 (1936)
- [56] Rumelhart, D., Hinton, G. & Williams, R. Learning representations by back-propagating errors. *Nature*. **323**, 533-536 (1986)
- [57] Cover, T. & Hart, P. Nearest neighbor pattern classification. *IEEE Transactions On Information Theory*. **13**, 21-27 (1967)
- [58] Zhu, X., Ghahramani, Z. & Lafferty, J. Semi-supervised learning using gaussian fields and harmonic functions. *Proceedings Of The 20th International Conference On Machine Learning (ICML-03)*. pp. 912-919 (2003)
- [59] Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. *Proceedings Of The 22nd Acm Sigkdd International Conference On Knowledge Discovery And Data Mining*. pp. 785-794 (2016)
- [60] Stone, M. Cross-validation: A review. *Statistics: A Journal Of Theoretical And Applied Statistics*. **9**, 127-139 (1978)
- [61] Geisser, S. The predictive sample reuse method with applications. *Journal Of The American Statistical Association*. **70**, 320-328 (1975)
- [62] Kohavi, R. & Others A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*. **14**, 1137-1145 (1995)
- [63] Powers, D. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *ArXiv Preprint ArXiv:2010.16061*. (2020)
- [64] Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters*. **27**, 861-874 (2006)
- [65] Sokolova, M. & Lapalme, G. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*. **45**, 427-437 (2009)