Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 31 No. 2 (2024)

Integration of Complex Mathematical Approaches and Statistical Methodologies in Machine Learning for Design of Robust Prediction Models

Sarita Silaich[1], Dr Rajesh yadav[2]

[1] PhD Scholar, Department of Computer Science and Engineering, Mody University of Science and Technology, Rajasthan, India

[2] Assistant Professor, Department of Computer Science and Engineering, Mody University of Science and Technology, Rajasthan, India

Email: sarita.bits@gmail.com [1], yadav.rajesh27@gmail.com [2]

Article History:

Received: *12-02-2024*

Revised: *26-04-2024*

Accepted: *06-05-2024*

Abstract:

The disease known as breast cancer continues to be one of the most common and potentially fatal diseases that affect women all over the world. In order to provide effective therapy and more favorable outcomes for patients, early detection and correct diagnosis are absolutely necessary. This research explores the integration of operational research (OR) methodologies and advanced statistical methods within machine learning frameworks to enhance the prediction accuracy of breast cancer outcomes. By harnessing the power of nonlinear optimization and statistical techniques, including regression analysis and probability distribution models, we aim to refine the predictive capabilities of existing algorithms. The research employs a comprehensive dataset derived from clinical trials and patient records, analyzed through a series of machine learning models that incorporate elements of combinatorial optimization, decision analysis, and stochastic modeling. Key performance metrics, such as accuracy, sensitivity, and specificity, are evaluated against standard benchmarks to determine the efficacy of the integrated approaches. Preliminary results indicate that incorporating OR and statistical methods significantly improves model robustness and predictive accuracy. The study not only demonstrates the potential of applied mathematics in medical diagnostics but also provides a framework for future research in enhancing machine learning models for health outcomes prediction through mathematical innovations. This investigation contributes to the field of mathematical oncology by demonstrating how applied nonlinear analysis can bridge the gap between theoretical mathematical approaches and practical clinical applications, offering new pathways for early and more accurate detection of breast cancer.

Keywords: Operational Research, Advanced Statistical Methods, Machine Learning, Breast Cancer Prediction, Nonlinear Optimization, Regression Analysis, Probability Models, Combinatorial Optimization, Mathematical Oncology, Clinical Applications.

1. INTRODUCTION

Breast cancer is a major public health issue globally, with early detection significantly increasing the chances of successful treatment and survival. Traditional diagnostic methods include mammography, ultrasound, and biopsies, which are often complemented by predictive modeling to identify high-risk cases early. Machine learning (ML) models have increasingly been applied to improve the accuracy and efficiency of these predictions. Operational research

(OR) and advanced statistical methods provide robust frameworks to enhance these machine learning models. OR, primarily concerned with optimizing complex operations and decision-making processes, applies various mathematical techniques to maximize efficiency and outcomes. In the context of breast cancer prediction, OR can optimize how predictive models handle data, make classifications, and even determine the best sequences of diagnostic tests. Statistical methods, especially those involving advanced calculations like logistic regression, Bayesian inference, and survival analysis, are pivotal in interpreting medical data. These methods help in understanding the relationships between various risk factors and the likelihood of developing breast cancer. The integration of these statistical methods into machine learning models ensures that the predictions are not only based on patterns in the data but are also statistically sound, reflecting true correlations and causations. The synergy between machine learning, operational research, and advanced statistics is potentiated through several key areas:

1. **Data Optimization:** OR techniques can optimize data preprocessing, selection, and reduction to enhance the quality and speed of machine learning algorithms.
2. **Model Selection and Tuning:** Advanced statistical methods aid in selecting the right model and tuning parameters to improve prediction accuracy and reduce overfitting.
3. **Algorithm Enhancement:** By incorporating OR algorithms such as linear programming, decision trees, and network flows, machine learning models can be refined to handle specific complexities of breast cancer data more effectively.

In developing robust predictive models, it is essential to apply a structured approach to integrate these disciplines. The following equations provide a mathematical basis for this integration, illustrating how various operational research and statistical techniques can be applied to refine machine learning algorithms specifically tailored for breast cancer prediction.

The general logistic regression model for binary outcomes, where *Y* is the binary response (breast cancer occurrence or not) and **X** represents the input features (e.g., age, genetics, lifestyle factors):

$$Y = \frac{1}{1+e^{-|x_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n|}} \quad (1)$$

The likelihood function for the logistic regression, used to estimate the parameters β :

$$L(\beta) = \prod_{i-1}^{n} p_i^{y_k}(1-p_i)^{1-\beta} \quad (2)$$

The log of the likelihood function, which is often used because it is simpler to maximize:

$$\log L(\beta) = \sum_{i-1}^{n} [y_i \log p_i + (1-y_i)\log(1-p_i)] \quad (3)$$

The first derivative of the log-likelihood function, used to find the maximum likelihood estimates:

$$S(\beta) = \frac{\partial \log L(\beta)}{\partial \beta} - \sum_{i-1}^{n} x_i(y_i - p_i) \quad (4)$$

The second derivative of the log-likelihood function, which assesses the curvature of the log-likelihood surface:

$$H(\beta) = -\sum_{i-1}^{n} x_i x_i^T p_i(1-p_i) \quad (5)$$

The update rule in the Newton-Raphson method for finding the maximum likelihood estimates:

$$\beta^{(\text{new})} - \beta^{(\text{old})} - H^{-1}\left(\beta^{(\text{old})}\right) S\left(\beta^{(\text{old})}\right) \quad (6)$$

The variance-covariance matrix of the estimator β, assuming the model is correctly specified:

$$\text{Var}\,(\hat{\beta}) = -[H(\hat{\beta})]^{-1} \quad (7)$$

Wald Test Statistic is used for hypothesis testing of coefficients:

$$W = \left(\hat{\beta}_j - 0\right)^2 \left[\text{Var}\left(\hat{\beta}_j\right)\right]^{-1} \quad (8)$$

Receiver Operating Characteristic (ROC) Curve is a tool used to evaluate the performance of a binary classifier system:

$$\text{TPR} = \frac{\text{TP}}{\text{TP + IN}} \quad (9)$$

$$\text{FPR} = \frac{\text{FP}}{\text{TN+FT}} \quad (10)$$

Area Under the ROC Curve (AUC) is a scalar measure to assess the overall performance of the diagnostic tests:

$$\text{AUC} = \int_0^1 \text{TPR}(t)dt \quad (11)$$

K-fold cross-validation procedure to evaluate model stability:

$$CV_{\text{k- fold}} = \frac{1}{k}\sum_{i-1}^{k} \text{Accuracy}_{\ i} \quad (12)$$

Akaike Information Criterion (AIC) used for model selection among a set of models:

$$\text{AIC} = 2k - 2\log(L) \quad (13)$$

Bayesian Information Criterion (BIC) is an another criterion for model selection:

$$\text{BIC} = \log(n)k - 2\log(L) \quad (14)$$

Sensitivity Analysis formula is used to calculate sensitivity, or true positive rate:

$$\text{Sensitivity} = \frac{TP}{\text{TP + FN}} \quad (15)$$

Specificity Analysis is used to calculate specificity, or true negative rate:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FN}} \quad (16)$$

Positive Predictive Value is the probability that subjects with a positive screening test truly have the disease:

$$\text{PPV} = \frac{\text{TP}}{\text{TP+FP}} \quad (17)$$

Negative Predictive Value is the probability that subjects with a negative screening test truly don't have the disease:

$$\text{NPV} = \frac{\text{TN}}{\text{TN+FN}} \quad (18)$$

F1Score is equal to the harmonic mean of precision and recall:

$$F1 = 2 \cdot \frac{\text{PPV.Sensitivity}}{\text{PPV} + \text{Sensitivity}} \quad (19)$$

Decision Tree Split Criterion is Used in decision tree algorithms to choose the best split:

$$G - 1 - \sum_{i-1}^{c} p_i^2 \quad (20)$$

Information Gain is used in decision tree algorithms, measuring the effectiveness of an attribute in classifying the training data:

$$IG(T, A) = H(T) - H(T \mid A) \quad (21)$$

Entropy which is a measure of the amount of uncertainty in the dataset T :

$$H(T) = -\sum_{i-1}^{c} p_i \log_2 (p_i) \quad (22)$$

Conditional Entropy is equal to Entropy of the dataset after using attribute A for splitting:

$$H(T \mid A) = \sum_{j-1}^{v} \frac{|T_j|}{|T|} H(T_j) \quad (23)$$

Support Vector Machine Margin Maximization of the objective function for SVM focusing on margin maximization:

$$\min_{\mathbf{w},b} \frac{1}{2} \| \mathbf{w} \|^2 \quad (24)$$

Kernel Trick for Non-Linear Separation is a transformation used in SVM for non-linear classification:

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j) \quad (25)$$

After establishing the mathematical framework, it's crucial to understand the implications of these equations and how they can be practically applied in the field of breast cancer prediction. Each equation plays a critical role in enhancing the robustness and accuracy of predictive models.

The logistic regression model, for instance, is foundational in medical statistics, allowing for the estimation of probabilities directly linked to patient outcomes based on multiple risk factors. The optimization of this model through operational research techniques, such as the Newton-Raphson method (Equation 1.6), significantly refines parameter estimation, making the model more responsive to subtle variations in patient data.

The Hessian matrix and variance-covariance calculations are critical for understanding the confidence intervals around estimated parameters, providing insights into the reliability of predictions and the stability of the model under various conditions.

Performance metrics derived from statistical analysis, such as the AUC of the ROC curve, are indispensable for evaluating the effectiveness of breast cancer prediction models. They help in assessing how well the model can distinguish between patients with and without breast cancer, which is crucial for clinical decision-making.

Moreover, cross-validation techniques ensure that the model is not just fitting the data well but also generalizes effectively to new, unseen data, thus preventing overfitting. This is particularly important in medical applications where the cost of a wrong prediction can be very high.

Finally, advanced machine learning techniques such as support vector machines and decision trees utilize operational research and statistical methods to find non-linear patterns and complex relationships in data that might not be apparent through traditional statistical methods alone.

In conclusion, the integration of operational research and advanced statistical methods into machine learning creates a powerful tool for predicting breast cancer. This integrated approach not only improves the accuracy and efficiency of predictions but also offers a deeper understanding of the underlying patterns and relationships in medical data. As research progresses, these methods will continue to evolve, providing ever more sophisticated tools that can be used to fight breast cancer more effectively. This ongoing development underscores the importance of interdisciplinary approaches in medical research, leveraging the strengths of each field to tackle complex health challenges

2. RELATED STUDY

In fact, breast cancer has been recognized as the fifth largest cause of cancer-related death worldwide in 2020 [1]. Breast cancer is the most often diagnosed form of cancer overall and among women all over the world. On a global scale, it is believed to be the most frequent kind of cancer [2]. There are a number of different tests that are utilized in the process of screening for breast cancer and diagnosing the disease. These tests include mammography, breast inspection, and a biopsy. The identification of breast cancer has been accomplished by the utilization of a variety of imaging modalities, including mammography, ultrasound (US), magnetic resonance imaging (MRI), histology pictures, and infrared thermography. For the screening of breast cancer, mammography is the method that is most widely employed. As an illustration, it is advised that women who are forty years old or older go through a mammographic screening [3, 4]. The digital mammogram and the digital breast tom synthesis (DBT) are the two instruments that make up the majority of mammography. However, it has been shown that digital mammography is less successful in individuals who have thick breasts and are less sensitive to tiny tumors (tumors with a volume of less than 1 mm [5]). This is despite the fact that the digital mammogram is the most often used detection method for breast cancer. On the other hand, these drawbacks are circumvented by DBTthe three-dimensional mammogram, which is a more advanced method of mammography, is another name for this examination.. In general, it offers a greater level of diagnostic accuracy compared to the two-dimensional mammogram [6]. On the other hand, when these two methods were utilized for screening purposes, there was not a discernible difference between them [7]. There are hopes that machine learning will lead to better health care, especially in specialized medical fields like pathology, ophthalmology, diagnostic imaging, and cardiology [8]. The faster use of machine learning in many medical areas will be caused by a number of things, such as the easy access to large amounts of medical data and the progress of computer technology. However, even with these good gains, it is still not clear how machine learning can be used in a

therapeutic setting [9–11]. People are worried about their privacy, don't trust the technology, and think that machine learning might be biased without meaning to be [8, 12–14]. These are some of the problems that haven't been fully looked into yet. Researchers have looked into how machine learning can be used in the field of breast cancer for a number of reasons, such as to predict and screen for the disease [15], to predict when cancer will come back [16], to predict how long a patient will live [17], to predict breast density [18], and to help with treatments and management of the disease [19]. Researchers have looked into a number of different data sources and machine learning methods to see how they might be useful in different breast cancer clinical situations. These data sources include sociodemographic and clinical data, genetic data, imaging data, and more. It can basically divide the use of machine learning in this field of study into three main groups: as a screening tool, a diagnostic tool, or a prediction tool. It's just that most studies don't make it clear what role their machine learning model plays in the clinical setting or how it can be used in real life. The reason for this is that these different tasks of machine learning will have an impact on how the model is built and used.Meena et al. (2023) [20] planned a breast cancer uncovering model using the curvelet transform for feature extraction, adaptive particle swarm optimization for feature selection, and support vector machines for classification, achieving higher accuracy rates compared to previous approaches. Nurhayati et al. (2020) [21] utilized PSO for feature selection in various classification algorithms to improve breast cancer diagnosis, highlighting its effectiveness but noting that it couldn't surpass the performance of genetic algorithms. Sannasi Chakravarthy et al. (2022) [22] designed a computer-aided diagnosis (CAD) system for breast cancer diagnosis using the Ebola Optimization Algorithm (EOA) for feature selection and achieved a maximum accuracy of 97.19% with mK. - SVM Harish et al. (2022) [23] They used medical image processing methods like convolutional neural networks (CNN), particle swarm optimization (PSO), and support vector machines (SVM) to find breast cancer. In 2023, Momtahen et al. [24] suggested a DOB-Scan probe that uses ensemble learning to find breast cancer earlier and achieve high success rates with different regression algorithms. In 2020, Baskaran et al. [25] looked at GA and PSO for planning thermal treatment for breast cancer and found that GA did better for global optimization than PSO. Mani et al. (2020) [26] used a decision tree classifier on gene expression data for breast cancer diagnosis, enhancing results with elephant herding optimization for feature transformation. Vilohit et al. (2022) [27] implemented a decision tree classifier on gene expression data, enhancing performance with elephant herding optimization for feature transformation and principal component analysis for dimensionality reduction. Mitra et al. (2023) [28] applied particle swarm optimization to identify disease-causing genes and developed hybrid algorithms for the classification of triple negative breast cancer, achieving high accuracy rates. Many years ago, Aouragh et al. [29] looked at different machine learning methods for classifying breast cancer and improved them by balancing the data, choosing the right features, and optimizing hyperparameters. The results were impressive, with over 98% accuracy across all measures.

3. MATERIALS AND METHODS

The following subsections briefly summarize this paper's research materials and methods.

Dataset and Tools

Breast Cancer Wisconsin (Diagnostic) Data Set from UCI Machine Learning Repository is being tested [28]. Wisconsin Breast Cancer diagnosis from UCI repository having 569 samples 357 benign and 212 malignant..

Methodology for the Proposed System

The suggested approach distinguishes malignant from benign cells. We improved breast cancer diagnosis machine learning classification models in our research. To compare classifier accuracy, all characteristics and chosen features were examined independently. We employed wrapper-based feature selection, nature-inspired algorithms like (PSO), and a hybrid of PSO and grey wolf optimizer to discover key features. Popular machine learning classifiers SVM, KNN, LR, and RF were employed on these features. The suggested system has five stages: (1) Pre-processing of the Data, (2) Data imbalance management, (3) Feature Selection, (4) Classes derived by machine learning, as well as (5) The evaluation of the performance of the classifier.

Data Pre-Processing and cross validation

Preprocessing data is a necessary step in order to represent data efficiently. This phase encompasses the elimination of absent values and the discretization of features, which involves converting numeric data to nominal. Discreteization facilitates the generation of comprehensible branches for the decision tree, as opposed to branches that rely on numerical values. The feature row containing missing values is eliminated from the dataset. Cross-validation is a method employed to evaluate the efficiency of a machine learning model and to alleviate concerns, including overfitting. The procedure necessitates dividing the dataset into many folds, which are subsets; the model is subsequently trained on a subset of the folds and assessed on the remaining folds. Each time this procedure is replicated, distinct subsets are utilized for instruction and evaluation.

Feature Selections

the proposed study on breast cancer prediction using supervised learning methods, feature selection plays a crucial role in identifying the most relevant and informative features from the dataset. Effective feature selection can improve model performance, reduce overfitting, and enhance interpretability. In the context of feature selection for breast cancer prediction, both Particle Swarm Optimization (PSO) and Grey Wolf Optimizer (GWO) can be employed as metaheuristic algorithms to efficiently search for the optimal subset of features. These algorithms aim to select the most relevant features while minimizing redundancy, thus improving the performance of the predictive models. Here's how PSO and GWO can be applied for feature selection:

A-Particle Swarm Optimization

PSO [29] is metaheuristic algorithms that illustrate inspiration from swarm performance observed in nature, specially the flocking of birds. In 1995, Kennedy and Eberhart put forth the proposition. PSO is a stochastic optimization technique that manipulates populations and draws inspiration from the social dynamics observed in fish schooling or avian flocking. The

proposition was initially put forth in 1995 by Kennedy and Eberhart. PSO attempts to improve the fitness of a candidate solution iteratively through simulation of the social behavior of particles traversing a search space. Here is the typical operation of PSO:

Initialization: Particle Swarm Optimization (PSO) starts by randomly selecting a population of particles to initially populate the search space. Within the context of the optimization problem, each particle is a potential solution.

Velocity and Position Update: Each time through the loop (or generation), each particle's speed and location are changed based on its current speed and location as well as the best locations found by it and its neighbors.

Velocity Update: The following formula is used to update each particle's velocity:

$$v_{ij}^{t+1} = wv_{ij}^{t} + c_1 r_1\left[pBest_{i\,j}^{t} - x_{ij}^{t}\right] + c_2 r_2\left[gBest_{ij}^{t} - x_{ij}^{t}\right] \quad (26)$$

$$x_{ij}^{t+1} = x_{ij}^{t} + v_{ij}^{t+1} \quad (27)$$

v_{ij}^{t} is velocity of i[th] partical of j[th] dimension at time t and x_{ij}^{t} is position of same, *w* is Inertia weight, c_1 c_2 are cognitive learning factor, r_1 r_2 Uniformly distributed random number between 0 and 1, pBest is its personal best value and gBest is global best value. Then position is converted in binary using sigmoid function.

$$x_{ij}^{t+1} = \begin{cases} 0 \; if \; rand() \geq \text{Sigmoid}(v_{ij}^{t+1}) \\ 1 \; if \; rand() < \text{Sigmoid}(v_{ij}^{t+1}) \end{cases} \quad (28)$$

$$\text{Sigmoid}\,(v_{ij}^{t+1}) = 1 / (1+e^{-v_{ij}^{t}}) \quad (29)$$

Where rand **()** is a random number between 0 and 1, and the sigmoid function transforms the velocity value into a probability between 0 and 1

Evaluation: After updating the positions, the fitness of each particle is evaluated based on the objective function of the optimization problem.

Update Personal and Global Best: It is possible for each particle to change its personal best position (pbest) if the new position makes it more fit. Also, if a particle finds a better answer than the current global best, it changes the global best position (gbest).

Termination: PSO iterates indefinitely until a termination condition is satisfied, which may be the attainment of a satisfactory solution or the completion of a limit number of iterations.

B-Grey Wolf Optimizer (GWO)

Grey Wolf Optimizer (GWO) A population-based metaheuristic optimization method was created by looking at the social structure and hunting habits of grey wolves. As an alternate optimization technique for the purpose of resolving complex optimization issues, GWO was presented by Mirjalili et al. in the year 2014.'[26]

Initialization: At the start, GWO creates a population of possible answers at random, which is shown as a pack of grey wolves.

Hierarchy Formation: Based on their fitness values, the grey wolves are classified as alpha, beta, delta, or omega wolves in each iteration. The alpha wolf symbolizes the most optimal solution thus far, while the beta wolf and delta wolf represent the second-best and third-best solutions, respectively. The omega wolf denotes the worst solution thus far.

Update Positions: To simulate the way that wolves hunt, each individual wolf will modify its position in accordance with the positions of the alpha, beta, and delta wolves. This is done using specific equations that determine the movement of each wolf towards the alpha, beta, and delta wolves, as well as exploration and exploitation phases.

Fitness Evaluation: The fitness of each wolf is evaluated based on the objective function of the optimization problem after their positions have been updated by the algorithm.

Update Alpha, Beta, and Delta Wolves: The fitness of the wolves in the current cycle is used to change the alpha, beta, and delta wolves. If another wolf comes up with a better idea than the current alpha, beta, or delta wolf, it takes their place.

$$D = |CX_p - AX(t)| \quad (30)$$

$$X(t+1) = X_p(t) - AD \quad (31)$$

X_p is position of prey at current iteration t, X is the position vector of a wolf, A and C are coefficient vectors given as:

$A=2ar_1-a$

$C=2r_2$

r_1 and r_2 are random vectors $\in [0, 1]$ and a linearly varies from 2 to 0

$D_\alpha = |C_1.X_\alpha - X|,$

$D_\beta = |C_2.X_\beta - X|,$

$D_\delta = |C_3.X_\delta - X| \quad (32)$

$X_1 = X_\alpha - A_1 D_\alpha ,$

$X_2 = X_\beta - A_2 D_\beta,$

$X_3 = X_\delta - A_3 D_\delta \quad (33)$

$X(t+1) = (X_1 + X_2 + X_3) \quad (34)$

As it go through the iterations, a changes from 2 to 0, which is the linear value of the α, β, and ε wolfs.

Termination: The algorithm will continue to iterate until a termination condition is satisfied, which could be reaching a maximum number of iterations or achieving a solution that is satisfactory.

C-Hybrid algorithm

A hybrid algorithm that combines PSO and GWO for breast cancer detection involves integrating the search mechanisms of both algorithms. Here's a conceptual outline along with mathematical expressions:

Initialization: Initialize the population of particles for PSO and the pack of wolves for GWO randomly within the search space.

Objective Function: the objective function that represents the fitness of a solution based on its ability to classify breast cancer accurately. Precision, specificity, sensitivity, and the ROC area are just a few of the possible factors that might be incorporated into this function.

GWO position update equations They do not make use of traditional mathematical formulae; rather, we make use of the inertia constant to guide the exploration and exploitation of the grey wolf within the bounds of the search space. Equation (5), after being modified, is now the following:

$$D_\alpha = |C_1.X_\alpha - w*X|,\ D_\beta = |C_2.X_\beta - w*X|,\ D_\delta = |C_3.X_\delta - w*X| \quad (35)$$

Particle and Wolf Movement:

Update the velocity of each particle in PSO and the position of each wolf in GWO based on their current positions and velocities, similar to standard PSO and GWO algorithms.

PSO velocity updates equation:

$$v_{ij}^{t+1} = w(v_{ij}^{t} + c_1\, r_1(X_1 - X_i^t) + c_2\, r_2(X_2 - X_i^t) + c_3\, r_3(X_3 - X_i^t)) \quad (36)$$

Then position is calculated using updated velocity as in PSO (37)

$$x_i^{t+1} = x_i^t + v_i^{t+1} \quad (38)$$

Fitness Evaluation: The goal function should be used to determine the fitness of each wolf and particle.

Update Best Positions: Update the personal best positions (pbest) of particles in PSO and the alpha, beta, and delta positions of wolves in GWO based on the fitness evaluations.

Hybridization: Combine the movement strategies of PSO and GWO, possibly by assigning different weights or probabilities to each algorithm's update equations. For example; you could use a weighted average of the velocity update from PSO and the position update from GWO to update the position of each particle or wolf.

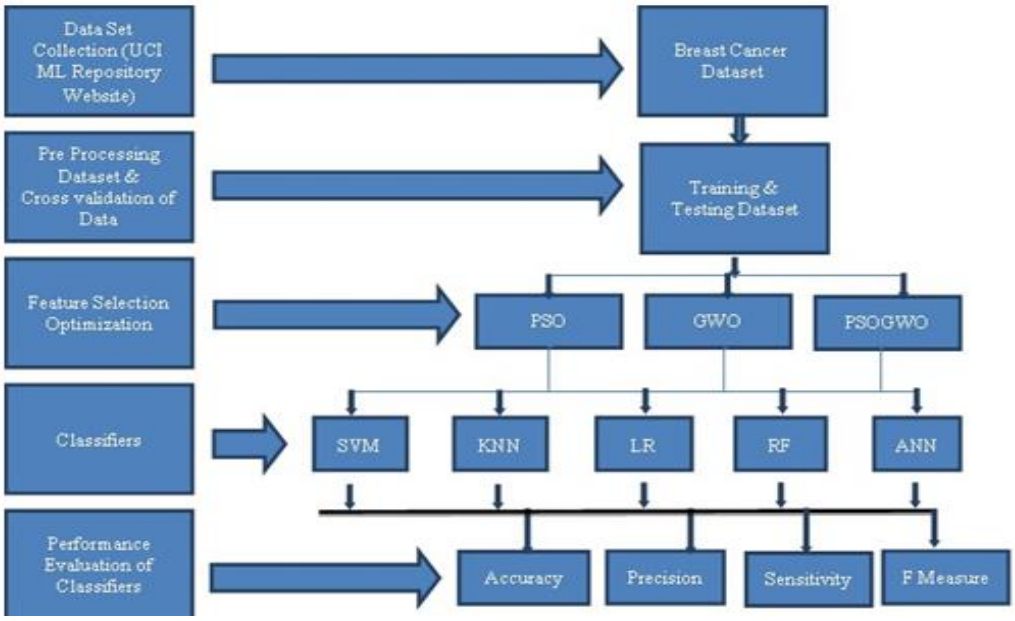

Fig.1 Proposed Flow Diagram

Termination: Iterate through steps 3–6 until a termination condition is met, like as when the maximum number of iterations is reached or when a satisfactory solution is found.

Classification: Use the final positions of particles or wolves as the selected features for breast cancer detection. Apply a classification algorithm like SVM, KNN, LR, and Random Forest RF to classify the breast cancer instances based on the selected features.

4. EXPERIMENT SETUP AND PROPOSED SYSTEM

During the course of this research endeavor, the Breast Cancer dataset, which was obtained from the UCI Machine Learning Repository, was utilized.The dataset was then subjected to pre-processing procedures in order to get it ready for analysis, which included dealing with any missing variables. Following preprocessing, we separated the dataset into training and testing sets, which were used for the construction of the model and the evaluation of the model, respectively. For optimization algorithms such as PSO, GWO, and a hybrid approach combining PSO with GWO. These algorithms were used to identify the most relevant features from the dataset, thereby improving the efficiency and effectiveness of our models. The results of our study demonstrated the effectiveness of different feature selection and classifier combinations in predicting breast cancer outcomes. By comparing the performance metrics of each model, we were able to identify the most accurate and reliable classifiers for breast cancer prediction.

5. RESULT & DISCUSSION

This section discusses ML classification models and outcomes from diverse methodologies. Initially, we used ML classifiers to eliminate missing values and undesirable data from pre-processed data. During the second stage, we used PSO, GWO, and a hybrid strategy, as well as Wrapper approaches including SVM, KNN, LR, and RF, ANN, on both preprocessed and unprocessed datasets. The total number of characteristics that were chosen using these methods. The comparison of the accuracy of these classifiers with the accuracy (percentage) results of virtual machine models that were trained with a variety of data splitting ratios (90-10, 80-20, 70-30, and 60-40) and optimized with a number of different optimization algorithms

Performance evaluation metrics

Performance evaluation metrics are crucial for assessing the effectiveness of machine learning classifiers. The commonly used metrics include classification accuracy, precision, recall

Table 1 performance of different feature selection and classification techniques

	Train Test Ratio	All Features			PSO			GWO			HPSOGWO		
		Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
SVM	90-10	96.49	97.22	97.22	92.98	97.06	91.67	96.49	97.22	97.22	92.08	97.06	91.67
	80-20	98.25	98.60	98.60	95.61	97.18	95.83	97.37	97.26	98.60	95.61	97.18	95.83
	70-30	97.66	98.13	98.30	97.49	98.13	97.20	97.66	97.20	99.07	97.08	97.22	98.13
	60-40	96.93	96.58	98.60	97.81	97.92	98.60	97.81	97.26	99.30	98.25	97.28	100.00
KNN	90-10	98.25	97.30	99.00	96.49	97.22	97.22	98.25	97.30	100.00	94.74	97.14	94.44
	80-20	98.25	97.30	98.00	93.86	97.10	93.06	93.86	95.77	94.44	93.86	97.10	93.06
	70-30	97.60	96.40	98.00	94.74	95.37	96.26	95.32	94.59	98.13	95.32	95.41	97.20
	60-40	96.05	95.89	97.90	95.18	95.21	97.20	95.18	94.59	97.90	96.49	96.55	97.90
LR	90-10	96.49	97.22	97.22	98.25	97.30	99.00	92.98	97.06	91.67	96.49	97.22	97.22
	80-20	95.01	94.67	98.61	96.49	95.95	98.61	94.74	95.83	95.83	93.86	93.33	97.22
	70-30	94.74	92.98	99.07	94.74	92.24	99.00	94.50	92.92	98.13	93.57	92.11	98.13
	60-40	95.80	94.00	98.60	95.61	94.04	99.30	95.61	94.63	98.63	94.74	93.38	98.60
RF	90-10	96.49	97.22	97.22	94.74	97.14	94.44	92.98	97.06	91.67	94.74	97.14	94.44
	80-20	94.74	95.83	95.83	95.61	97.18	95.83	96.49	95.95	98.60	94.74	95.83	95.83
	70-30	95.32	95.41	97.20	95.32	95.14	97.20	94.15	95.33	95.33	94.15	94.50	96.26
	60-40	94.74	95.17	96.50	96.49	96.55	97.90	95.18	95.83	96.50	93.86	94.48	95.80
ANN	90-10	96.49	97.22	97.22	92.98	97.06	91.67	91.23	96.97	98.89	94.74	97.14	94.44
	80-20	97.37	98.59	97.22	95.61	97.18	95.83	94.74	95.83	95.83	96.49	98.57	95.83
	70-30	98.25	99.06	98.30	95.91	96.30	97.20	95.32	95.41	97.20	97.08	98.11	97.20
	60-40	98.68	99.30	98.60	96.61	96.50	96.50	95.60	95.85	97.20	96.93	97.89	97.20

Classification Accuracy: The accuracy of classification is determined by determining the percentage of data points that have been successfully classified out of the total number of data points. Calculated by taking the total number of data points and dividing it by the sum of true positives (TP) and true negatives (TN), it includes the following:

$$\textbf{Accuracy}=\frac{\textbf{TP+TN+FP+FN}}{\textbf{TP+TN}}$$

Precision: Accuracy can be defined as the ratio of the actual positive to the total number of positives anticipated. A model with a high precision has a low false positive rate, which means that it only sometimes incorrectly identifies negative occurrences as positive. Conversely, a low precision suggests that the model tends to make a significant number of false positive

predictions, which can be problematic in scenarios where false positives are costly or undesirable as in fraud detection, and spam filtering.

$$\textbf{Precision} = \frac{\mathbf{TP}}{\mathbf{TP+FP}}$$

Recall,-Recall, In classification tasks, a performance indicator that is also known as sensitivity or true positive rate is utilized to evaluate the capability of a model to accurately identify all positive cases from the entire number of actual positive examples that are contained inside the dataset. A formula that is used to compute it is as follows:

$$\textbf{Recall-=}\frac{\mathbf{TP}}{\mathbf{TP+FN}}$$

Table 1 shows a full comparison of how well different machine learning classifiers (SVM, KNN, LR, RF, and ANN) worked when trained with different optimization algorithms (PSO, GWO, and HPSOGWO) and when the train-test ratios were 90-10, 80-20, 70-30, and 60-40. For each combination of classifier, optimization algorithm, and train-test ratio, the table reports accuracy, precision, and recall values. Fig. 2, Fig. 3, and Fig. 4 show the support vector machine (SVM) performance. Across all train-test split ratios, SVM classifiers trained with the GWO optimization algorithm consistently achieved the highest accuracy, precision, and recall values. In the 90-10 split, GWO attained an accuracy of 96.49%, precision of 97.22%, and recall of 97.22%. And Fig. 5, 6, and 7 shows the K-Nearest Neighbors (KNN) performance, with GWO consistently demonstrating strong accuracy, precision, and recall rates across different split ratios. In the 90-10 split, GWO achieved an accuracy of 98.25%, precision of 97.30%, and recall of 99.00%. And Fig. 8; Fig. 9; and Fig. 10 show the Logistic Regression (LR) performance. LR classifiers trained with the PSO optimization algorithm consistently achieved high accuracy and precision values across various split ratios. For instance, in the 90-10 split, PSO attained an accuracy of 98.25% and precision of 97.30%. And Fig. 11, 12, and 13 show the Random Forest (RF) performance. RF classifiers trained with the PSO and HPSOGWO algorithms demonstrated competitive precision rates across all split ratios. GWO also performed well, particularly in achieving high recall rates. In Fig. 14, Fig. 15, and Fig. 16, showing the artificial neural network performance, ANN classifiers exhibited strong performance across different optimization algorithms and split ratios. HPSOGWO consistently achieved high accuracy, precision, and recall rates across all splits, showcasing its effectiveness.

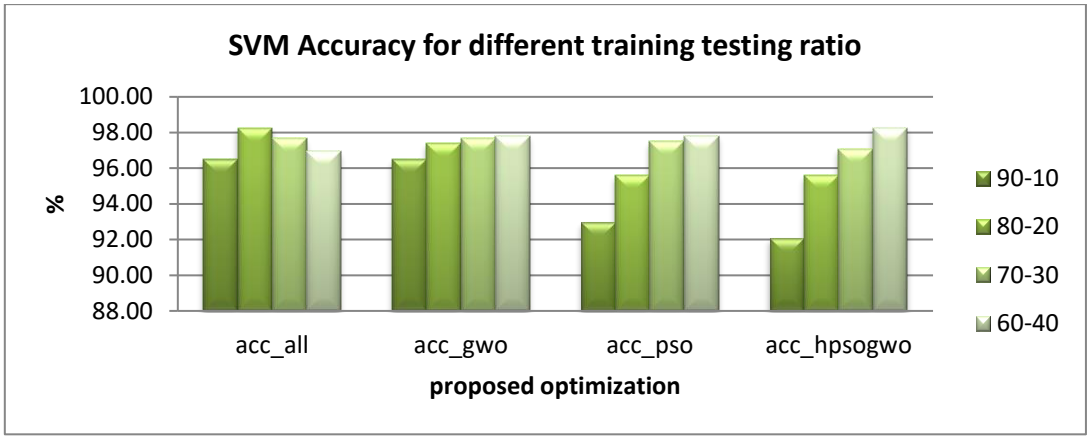

Fig. 2 SVM Accuracy for different training testing ratio

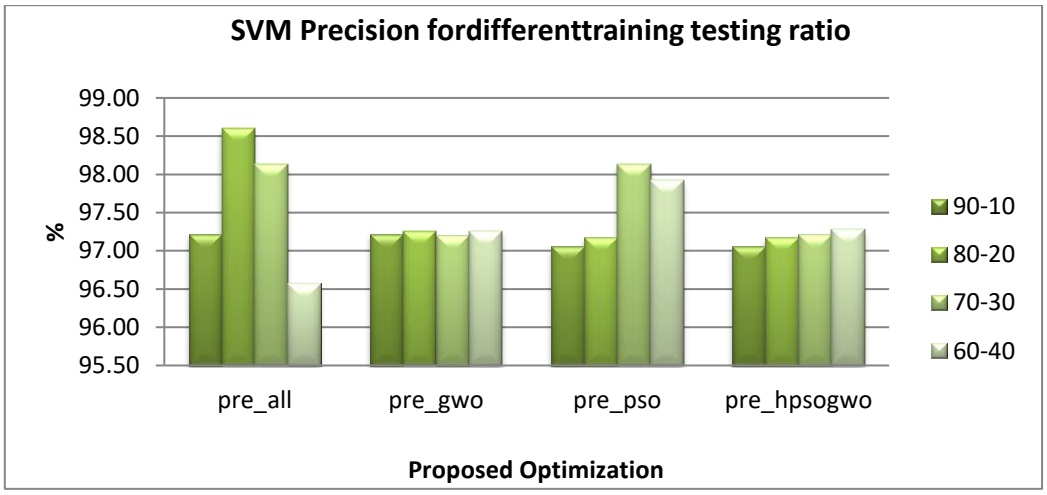

Fig. 3 SVM Precision for different training testing ratios

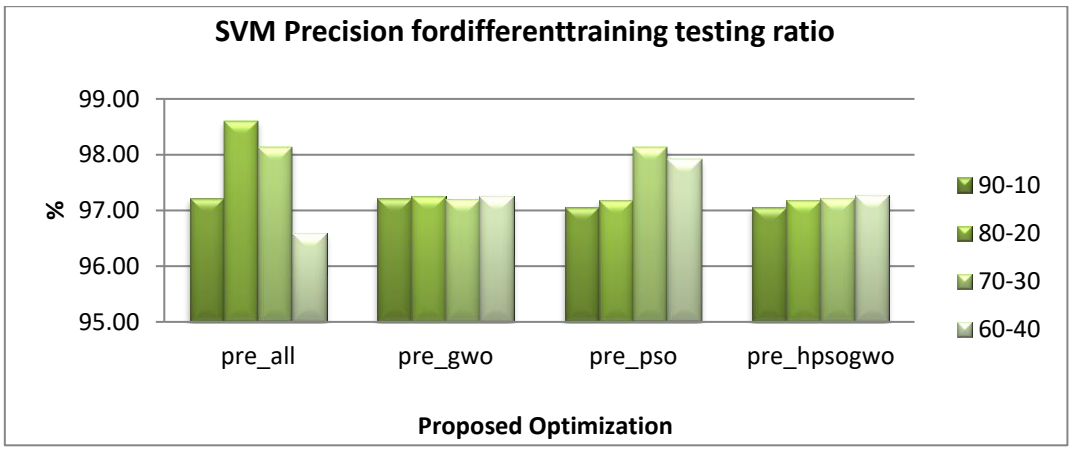

Fig. 4 SVM Recall for different training testing ratio

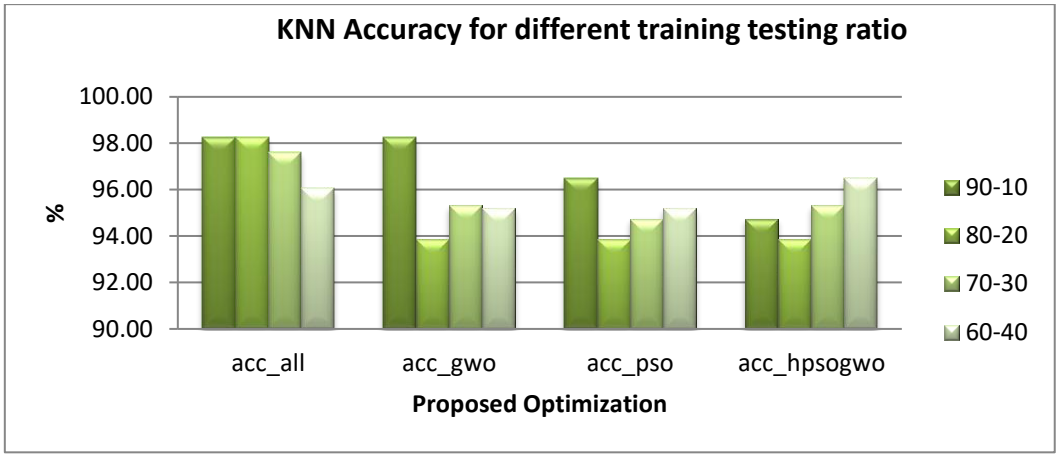

Fig.5 KNN Accuracy for different training testing ratio

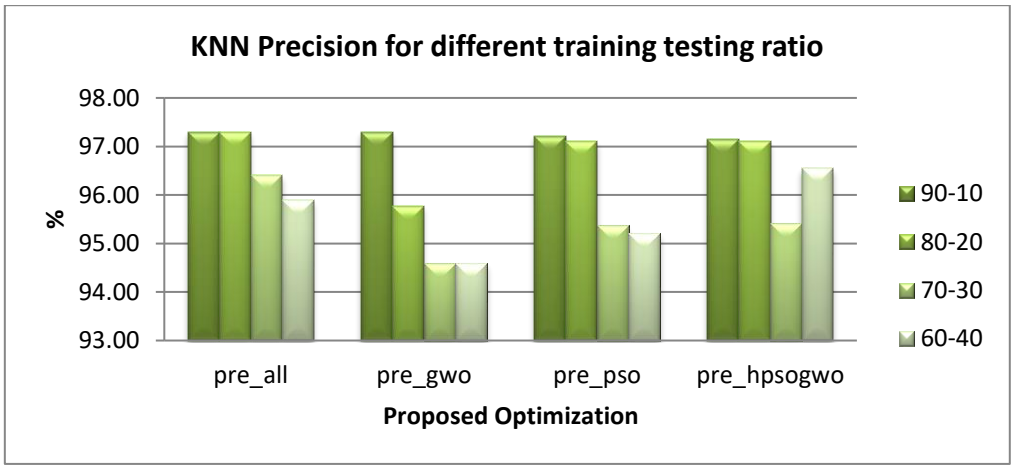

Fig.6 KNN Precision for different training testing ratio

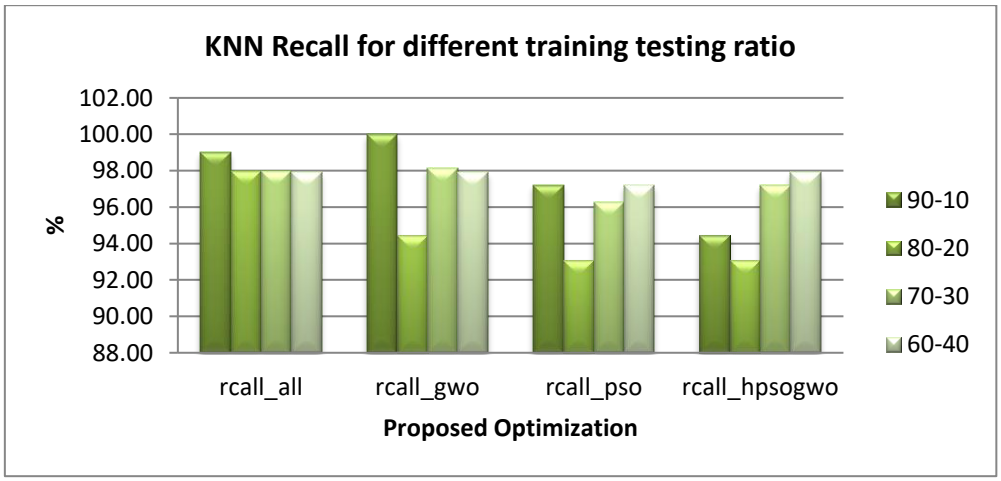

Fig.7 KNN Recall for different training testing ratio

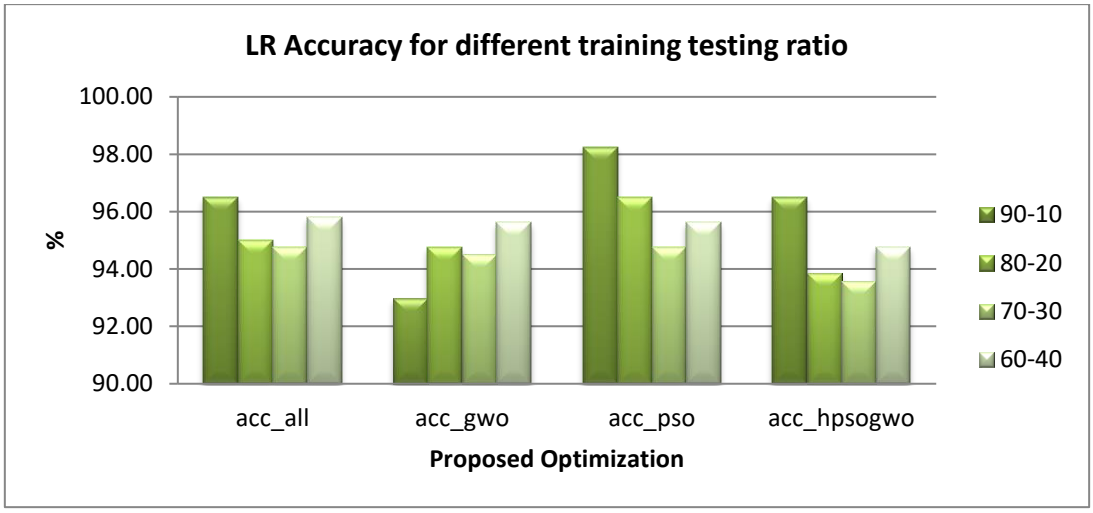

Fig.8 LR Accuracy for different training testing ratio

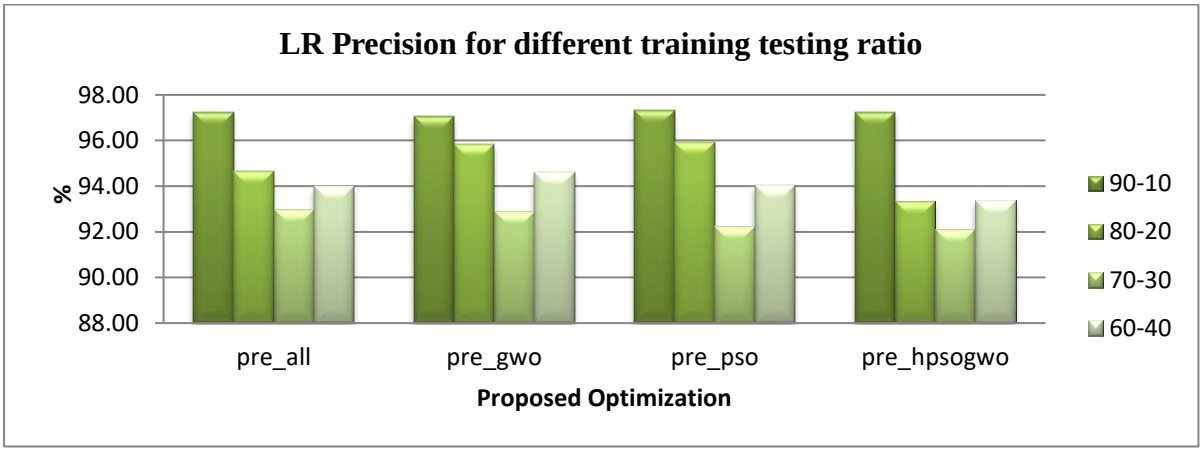

Fig.9 LR Precision for different training testing ratio

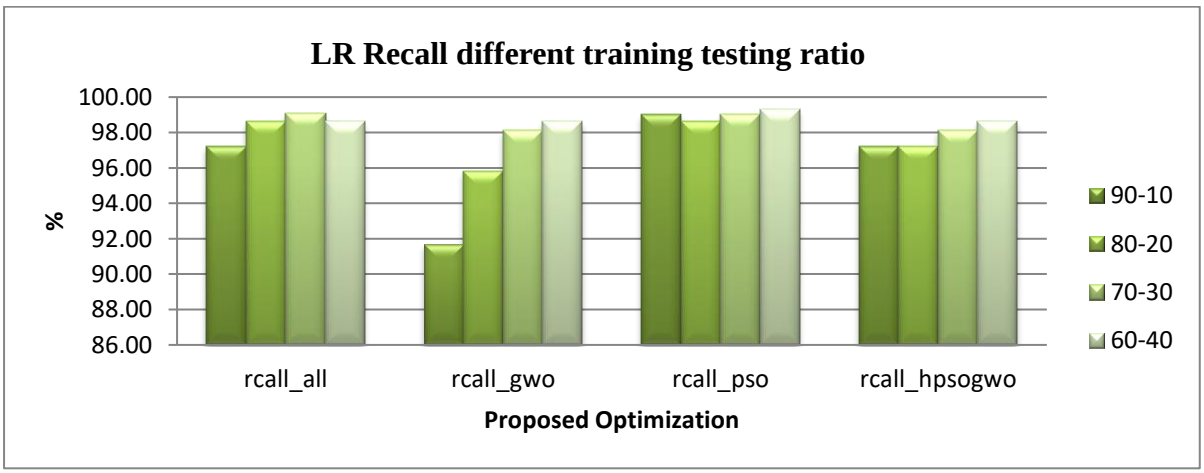

Fig.10 LR Recall for different training testing ratio

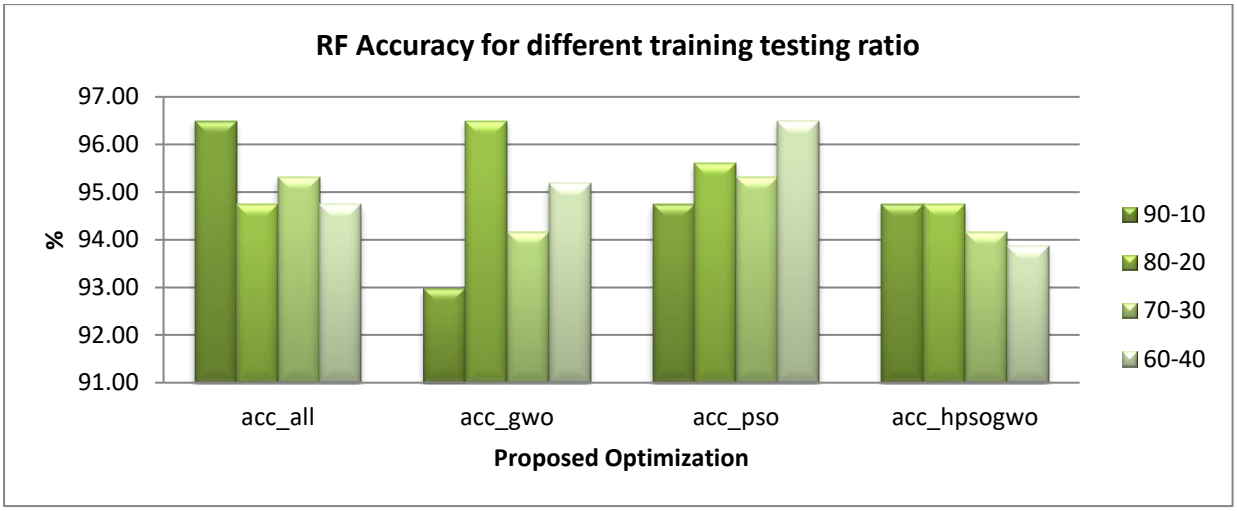

Fig.11 RF Accuracy for different training testing ratio

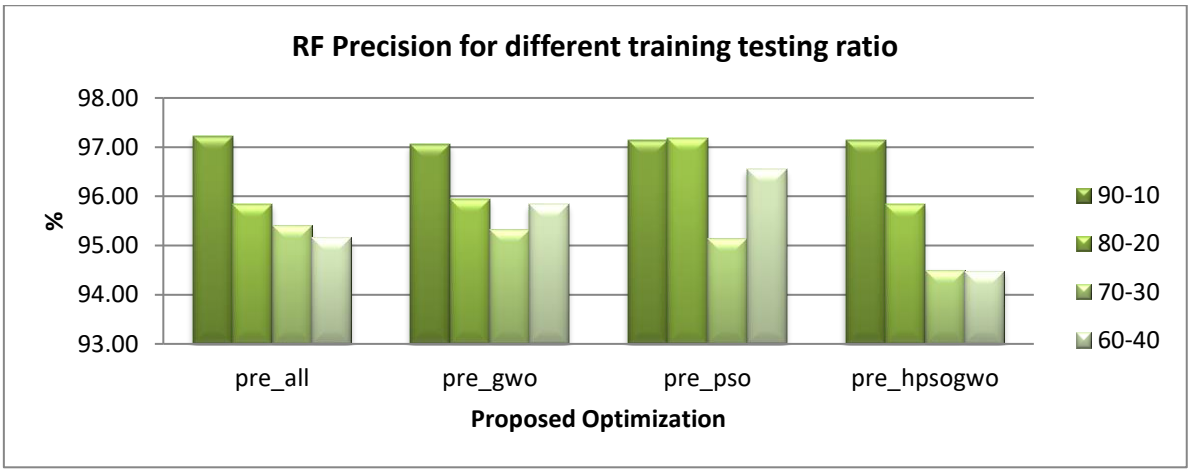

Fig.12 RF Precision for different training testing ratio

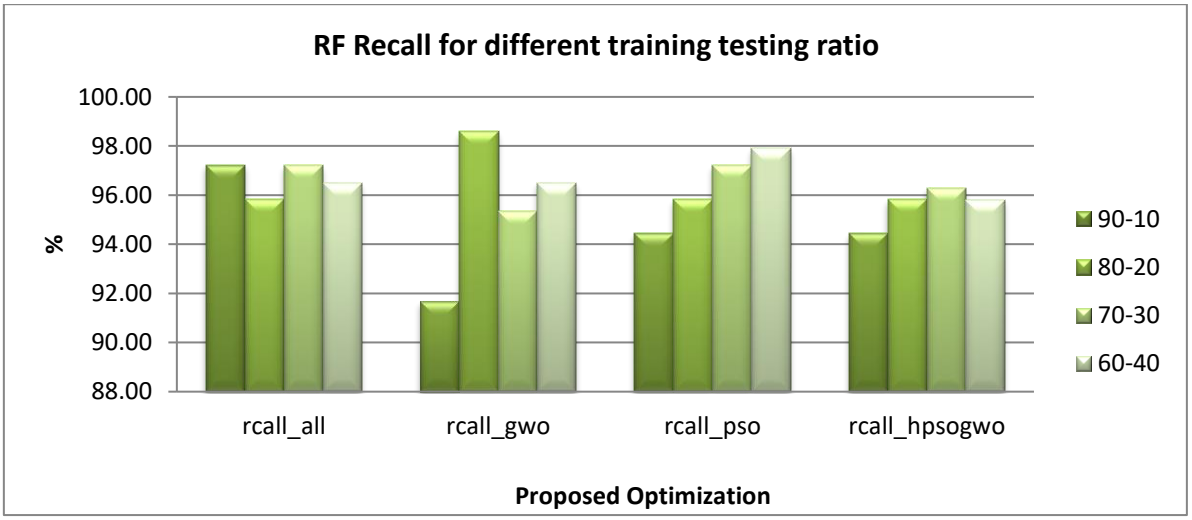

Fig. 13 RF Recall for different training testing ratio

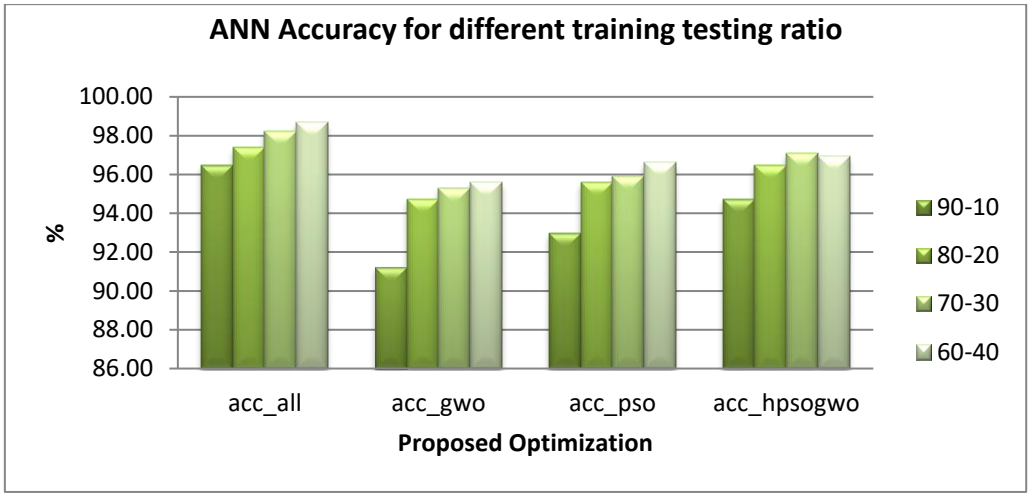

Fig. 14 Accuracy for different training testing ratio

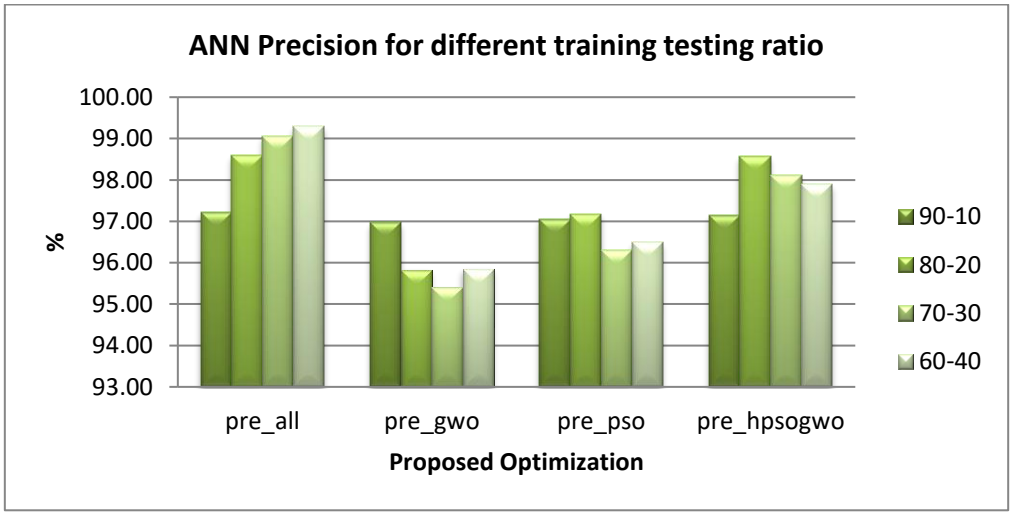

Fig. 15 ANN Precision for different training testing ratio

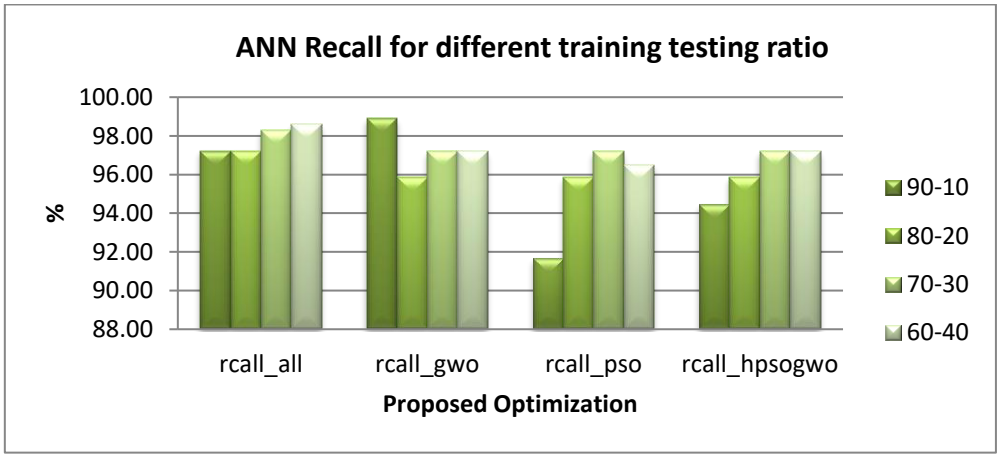

Fig.16 ANN Recall for different training testing ratio

Table 2 number of features selected by three different feature selection methods

	PSO	GWO	HPSOGWO
SVM	13	13	12
KNN	13	10	13
LR	15	8	14
RF	11	10	15
ANN	12	9	16

Table 2 provides the number of features selected by three different feature selection methods. For SVM, PSO and GWO select 13 features each, while HPSOGWO selects 12 features. For KNN, PSO and HPSOGWO select 13 features each, while GWO selects 10 features. For LR, PSO selects 15 features, GWO selects 8 features, and HPSOGWO selects 14 features. For RF, PSO and GWO select 11 features each, while HPSOGWO selects 15 features. For ANN, PSO selects 12 features, GWO selects 9 features, and HPSOGWO selects 16 features.

6. CONCLUSION

This research work provides valuable insights by utilizing various machine learning and feature selection techniques in the application of breast cancer prediction and exploring various supervised learning methods, including logistic regression, SVM, KNN, random forests, and ANN. The choice of feature selection method, such as PSO, GWO, or a hybrid PSO-GWO approach, significantly influenced the performance of the classifiers. These optimization algorithms helped identify the most relevant features from the dataset, improving the efficiency and effectiveness of our predictive models. Furthermore, our study demonstrated the effectiveness of different classifiers in predicting breast cancer outcomes. SVM showed promise in handling high-dimensional data and finding optimal hyperplanes for classification. LR, KNN, and RF also performed well, each offering unique advantages in terms of simplicity, interpretability, and ability to handle nonlinear relationships in the data. Performance evaluation using metrics such as accuracy, precision, recall (sensitivity), and specificity provided a comprehensive assessment of the models' predictive capabilities. Our findings underscore the potential of machine learning for improving breast cancer diagnosis and treatment. By leveraging advanced computational techniques and optimization algorithms, we can enhance the accuracy and efficiency of breast cancer diagnostic systems, ultimately leading to improved patient care and outcomes in the fight against this debilitating disease.

REFERENCES

[1] Sung, H. Ferlay, J. Siegel, R.L. Laversanne, M.Soerjomataram, I. Jemal, A.; Bray, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. 2021, 71, 209–249.

[2] World Health Organization. Breast Cancer. Available online: https://www.who.int/news-room/fact-sheets/detail/breast-cancer (accessed on 19 July 2021).

[3] Hamashima, C. Hattori, M. Honjo, S.; Kasahara, Y.; Katayama, T. Nakai, M.Nakayama, T. Morita, T. Ohta, K. Ohnuki, K.et al. The Japanese guidelines for breast cancer screening. Jpn. J. Clin. Oncol. 2016, 46, 482–492.

[4] Duffy, S.W. Tabár, L.Yen, A.M.F. Dean, P.B.; Smith, R.A.Jonsson, H.Törnberg, S.; Chen, S.L.S. Chiu, S.Y.H. Fann, J.C.Y et al. Mammography screening reduces rates of advanced and fatal breast cancers: Results in 549,091 women. Cancer 2020, 126, 2971–2979.

[5] Wang, L. Early diagnosis of breast cancer. Sensors 2017, 17, 157

[6] Gilbert, F.J. Pinker-Domening, K. Diagnosis and staging of breast cancer: When and how to use mammography, tomosynthesis, ultrasound, contrast-enhanced mammography, and magnetic resonance imaging. 2019; pp. 155–166. ISBN 9783030111496.

[7] Hofvind, S. Holen, Å.S. Aase, H.S.Houssami, N. Sebuødegård, S.Moger, T.A.;Haldorsen, I.S. Akslen, L.A. Two-view digital breast tomosynthesis versus digital mammography in a population-based breast cancer screening programme (To-Be): A randomised, controlled trial. Lancet Oncol. 2019, 20, 795–805.

[8] . Ahuja, A.S. The impact of artificial intelligence in medicine on the future role of the physician. Peer J 2019, 7, e7702.

[9] Abdullah, R. Fakieh, B. Health care employees' perceptions of the use of artificial intelligence applications: Survey study. J. Med. Internet Res. 2020, 22, 1–8.

[10] Doraiswamy, P.M. Blease, C.Bodner, K. Artificial intelligence and the future of psychiatry: Insights from a global physician survey. Artif. Intell. Med. 2020, 102, 101753.

[11] . Blease, C. Kaptchuk, T.J.Bernstein, M.H. Mandl, K.D. Halamka, J.D. DesRoches, C.M. Artificial intelligence and the future of primary care: Exploratory qualitative study of UK general practitioners' views. J. Med. Internet Res. 2019, 21, 1–10.

[12] Meskó, B. Görög, M. A short guide for medical professionals in the era of artificial intelligence. NPJ Digit. Med. 2020, 3, 126.

[13] Kelly, C.J. Karthikesalingam, A.Suleyman, M. Corrado, G. King, D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019, 17, 195.

[14] Asan, O. Bayrak, A.E. Choudhury, A. Artificial intelligence and human trust in healthcare: Focus on clinicians. J. Med. Internet Res. 2020, 22, 1–7.

[15] Sadoughi, F. Kazemy, Z. Hamedan, F. Owji, L. Rahmanikatigari, M. Azadboni, T.T. Artificial intelligence methods for the diagnosis of breast cancer by image processing: A review. Breast Cancer 2018, 10, 219–230.

[16] Abreu, P.H. Santos, M.S.Abreu, M.H. Andrade, B. Silva, D.C. Predicting breast cancer recurrence using machine learning techniques: A systematic review. ACM Comput. Surv. 2016, 49, 1–40.

[17] Li, J. Zhou, Z.Dong, J. Fu, Y. Li, Y. Luan, Z. Peng, X. Predicting breast cancer 5-year survival using machine learning: A systematic review. PLoS ONE 2021, 16, 1–23.

[18] Tabl, A.A.Alkhateeb,A.ElMaraghy, W.Rueda, L.Ngom, A. A Machine Learning Approach for Identifying Gene Biomarkers Guiding the Treatment of Breast Cancer. Front. Genet. 2019, 10, 256.

[19] Alaa, A.M. Gurdasani, D. Harris, A.L. Rashbass, J.van der Schaar, M. Machine learning to guide the use of adjuvant therapies for breast cancer. Nat. Mach. Intell. 2021, 3, 716–726.

[20] L. C Meena;P. M Joe Prathap;S Sankara Narayanan (2023) Detection of Breast Cancer using Curvelet Transform and Adaptive Particle Swarm Optimization Technique 2023 12th International Conference on Advanced Computing (ICoAC) Year: 2023

[21] Nurhayati;Fajar Agustian;Muhammad Dzil Ikram Lubis(2020) Particle Swarm Optimization Feature Selection for Breast Cancer Prediction 2020 8th International Conference on Cyber and IT Service Management (CITSM) Year: 2020

[22] Sannasi Chakravarthy S R;Harikumar Rajaguru;Sundaresan Chidambaram (2022) Processing of Wisconsin Breast Cancer Data using Ebola Optimization Algorithm with Mixture Kernel SVM 2022 Smart Technologies, Communication and Robotics (STCR) Year: 2022

[23] Harish H;Bharathi D S;Pratibha M;Deeksha Holla;Ashwini K B;Keerthana K R (2022) Particle Swarm Optimization for Predicting Breast Cancer 2022 International Conference on Knowledge Engineering and Communication Systems (ICKES) Year: 2022

[24] Maryam Momtahen;Shadi Momtahen;Ramani Remaseshan;Farid Golnaraghi(2023) Early Detection of Breast Cancer using Diffuse Optical Probe and Ensemble Learning Method 2023 IEEE MTT-S International Conference on Numerical Electromagnetic and Multiphysics Modeling and Optimization (NEMO) Year: 2023

[25] Divya Baskaran;Kavitha Arunachalam (2020) Comparison of two global optimization techniques for hyperthermia treatment planning of breast cancer: Coupled electromagnetic and thermal simulation study 2020 IEEE MTT-S International Microwave Biomedical Conference (IMBioC) Year: 2020

[26] Farhad Imani;Zihang Qiu;Hui Yang (2020) Markov Decision Process Modeling for Multi-stage Optimization of Intervention and Treatment Strategies in Breast Cancer 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) Year: 2020

[27] K Vilohit;Bharanidharan N;Harikumar Rajaguru (2022) Improvisation of Decision Tree Classification Performance in Breast Cancer Diagnosis using Elephant Herding Optimization 2022 Smart Technologies, Communication and Robotics (STCR) Year: 2022

[28] Suman Mitra;Sriyankar Acharyya (2023) Identification of disease critical Genes for Triple Negative Breast Cancer, TNBC, using Particle Swarm Optimization, PSO, with Machine Learning 2023 Fifth International Conference on Electrical, Computer and Communication Technologies (ICECCT) Year: 2023

[29] Abd Allah Aouragh;Mohamed Bahaj (2023) Advancing Breast Cancer Diagnosis with Machine Learning: Exploring Data Balancing, Feature Selection, and Bayesian Optimization 2023 IEEE 6th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech) Year: 2023

[30] Alaria, S. K. "A.. Raj, V. Sharma, and V. Kumar.“Simulation and Analysis of Hand Gesture Recognition for Indian Sign Language Using CNN”." International Journal on Recent and Innovation Trends in Computing and Communication 10, no. 4 (2022): 10-14.

[31] Ashwini, K., Raj, A., & Gupta, M. (2016, December). Performance assessment and orientation optimization of 100 kWp grid connected solar PV system in Indian scenario. In 2016 International conference on recent advances and innovations in engineering (ICRAIE) (pp. 1-7). IEEE.

[32] Alaria, Satish Kumar, Ashish Raj, Vivek Sharma, and Vijay Kumar. "Simulation and analysis of hand gesture recognition for indian sign language using CNN." International Journal on Recent and Innovation Trends in Computing and Communication 10, no. 4 (2022): 10-14.

[33] [Yogi, Jyoti, Upendra Singh Chauhan, Ashish Raj, Manoj Gupta, and Simranjeet Singh Sudan. "Modeling simulation and performance analysis of lightweight cryptography for iot-security." In 2018 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE), pp. 1-5. IEEE, 2018.

[34] Singh, Pushpendra Pratap, M. Ram Kumar Raja, Ashish Raj, and Mohammed Abdul Muqeet. "Solution to Interfacing Problems of Programmable Logic Controller in Hardware Replacement." In 2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE), pp. 1-7. IEEE, 2020.