Journal of Computer Science

Research Article

Optimization-Based Feature Selection and Ensemble Machine Learning Algorithms for Breast Cancer Classification

[1,2]Satyabrata Patro, [1]Jyotirmaya Mishra and [2]Bhavani Sankar Panda

[1]*Department of Computer Science and Engineering, GIET University, Gunupur, Odisha, India*
[2]*Department of Computer Science and Engineering, Raghu Engineering College, Visakhapatnam, Andhra Pradesh, India*

Article history
Received: 22-11-2024
Revised: 12-01-2025
Accepted: 07-04-2025

Corresponding Author:
Satyabrata Patro
Department of Computer Science and Engineering, GIET University, Gunupur, Odisha, India;
Department of Computer Science and Engineering, Raghu Engineering College, Visakhapatnam, Andhra Pradesh, India
Email: satyabrata.patro@giet.edu

Abstract: Breast cancer, which originates in a woman's breast tissue, is acknowledged to be a significant study topic in the medical field. For a long time, there has been a serious concern with the classification of breast cancer. Thus, to effectively categorize the breast cancer dataset, machine learning methods are designed and implemented. In previous research, the algorithms have classification accuracy and time complexity issues. This study proposes the use of Enhanced Cuckoo Search Optimization combined with Ensemble Machine Learning Classifiers (EMLC) to tackle the identified challenges and improve the accuracy of breast cancer classification. The system is structured into four key stages: pre-processing, feature extraction, feature selection, and classification. During pre-processing, statistical correlation analysis is applied to eliminate noise from the dataset, thereby enhancing classification performance. The feature extraction phase then derives more informative features from the cleaned data to support more accurate classification. It is performed using Improved Principal Component Analysis (IPCA), which extracts the prominent features from the breast cancer dataset. Then, utilizing the best fitness values of cuckoos, the ECSO algorithm is utilized to identify the relevant and useful characteristics. Finally, using a training and testing model, the EMLC algorithm is employed for classification. It classifies the features more accurately using ensemble Enhanced Granular Neural Network (E-GNN), Adaptive Neural Fuzzy Inference System (ANFIS) and Weighted Support Vector Machine (WSVM) algorithms. The experimental findings show that the proposed EMLC algorithm achieves superior performance compared to existing approaches, offering improved precision, recall, F-measure, accuracy, ROC curve results, AUC scores, and lower time complexity.

Keywords: Breast Cancer Classification, Feature Extraction, Feature Selection, Enhanced Cuckoo Search Optimization and Ensemble Machine Learning Classifiers (EMLC)

Introduction

The second most prevalent illness affecting women in India is breast cancer, which is growing every year. Poor survival rates are still caused by a lack of awareness programs, scheduled viewing, and affordable treatment options. There is the best chance of recovery if these anomalies in breast cancer are found early. We can use mammography for this early prediction. One of the most popular and efficient methods for identifying and screening breast cancer is this one. The fatal risk may be decreased by early identification and effective cancer therapy (Li *et al.*, 2022; Aljuaid *et al.*, 2022). While diagnosing a condition, medical personnel are susceptible to error. Regular detection of breast cancer and subsequent adequate cancer treatment may lower the risk. Every 4-6 weeks, a tumour assessment test is advised. That makes it crucial to distinguish between benign and malignant growths using categorization characteristics (Wu *et al.*, 2020).

Breast cancer is an internal tumour that may be benign or malignant and is caused by unchecked cell division. Researchers have investigated the cause of breast cancer because many risk factors increase a woman's risk. Breast cancer factors include age, genetic risk, and family history (Wuniri *et al.*, 2019). Both local and systemic treatments are available for breast cancer. Surgery, radiation, and other local treatments differ from systemic therapies like chemotherapy and hormone

therapy. Depending on the condition and disease severity, various combinations of both therapies are employed for the best results. Before mining occurs, these records are filtered and cleaned to clear any unnecessary information from the warehouse. Thus, across the breast cancer dataset, feature selection is crucial.

Finding effective feature subsets (more discriminant) and enhancing dataset quality (better and quicker results) are crucial steps in the feature selection process. To improve representation, several subsets of the retrieved characteristics from the provided datasets have been tested (Subasree *et al.*, 2022). The feature selection process and the classification algorithm determine how accurate the breast cancer classifier will be. For the supplied breast cancer dataset, the classifier may get confused and produce inaccurate results if irrelevant and improper characteristics are used Liu & Tang (2014); Singh *et al.* (2023). By minimizing redundant breast cancer dataset characteristics and feature subset selection, an optimization-based solution to this issue is achieved. When estimating kernel density to diagnose breast cancer, it chooses the feature subset and sets the kernel bandwidth.

Predicting categorical labels is the process of classification. Based on the training set and the values of a classifying characteristic, it is used to categorize the data. One of the earliest diagnostic procedures used to detect breast cancer is mammography. Breast cancer recovery rates are known to be significantly increased by early identification. Most medical facilities entrust the analysis of mammograms to skilled radiologists. However, human mistakes may happen at any time. Fatigue of the observer is commonly a factor in errors, which may lead to interobserver and interrater differences. The quality of the images affects the sensitivity of mammographic screening as well (Nasir Khan *et al.*, 2019; Ganesan *et al.*, 2014; Malebary & Hashmi, 2021). Automated methods for detecting and classifying breast cancer images are being developed in an attempt to reduce variability and standardize diagnostic processes. A one-class classification pipeline is used to classify images of breast cancer as either benign or malignant. The sparse distribution of abnormal mammograms simplifies the one-class outlier detection problem. The features were extracted using trace, which is a generalization of the Radon transform. A variety of recently developed and deployed mammography image analysis methods have been created to provide clinically relevant features. There are several different classifiers available, including the Gaussian Mixture Model (GMM), nearest classifier, nearest neighbor, linear discriminant classifier, and quadratic discriminant classifier.

Machine Learning (ML) is widely used for classifying breast cancer patterns because of its ability to effectively identify important features within complex breast cancer datasets (Naseem *et al.*, 2022). To determine whether the data is related to breast cancer or not, SVM, a supervised pattern classification technique, has been used in a dataset of breast cancer cases. The SVM algorithm performs binary classification on the given breast cancer dataset. Using a hyperplane, creating a binary classifier that separates class members from non-members in the input space is easy. After mapping to a higher-dimensional feature space, SVM separates using a maximum margin hyperplane and finds a nonlinear decision function in the input space (Asri *et al.*, 2016). Sparsely linear points form the separation hyperplane and are represented by the system using support vectors, which are automatically recognized as a subset of informative points.

This research project mainly aims to categorize breast cancer using an ensemble machine-learning algorithm. The accuracy of detection is not guaranteed despite much research and techniques. Current approaches suffer from noise and misclassification. The Ensemble Machine Learning Classifiers (EMLC) method and Enhanced Cuckoo Search Optimization (ECSO) technique are developed in this study to address the problems mentioned above and enhance classification performance overall. The study's main contributions include pre-processing, feature extraction, selection, and classification. With the help of efficient algorithms, the suggested strategy produces more accurate findings for the provided collection of breast cancer images.

Related Work

Sameti *et al.* (2009) suggested that a retrospective analysis of screening mammography performed before discovering a malignant mass be done to extract image features for detecting breast cancer early. In all, 58 individuals with breast cancer who had a positive biopsy were examined. Each mammography that was performed 10 to 18 months before cancer was discovered is assessed. There are two regions identified on each mammographic projection of the atypical breast: (1) Area one, which matched the location of the cancerous tumour that later formed, and (2) an area that, on the same mammography, resembled region one. In each projection of the normal breast, a third area corresponding to region one on the opposite breast was found. Sixty-two textural and photometric image properties are then computed for each highlighted place. According to a stepwise discriminant analysis, six characteristics may most effectively discriminate between normal and pathological areas. The average classification using the best linear classification function was 72%.

In Hamed *et al.* (2021), the INbreast dataset's full-field digital mammograms are generally subjected to three steps. The mammograms are pre-processed to eliminate further artefacts before being cropped into small, overlapping slices. Second, after establishing the

YOLO-V4 model, masses are located using two different methods: Whole mammograms and cropped slice detection. Third, different feature extractors are utilized to categorize the localized lesions and compare their performance to YOLO. These include ResNet, VGG, Inception, etc. YOLO-V4 as a detector with two channels for comprehensive mammography and cropped slices in a study to reduce data loss by reducing large mammograms using experimental results. Our methodology is more accurate than previously launched breast cancer screening techniques overall in locating the masses. Additionally, it has a more precise capacity to differentiate between benign and malignant tumours.

In Padmanabhan & Sundararajan (2012) MATLAB, a widely used simulation tool, and the MIAS dataset were used to improve the present accuracy (diagnostic) of digital mammograms. The method entails identifying and categorizing tumour cells according to various disease stages. We consider mammography's object detection, identification, and classification processes to distinguish between normal and malignant (benign or cancerous) cells. It is said that thick breasts might make reading conventional mammograms more challenging. There are hardly any places where you can get such costly digital mammograms, despite the claims made by newer mammography methods for greater detection in thick breast tissues. By employing the MATLAB numerical analysis program for image processing applications to analyze various breast structures (mammograms), this issue may be reduced to a minimum. Results showed an increase in accuracy of up to 91% from the current level of 70%. The technology has successfully detected early breast cancer in several breast tissues.

Baskaran *et al.* (2011) state that many healthcare applications have successfully used recently emerging machine learning-based prediction. A recent development is the use of machine learning to anticipate breast screening attendance before the mammography. New predictor characteristics for such an algorithm are presented. To make predictions, it discusses a novel hybrid technique that uses neural networks with radial basis functions and back-propagation. The algorithm was created in an environment that relies on open-source software. A 13-year dataset (1995-2008) was used to evaluate the technique. This study compares the algorithm, verifying its efficacy and accuracy across several platforms. The algorithm's results show a nearly 80% accuracy rate and an 88% positive predictive value and sensitivity. The 40-50% negative predictive value and high specificity need further study. The method should be tested on a larger scale since the first findings are encouraging and give plenty of justification.

Loizidou *et al.* (2022) introduced machine learning for the automated segmentation and categorization of masses and the removal of temporally consecutive digital mammograms. Two radiologists identified mass places on 320 images from 80 patients (two time periods and two breast views) and a dataset specifically constructed for this research was used to assess the algorithm's performance. In a leave-one-patient-out and k-fold cross-validation procedure, 96 features are retrieved and 10 classifiers are evaluated. Neural networks provided 99.9% accurate mass detection. Instead of state-of-the-art temporal analysis, the technique improved mass classification from 92.6 to 98%. The statistical analysis revealed a significant improvement, as shown by a p-value of 0.05. In conclusion, these findings show that detecting breast masses may be effectively accomplished by subtracting the results of temporally subsequent mammograms. According to a statement defining the algorithm's clinical and translational significance, the approach may help build automated breast cancer computer-aided diagnostic tools that improve patient prognosis.

In the study by Huang & Chen (2022), the authors proposed the Hierarchical Clustering Random Forest (HCRF) model, which integrates hierarchical clustering with decision tree analysis to enhance classification performance. The methodology involves applying hierarchical clustering to a set of decision trees within a random forest, enabling the identification and selection of representative trees from distinct clusters characterized by low inter-tree similarity and high predictive accuracy. Additionally, the Variable Importance Measure (VIM) technique is used to enhance feature selection in order to improve the prediction of breast cancer. The UCI Machine Learning Repository's Wisconsin Breast Cancer (WBC) and Wisconsin Diagnostic Breast Cancer (WDBC) datasets were used to assess the model. For evaluation, performance criteria including the Area Under the ROC Curve (AUC), sensitivity, specificity, accuracy, and precision were used. Results from experiments show that the HCRF model produces better classification accuracy when paired with VIM-based feature selection.

In the study by Zhou *et al.* (2020), the authors introduced a novel Multi-Objective Feature Selection (MO-FS) method that simultaneously considers sensitivity and specificity as dual objective functions during the feature selection process. To enhance the efficiency and adaptability of the algorithm, they proposed a modified entropy-based termination criterion that dynamically determines the stopping point of the optimization process, eliminating the need for a predefined number of generations. Using the evidential reasoning approach, we also developed a technique for automatically choosing the best option from the Pareto-optimal set for multi-objective learning. Additionally, we created an adaptive mutation procedure to calculate the mutation probability in MO-FS automatically. Results: We assessed the MO-FS to determine the malignancy of breast lesions in digital breast tomography and lung nodules in low-dose CT. According to the experimental findings, MO-FS's feature set outperformed feature

selections made by other widely used approaches regarding classification performance. The technique is more flexible and successful in selecting radionics features. In Lai *et al.* (2014), a novel multiscale 2-D singular spectrum analysis with PCA (2-D-MSSP) for noise-robust feature extraction and classification of Hyperspectral Images (HSI). Unlike traditional methods that require pre-corrected data, 2-D-MSSP effectively extracts spatial-spectral features while removing noise in both domains. It achieves high classification accuracy even on uncorrected datasets and with limited training samples. Experimental results validate its robustness against 10 state-of-the-art classifiers.

Two sophisticated automated Breast Cancer (BC) classification methods were created in the study by Elkorany *et al.* (2022) by combining the Dragonfly Algorithm (DA) and Whale Optimization Algorithm (WOA) with Radial Basis Function Kernel Support Vector Machines (RBF-SVM). The main goal was to increase classification accuracy through effective SVM parameter optimization. The Wisconsin Diagnostic Breast Cancer (WDBC) and Wisconsin Breast Cancer Database (WBCD) datasets were used to assess the performance of the suggested WOA-SVM and DA-SVM models. Classification accuracy (CA), confusion matrix, area under the ROC curve (AUC), sensitivity, and specificity were among the evaluation criteria. Additionally, the study contrasted the suggested models with well-known optimization techniques including Genetic Algorithm (GA) and Particle Swarm Optimization (PSO), which are frequently used to train SVM and Artificial Neural Network (ANN) classifiers. In order to evaluate the efficacy of the WOA-SVM and DA-SVM approaches, their feature selection skills were also examined and compared to those of other models. The experimental results demonstrate that the WOA-SVM approach performs better than previous classification methods on the WBCD dataset.

Materials

The experimental design of this study followed a structured pipeline comprising dataset preparation, pre-processing, feature extraction, feature selection, and classification using ensemble machine learning models. The CBIS-DDSM breast cancer dataset from Kaggle, containing annotated digitized mammograms with Regions of Interest (ROIs), was utilized for training and evaluation. To enhance input quality, a Statistical Correlation Coefficient (SCC)-based pre-processing technique was employed to remove noisy and irrelevant data. Feature extraction was conducted using Improved Principal Component Analysis (IPCA) to reduce dimensionality while preserving significant information. Subsequently, the Enhanced Cuckoo Search Optimization (ECSO) algorithm was applied to select the most relevant features based on a fitness function incorporating classification accuracy, relevance, and redundancy. For classification, an ensemble machine learning classifier (EMLC) was developed by combining Enhanced Granular Neural Network (E-GNN), Adaptive Neuro-Fuzzy Inference System (ANFIS), and Weighted Support Vector Machine (WSVM), with final predictions derived through a voting mechanism. Model performance was assessed using accuracy, precision, recall, F1-score, ROC, AUC, and execution time, and k-fold cross-validation was implemented to ensure result robustness and generalizability.

Proposed Methodology

This research hopes to improve the classification process by examining breast tumour features. This study first incorporated a statistical correlation analysis-based pre-processing to improve the classifier's performance. Then, the features are extracted from the Region of Interest (ROI) mammogram images by Improved Principal Component Analysis (IPCA). Then, the prominent features are selected via the Enhanced Cuckoo Search Optimization (ECSO) algorithm. We employed ensemble learning in the suggested strategy to increase accuracy. The Enhanced Granular Neural Network (E-GNN), Adaptive Neural Fuzzy Inference System (ANFIS), and Weighted Support Vector Machine (WSVM) algorithms make up the three Machine Learning (ML) classifiers that make up the ensemble voting system.

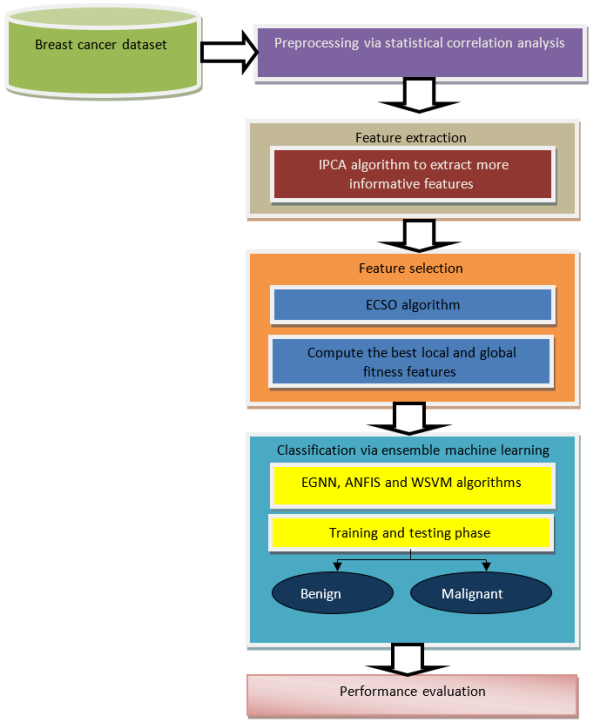

Fig. 1: Overall methodological approach suggested

Figure (1) illustrates the overall process of the proposed methodology. The data are taken from the Breast Cancer Kaggle Dataset (Kaggle, 2021). The first

step in processing any dataset is cleaning it to remove typical issues like noise, outliers, and missing values that are frequent in healthcare data. Next, we use feature scaling and normalization approaches to ensure that every feature has an equal impact on the model training. Improved Principal Component Analysis (IPCA) may reduce the data's dimensionality and extract useful features while keeping the data's fundamental structure intact. Ensuring patient confidentiality and conformity with data protection standards like HIPAA or GDPR are essential ethical issues in this study. For security reasons, no personally identifying information is used and the dataset is anonymized to make it impossible to abuse. To prove that the ECSO-EMLC technique is effective, it is compared against baseline approaches, including decision trees, classical logistic regression, and standalone Support Vector Machines (SVM). The technology has the potential to surpass traditional diagnostic methods in detecting breast cancer, as shown by the findings that demonstrate substantial increases in classification accuracy, sensitivity, and specificity.

Pre-Processing Using Statistical Correlation Coefficient (SCC)

As the average of all related values of the second variable (the dependent, result, response, or variable being explained) and the degree of the link between the two variables, the correlation may be defined (Xiong *et al.*, 2005). When the variable x serves as a random covariate to the variable y meaning both x and y are continuous variables that vary together, the sample correlation coefficient r_{xy} (r) (denoted as r) is employed to evaluate the strength and direction of their linear relationship. In this context, statistical analysis focuses on assessing the linear association between the variables rather than predicting one from the other.

Graphing and numerical approaches may evaluate a regression equation's fit. To test a linear regression fit with any number of predictors, the sample coefficient of determination (R^2) is useful. For a homoscedastic model (wi = 1) with a constant term, R^2 is the regression sum of the ratio of the square (SS_{Reg}) to the total sum of squares of mean deviation (S_{YY})*:*

$$R^2 = \frac{SS_{Reg}}{S_{YY}} = \frac{S_{YY}-SSE}{S_{YY}} = 1 - \frac{SSE}{S_{YY}} = 1 - \frac{(y_i-\hat{y}_i)^2}{(y_i-\overline{y_i})^2} \quad (1)$$

Where ŷ represents the anticipated value of y and ŷ represents the mean of y's values, summed across i = 1, 2,... n. SSE is a residual sum. R^2 = 1−SSE/SST, where *SST* is the sum of squares *of* y^2 in a model without a constant term. The R^2 in Eq. (1) indicates the percentage of variance around the mean $\overline{y}$ explained by the regression. Thus, adding the independent variable x reduces the total variance of $\overline{y}$ when R^2 is large, multiplying by 100 gives it as a percentage. Since 0 ≤ *SSE* ≤ *SSYY*, R^2 does not exceed 1. The correlation between y and ŷ is *R*:

$$R = r_{y\hat{y}} = \frac{\sum(y_i-\hat{y}_i)(y_i-\overline{y_i})}{[\sum(y_i-\hat{y}_i)][(y_i-\overline{y_i})]^{\frac{1}{2}}} \quad (2)$$

Usually termed the multiple correlation coefficient. Comparing R^2 of equations with differing coefficient counts from the same data set is inappropriate. Even so, we appreciate R^2 in the regression printout. The square of the correlation coefficient between x and y is used as the coefficient of determination in simple regression with a constant term, thus the notation:

$$r_{xy} = \pm\sqrt{R^2 = \sqrt{1 - \frac{S_{YY}-a_1^2 S_{XX}}{S_{YY}}}} = a_1\sqrt{\frac{S_{XX}}{S_{YY}}} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} \quad (3)$$

The slope, a1, of the fitted regression line determines whether this metric is positive or negative. If R^2 is unity, each regression line point and variance are fitted. A zero coefficient means that the regression line is horizontal and y is not a function of x. Regression coefficients are similarly connected to the r_{xy} correlation in more generic regression situations, although more complicatedly.

The x and y covariance is the expected value of the product of their deviations from their expected values with a joint normal distribution. It measures the correlation between two items. The sample covariance is:

$$cov\,(x,y) = \frac{1}{n-1}\sum(x_i-\overline{x})\,(y_i-\overline{y}) \quad (4)$$

$$cov\,(x,y) \leq s_x s_y \quad (5)$$

Which means now that $r \leq 1$.

An indicator of x-y correlation is covariance. Depending on the slope, the covariance of two linearly connected variables will be positive or negative. If x and y are uncorrelated, the covariance is zero. However, highly dependent random variables, frequently nonlinear, might have zero covariance (correlation). Variance is a random variable's covariance with itself, although introductory textbooks disregard it. The standard deviation (σ) is the square root of variance and is always positive for populations and samples. In circumstances of actual uncertainty, covariance must be considered.

Feature Extraction Using Improved Principal Component Analysis (IPCA)

In order to identify latent information and categorize data in order to differentiate between benign and malignant tumors, the Improved Principal Component Analysis (IPCA) algorithm is utilized for feature extraction in this study. By making deductions from characteristics like trouble interacting with others, unusual hobbies, a range of skill levels, atypical sensory responses (hypersensitivity or hyposensitivity to sight, touch, taste, smell, or hearing), and repetitive behaviors or body movements, the suggested system is intended to determine whether a condition is normal.

Principal Component Analysis (PCA) is utilized to reduce the high intrinsic dimensionality of the feature space by transforming it into a lower-dimensional subspace, particularly effective when variables exhibit strong correlations. This dimensionality reduction is achieved by eliminating irrelevant components, thereby enabling a more compact representation of the dataset. As a classical multivariate analysis technique, PCA facilitates linear feature extraction and is applied in this research to generate informative feature vectors for the breast cancer dataset.

However, standard PCA may result in the loss of critical feature information, especially when applied to small datasets, and may not effectively preserve class-relevant information during compression. To address these limitations, a modified version of PCA is introduced, enhancing its capability to retain essential discriminatory features and improve classification performance.

In IPCA, normalize the jth element y_{ij} of the ith feature vector y concerning its standard deviation, $\sqrt{\lambda_j}$, to minimize the impact of high eigenvalues. Thus, the new feature vector $y_i^{'}$ is expressed as:

$$y_i^{'} = \left[\frac{y_{i0}}{\lambda_0}, \frac{y_{i1}}{\lambda_1}, \ldots . \frac{y_{i(r-1)}}{\lambda_{r-1}}\right] \quad (6)$$

Normalized feature vectors are used to create a new feature subspace. By normalizing feature vectors by the square root of their eigenvalues, this technique calculates the separation between training and testing features.

This Equation may represent the linear transform (PCA):

$$Y = TX \quad (7)$$

X, Y, and *T* are the original, altered, and transformed matrix vectors. The Equation for solving the transform matrix T:

$$(\lambda I - S)\, U = 0 \quad (8)$$

Use the matrices I, S, U, and λ, the original image covariance matrix, a square matrix with unity along the diagonal, eigenvectors, and eigenvalues. To calculate U_j and λ_j $(j = 1, 2, \ldots m)$, use Eq. (2), with eigenvalues arranged as $\lambda_1 \geq \lambda_2 \geq \ldots . \geq \lambda_m$. You may write the eigenvectors as $U = [U_1, U_2, \ldots \ldots, U_m]$.

The IPCA generated the transformed matrix T' from breast cancer dataset training samples appropriate to a specified application. Here's the Equation:

$$Y = T'X \quad (9)$$

$$V_N = b_1 u_1 + b_2 u_2 + \ldots . + b_N u_N \quad (10)$$

When comparing Eqs. (7 and 8), The primary changes are the transform matrix, the covariance matrix of training samples, and the breast cancer dataset.

The primary advantage of the Improved Principal Component Analysis (IPCA) lies in its ability to reduce dimensionality without significant information loss by eliminating redundant data. PCA, when examined through statistical and mathematical methods such as eigenvalues and eigenvectors, provides a clearer understanding of its mechanism. The IPCA technique employs a mathematical framework that projects high-dimensional data into a lower-dimensional space, with the eigenvectors of the covariance matrix representing this reduced space. The error-minimizing and decorrelation properties of IPCA are particularly beneficial for this study, enabling the extraction of meaningful data and crucial acoustic features for breast cancer classification.. Mean and standard deviation are among the characteristics of the normal and breast cancer patient input data that were gathered:

$$Mean = \frac{Sum\ of\ No.\ of\ Data}{Total\ Number\ of\ Data} \quad (11)$$

Typical Variation Standard deviation is sometimes called root-mean-square since it indicates the square root of the squared difference from the arithmetic mean:

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} \quad (12)$$

Algorithm 1: IPCA

Input: Pre-processed breast cancer dataset
Output: Extracted informative features
Start

1. Calculate the mean value S' of the given breast cancer dataset S
2. Subtract the mean value from each data point in the dataset to center the data
3. Compute the covariance matrix C from the matrix A: $C = AA^T$
4. Find the eigenvalues $V_1, V_2, V_3, V_4, \ldots, V_N$ of the covariance matrix C
5. Calculate the eigenvectors corresponding to the eigenvalues of matrix C
6. Represent any vector S as a linear combination of these eigenvectors
7. Select only the largest eigenvalues to form a reduced-dimensional dataset based on their significance
8. Match the combination of features in the breast cancer dataset using the selected eigenvectors
9. Compute the features using the mean and standard deviation (11 and 12)
10. Extract the most informative features from the dataset based on the selected eigenvectors and computed statistics

End

The IPCA algorithm extracts the maximum and minimum occurring synchro states and obtains more informative features.

Feature Selection Via ECSO Algorithm

To choose important and pertinent features from the given breast cancer dataset, the Enhanced Cuckoo Search Optimization (ECSO) approach is used in this part. The algorithm selects features as efficiently as possible by

using the best fitness values. Inspired by nature, Cuckoo Search (CS) is a contemporary metaheuristic optimization algorithm that seeks out the best answers. The aggressive reproductive approach and unique sounds of Cuckoos are well-known. Brood parasitism is the term for the practice of certain species, such as the Ani and Guira cuckoos, laying their eggs in other birds' nests. This strategy increases the likelihood of their eggs hatching by taking advantage of the host bird's resources. If the host bird detects the foreign eggs, it may either destroy them or abandon the nest, prompting it to build a new one. This concept of brood parasitism forms the foundation of the Cuckoo Search Optimization (CSO) algorithm (Alia & Taweel, 2021; Mohsin *et al.*, 2020). A cuckoo egg represents a new answer for each egg in the nest. A cuckoo may replace the nest solution if it's preferable. Levy flights, not isotropic random walks, are better for improving this CSO method.

The conventional cuckoo search has been described using the three idealized rules below:

- The reproductive behaviour of cuckoos involves the deposition of a single egg in a nest selected at random
- The next generations will receive the most optimal habitats and the eggs of the highest excellence
- The quantity of available host nests remains constant and the host bird has a probability ranging from 0-1 of discovering the egg laid by the cuckoo. The bird host has two potential courses of action: Either expel the egg from its nest or elect to forsake it entirely, then construct a new nest.

A fraction of the n host nests is replaced with new nests, which can be utilized to approximate the final hypothesis. The CSO has a large search area and is relatively straightforward. Instead of using the typical random walk to do a global search, it employs a Levy flight, allowing CSO to scour the search area more effectively.

Numerous host nests with plenty of eggs might improve CSO (Gao *et al.*, 2019). The three types of nests that cuckoos typically choose for their egg-laying are as follows. The common cuckoo selects several host nests that lay eggs with similar characteristics. Other cuckoos choose a variety of host nests with eggs that are distinct from their own. In contrast to the light-coloured eggs of their host birds, certain other cuckoo species lay cryptic, black eggs. This tactic is utilized to keep the eggs hidden from the host in cuckoos that parasitize hosts and have dark, rounded nests.

Initial Population

This research uses each egg to represent a possible set of features to identify the data properly. The top-mated features from the breast cancer dataset statistical analysis determine the characteristics.

Finding New Solutions and Levy Flight

The Levy flight is a technique for discovering novel solutions to Eq. (4) based on ECSO. A levee walk around the best solution so far should be used to produce some of the new alternatives; this will speed up the local search (Ma *et al.*, 2021). The new solution for cuckoo i is obtained via Levy flying and is shown below as $x_i^{(t+1)}$:

$$x_i^{(t+1)} = x_i^{(t)} + C \oplus Levy\left(s, \lambda\right) \tag{13}$$

The step size is t. The step length follows the Levy distribution:

$$Levy\left(S, \lambda\right) \sim s^{-\lambda}, 1 < \lambda \leq 3 \tag{14}$$

Crossover and Mutation

- Common cuckoos produce two eggs in their nests via a crossover and the best egg is selected.
- Using crossover and the uniform mutation operator, a European-type cuckoo produces two eggs, from which it selects the better one.
- If otherwise, a random solution generates eggs (cryptic)

Fitness Function

The selection method heavily relies on the fitness function. The salient subset characteristics of the breast cancer dataset are successfully chosen by applying the optimal fitness function values. This means that the fitness function, which the Equation may represent, incorporates relevance and redundancy to direct CSO in finding the optimal feature subset:

$$fitness\left(fi\right) = \alpha \times D - \left(1 - \alpha\right) \times R \tag{15}$$

Where:

$$D = \frac{1}{|S|}\sum_i I\left(xi, C\right) \tag{16}$$

$$R = \frac{1}{|S|^2}\sum_i \sum_j I\left(xi, xj\right) \tag{17}$$

C stands for the class designation and X for the selected collection of qualities. Every selected feature and class label has a discrete random variable associated with it. Through pairwise calculations, D determines how relevant the selected feature subset is to the class labels. Each pair of selected features' mutual information is evaluated by R, which identifies any duplication in the subset of features chosen. In the chosen feature subset, fitness is a maximizing function to increase relevance (D) and concurrently reduce redundancy (R). $\alpha$is a constant value and $\epsilon\left[0, 1\right]$. The relevant significance in fitness function is indicated by α. (1- α) indicates the proportionality of redundancy reduction. Since relevance is deemed more significant than redundancy, we set α to be bigger (1- α) than in the fitness function. ECSO selects the key and essential subset characteristics from the dataset more effectively.

Algorithm 2: Enhanced Cuckoo Search Optimization (ECSO)

Input: Breast Cancer Dataset
Fitness function: Classification Accuracy
Output: Optimally Selected Feature Subset
Initialize:
Generate an initial population of *m* eggs distributed among *n* host nests (each egg represents a feature subset)
Repeat until stopping criteria are met (t < MaxGeneration or convergence condition):
For each nest **do**
Select a **cuckoo type** randomly:
If cuckoo_type == common cuckoo:
1. Apply crossover on the two best eggs in the nest to generate two offspring
2. Evaluate and select the better offspring
Else if cuckoo_type == European cuckoo:
1. Select any two eggs randomly from the nest
2. Perform crossover and uniform mutation to create two new eggs
3. Evaluate and select the better egg
Else (cryptic cuckoo):
Generate a random egg (solution)
Fitness Evaluation:
Evaluate the fitness f_i of the newly generated egg using classification accuracy (as per Eq. 15)
Selection and Replacement:
1. Identify the egg with the weakest fitness in the nest, say f_j
2. **If** $f_i > f_j$, replace egg *j* with the new egg *i*
Ranking and Sorting:
1. Rank all eggs in the population based on fitness values
2. Retain the best solution(s)
Levy Flight Mutation:
Abandon a portion of the worst-performing eggs
Generate new eggs using Levy flights (as per Eqs. 13 and 14)
End for each nest
Update the best egg (solution) across all nests
Feature Selection: Based on the final fitness ranking, select the most relevant and significant features for classification

According to the algorithm, subset features are chosen using objective function fitness values. The levy flight uses the best new solutions and reduces unneeded features. To make the breast cancer dataset more informative, the ECSO optimization technique is applied.

Classification Using Ensemble Machine Learning Classifiers (EMLC)

Ensemble methods are meta-algorithms that combine several machine learning techniques to create an ensemble model. Ensemble approaches have been successful in winning many prestigious machine learning competitions because they usually produce better predictions than a single model. Ensemble approaches in machine learning use a variety of algorithms to get better results and performance than any one method could. A machine learning ensemble has a finite number of possible models, as opposed to a statistical ensemble, which can theoretically have an infinite number. More flexibility in the ensemble methods' design is made possible by this differentiation (Dietterich, 2000).

A Voting Classifier is a machine learning model that aggregates the predictions from multiple individual models to produce a more accurate output. The predictions are made based on a voting mechanism, where the majority vote determines the final result. Unlike separate models that are individually trained and evaluated, ensemble learning combines these individual models into a unified model, where the final prediction is based on the majority vote. Voting classifiers are typically categorized into two types: Hard voting classifiers and soft voting classifiers. Ensemble algorithms have demonstrated exceptional performance on challenging datasets, often setting new performance benchmarks. In this study, three machine learning classifiers—EGNN, ANFIS, and WSVM are employed to classify the given dataset. Ultimately, the classification result is derived based on the voting outcomes.

Enhanced Granular Neural Network (E-GNN)

Artificial neural networks made to process numerical or granular data are known as Granular Neural Networks (GNN). Online incremental learning from data streams is the main goal of the GNN technique (Song & Pedrycz, 2013). As shown in Figure (2), there are two primary processes in the GNN learning process. First, the original numerical data is transformed into information granules, which are intervals or, more generally, fuzzy sets. Then, using this information rather than the actual raw input, the neural network learns, adapts, and refines. The neural network does not have to analyze all of the data because there is usually more data than the information granules that are required. Rather, it eliminates samples that don't provide fresh insights.

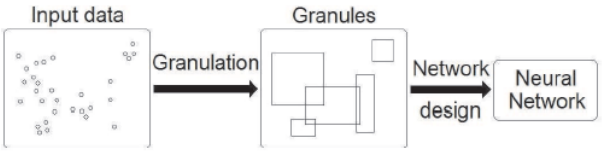

Fig. 2: Granular neural networks are designed in two steps

In essence, GNN manages data streams using a quick, incremental, one-pass learning process. Without being aware of the statistical characteristics of the data or classes beforehand, it might start the learning process. Fuzzy hyperboxes, which create decision boundaries between various classes, are used to granulate the feature space. The main features of GNN are listed below:

- Simultaneously handles samples with and without labels
- Adjusts its settings and structure to take into account new learnings, rejecting the old ones;
- Demonstrates the ability to perform nonlinear separation

- Supports the growth of lifetime learning through both destructive top-down and constructive bottom-up approaches
- Can deal with data uncertainty and drift detection techniques

GNN Structure and Processing

The Graph Neural Network (GNN) acquires knowledge from a continuous data stream denoted as x[h], where h represents the time index ranging from 1 to infinity. The presence or absence of a class label C[h] may or may not be seen in training instances. Each information granule, denoted as γ_i, belongs to a finite collection of granules γ = {γ1,...,γc} in the feature space X ⊆ Rn. These granules are connected with a specific class, Ck, which belongs to a finite collection of classes C = {C1,..., Cm} in the output space Y ⊆ N. Granules from the input stream and T-S neurons connect the feature and output spaces in the GNN.

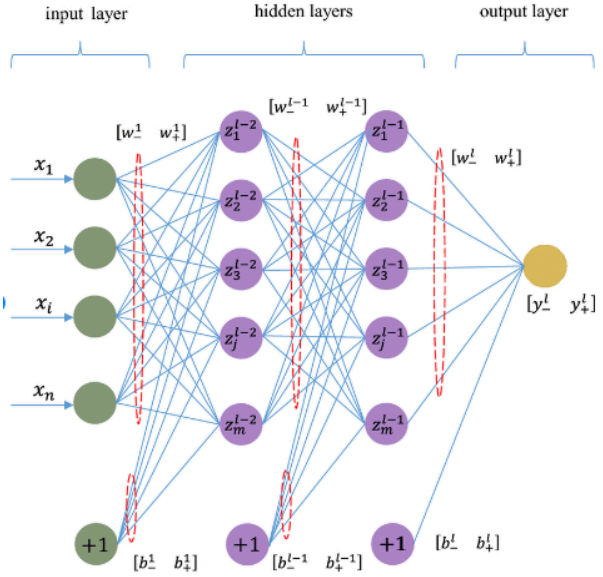

Fig. 3: Granular neural network categorization structure

Figure (3) shows the 5-layer neural network. The input layer feeds feature vectors $x[h] = (x1, ..., xj, ..., xn)[h], h = 1, ...$ into the network, whereas the granular layer contains information granules $\gamma i \forall$throughout the feature space. Null neurons $TSni \forall i$ are included in the aggregation layer for partly overlapping granules. After aggregating membership values, class compatibility values are compared by the decision layer o_i and output the class $\overline{C}_k$ associated with the granule γ_i whose compatibility rating is the greatest. Class label indications are shown in the output layer. $x[h], h = 1, ...,$ is input for all levels except the input layer.

The GNN classifier's parameters and structure can be altered to suit the needs of the particular application. The number of classes can be automatically handled once it is known. A maximum number of granules can be established inside the model structure if processing time and memory are constrained. The learning algorithm can automatically change the number of granules and classes in new situations.

The number of modules (or subgranules), the learning method, the goal error, the proportion of training data, the number of hidden layers, and the number of neurons in each layer are among the factors that are optimized in granular neural networks. Today, there are many different optimization techniques accessible, but choosing the right one for a certain application is essential to getting the best results. This study adds an adaptive mutation factor to improve GNN parameter optimization.

Adaptive Mutation Factor (AMF)

This study proposes an improved approach to fix these problems. This method uses an adaptive mutation relationship type that factors neighbouring data into its data. Two things affect this adaptive mutation factor: The data intensities, denoted as $\lambda\,(0 < \lambda < 1)$, and the spatial location of the neighbours, denoted as $\xi\,(0 < \xi < 1)$, are important factors to consider, it is also the neighbourhood's architecture also influences it. Taking into account the following definition of adaptive mutation:

$$sd^2\,(x_j, v_i) = \|x_j - v_i\|^2\,(1 - \lambda H_{ij} - \xi F_{ij}) \tag{18}$$

Where H_{ij} stands for the primary attraction's feature and F_{ij} for the secondary attraction's distance. The parameters modify the two nearby attractions' potency λ and ξ. To sum up, the best outcomes are provided by the E-GNN model.

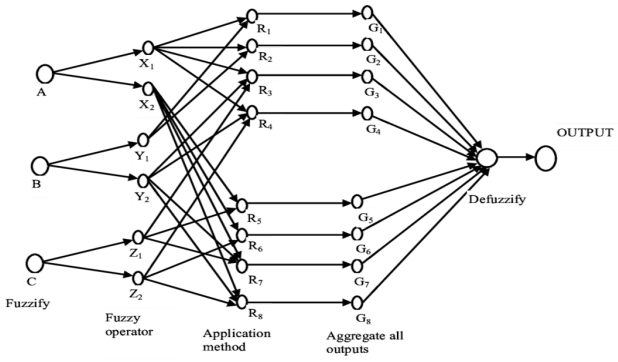

Fig. 4: Structure of ANFIS

Adaptive Neural Fuzzy Inference System (ANFIS) Algorithm

To get superior classification results, the Hybrid Optimized Adaptive Neuro Fuzzy Inference System (HOANFIS) is presented. The fuzzy inference system and the neural network are combined. The neural network lends the system a feeling of flexibility, while

the fuzzy logic accounts for the system's imprecision and uncertainty. This hybrid approach uses the rules to construct an initial fuzzy model and its input variables. The final HOANFIS model of the system is then created by using the neural network to refine the rules of the original fuzzy model. For the supplied database, it is utilized to achieve flexibility, rapid convergence, and high accuracy. The construction of ANFIS is shown in Figure (4).

Two sections make up ANFIS' internal organization. Using network-like regulations, these two parts are linked (Haznedar *et al.*, 2021). ANFIS uses a neural network to improve the fuzzy rules it initially discovered using the supplied input data set. It is possible to define a typical Takagi-Sugeno rule set as follows:

$$if\ x\ is\ A_1\ and\ y\ is\ B_1 then\ f_1 = \alpha_1 x + \beta_1 y + r_1 \quad (19)$$

$$if\ x\ is\ A_2\ and\ y\ is\ B_2 then\ f_2 = \alpha_2 x + \beta_2 y + r_2 \quad (20)$$

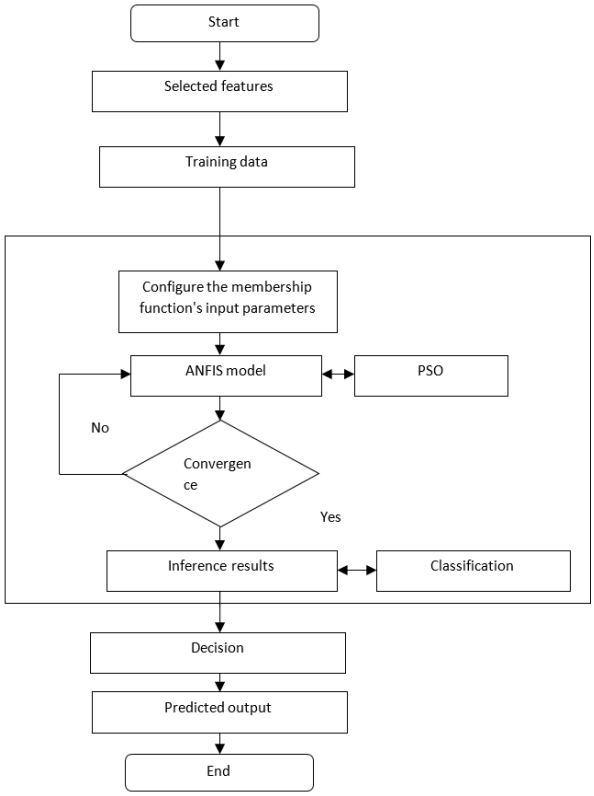

Fig. 5: HOANFIS algorithm

The linear output parameters are α, β, and r. See Figure (4) for the ANFIS structure. ANFIS learns optimum rules via hybrid learning. The learning process involves repeated updates of antecedent and consequent parameters, with one set held constant while the other is updated. The Least-Square Error (LSE) approaches are used by the Adaptive Neuro-Fuzzy Inference System (ANFIS) to optimize the Consequent Parameters (α,,r), which subsequently updates the Antecedent Parameters (a, b, c) in the backward pass. The computation of gradients at each iteration may pose significant challenges and carry the inherent possibility of encountering local and global minima. The Particle Swarm Optimization (PSO) method and ANFIS algorithm are combined to prevent sluggish convergence. It may provide an alternative remedy for this kind of issue. HOANFIS' proposed flowchart is seen in Fig. (5).

The PSO approach draws its inspiration from how birds search for food. According to their knowledge and that of others, particles in this model alter their placements and travel directions. The optimization method is based on particle competition and cooperation (Jiang *et al.*, 2007). One may track the states of the particles by their paths and velocities while using PSO to solve optimization issues.

A fitness function projects each particle to identify the best solution to the issue. To obtain the finest position globally and locally, the particle Pi screens its individual best, i.e., personal best named as $Pbest$ and global best named as $Gbest$ to inform its position and velocity. Each reiteration alters its position Xid, velocity Vid, and the dth dimension using the associated conditions:

$$V_{id}(t) = w * V_{id}(t-1) + c_1 r_1 * \left(X_{pbest_{id}} - X_{id}(t-1)\right) + c_2 r_2 * \left(X_{gbest_{id}} - X_{id}(t-1)\right) \quad (21)$$

$$X_{id}(t) = X_{id}(t-1) + V_{id}(t) \quad (22)$$

where, w = inertial weight

V_{id} is velocity in d dimension

c1 and c2 = constants named as acceleration factor

r1 and r2 = two dissimilar distributed arbitrary numbers in the [0,1] range regularly.

The modified technique is repeatedly repeated until one of two outcomes is reached: An acceptable $Gbest$ or a predetermined number of iterations, $tmax$. The ANFIS algorithm is used to improve classification accuracy and to produce optimum values. A greater weight is assigned to a training feature that the previous classifier has not taught. By lowering the necessary number of learning epochs, the suggested HOANFIS is intended to lower the computational cost of the traditional ANFIS when dealing with huge volumes of training data. Accuracy and training time metrics are considerably enhanced by changing the feature's weights in the training set.

The novelty of the HOANFIS algorithm is improving the convergence speed for the larger breast cancer dataset. Optimizing internal parameters using the PSO algorithm improves prediction accuracy, execution time, and AUC metrics for given data.

The above analysis shows that the proposed hybrid optimization-based ANFIS technique is highly efficient for analyzing breast cancer data.

Weight-Based Support Vector Machine (WSVM) Algorithm

In this investigation, WSVM classifies. It finds a linear separating hyperplane with the highest margin of separation between data in a higher-dimensional environment for categorizing data using machine learning. The normal SVM algorithm has a long training time. To overcome the issue, a weight-based SVM is introduced in this study.

WSVM delivers near-optimal class separation. Linear algebra and geometry may be used with nonlinearly separated data in WSVM's high-dimensional feature space. It increases the distance of either class from the hyperplane while using the hyperplane to segregate the greatest percentage of training data on the same class (Xie *et al.*, 2017). WSVM uses a variety of kernel functions to enable inner products to be created directly in feature space. This study's extracted features train the WSVM to classify the given dataset. The WSVM generally separates the high-dimensional space by building a hyperplane. A model of the training data is produced by the WSVM classifier using the labelled vector of displacements of each VM cost and time. This model is then used to classify displacements of unknown data dynamically. The training dataset model is used throughout the testing procedure. In training sets containing more interesting data, WSVM is a maximum margin hyperplane classifier with great classification performance. Figure (6) shows the structure of the SVM algorithm.

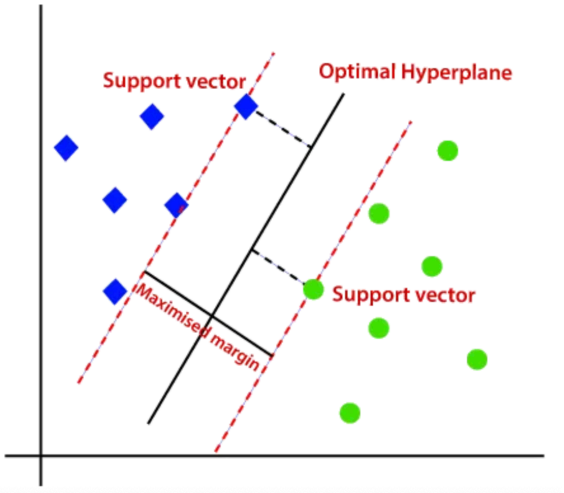

Fig. 6: Structure of the SVM algorithm

The basic idea behind the WSVM is to assign distinct weights to individual data points, reflecting their relative importance within the class. This approach enables different data points to contribute to the learning process of the decision surface to varying degrees. The training data set changes if the weights are specified.

$$\{(x_i, y_i, W_i)\}_{i=1}^{l} x_i \epsilon R^N, y_i \epsilon \{-1, 1), W_i \epsilon R \quad (23)$$

Where the scalar $0 \leq Wi \leq 1$ is a data point given a weight x_i.

To improve generalization, the WSVM increases the margin of separation while decreasing classification error, accuracy, and time. Building a cost function starts this procedure. WSVM weights the penalty term to reduce the effect of less important data points even if C is constant and all training data points are treated equally. The constrained optimization problem is expressed as follows:

$$Minimize\phi(w) = \frac{1}{2}w^T w + ATE\sum_{i=1}^{l} W_i \xi_i \quad (24)$$

where A is accuracy, T is time and E is error rate

Subject to:

$$\begin{aligned} &y_i(\langle w, \phi(x_i)\rangle + b) \geq 1 - \xi_i \quad i = 1, \ldots, l \\ &\xi_i \geq 0 \qquad\qquad\qquad\qquad\quad i = 1, \ldots, l \end{aligned} \quad (25)$$

It assigns the weight W_i to the data point x_i formulated above. Thus, the dual formulation is subject to:

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} (a_i a_j t_i t_j e_i e_j) \quad (26)$$

SVM upper limits for α_i are constants, whereas WSVM weight values $cteWi$ serve as dynamical boundaries. It is focused on understanding the pattern of breast cancer infection using weight values of SVM and also identifies the earlier stages of breast cancer disease, which improves the classification effectively.

Results and Discussion

To evaluate the effectiveness of the proposed method in distinguishing different categories of breast cancer images, two primary performance measures were employed. Throughout all experiments, the model parameters were optimized and hyperparameters were fine-tuned using the validation set and training procedures. The datasets used for experimentation were sourced from the Breast Cancer dataset available on Kaggle, specifically the CBIS-DDSM (Curated Breast Imaging Subset of the Digital Database for Screening Mammography), an enhanced version of the DDSM. This dataset includes 2,620 digitized mammographic images annotated with verified pathology information for normal, benign, and malignant cases. The curated dataset also features updated Region of Interest (ROI) segmentations, bounding boxes, and diagnostic labels, with all images converted to DICOM format to ensure consistency and standardization. A certified mammographer curated this subset from the full DDSM collection. True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) were the basic values used to calculate a number of binary classification metrics for performance evaluation. The percentage of accurately anticipated positive cases to all predicted positives was referred to as precision. The

percentage of true positives that the model properly recognized is known as recall (or sensitivity). The F1-score, which is the harmonic mean of precision and recall, was employed to provide a fair evaluation of performance because these two measures frequently trade off. As a measure of overall performance, accuracy was determined by dividing the number of properly predicted cases by the total number of predictions.

The performance estimate was made more robust and less biased by using k-fold cross-validation. Each iteration employed k−1 folds for training and the remaining fold for testing after the dataset was divided into k folds of equal size. Using a different fold as the test set, this procedure was carried out k times. A more accurate and broadly applicable assessment of the model's predictive power was provided by the final performance measures, which were derived by averaging the outcomes across all k iterations. The suggested ECSO-EMLC method is evaluated against GONN, BI-RADS, FS-ResNet CNN, and EMLC classification for breast cancer classification in terms of recall, precision, accuracy, f-measure, AUC, ROC, and execution time.

Calculating precision:

$$Precision = \frac{True\ positive}{True\ positive + False\ positive} \quad (27)$$

While recall denotes the completeness of the retrieved pertinent instances, precision relates to the caliber of the categorization findings. A high precision value in classification tasks indicates that the algorithm properly detects a higher percentage of pertinent cases in comparison to the total number of examples it projected to be affirmative. Precision, which measures the percentage of correctly predicted positive instances among all of the model's positive predictions, is specifically defined as the ratio of true positives (TP) to the total of true positives plus false positives (TP + FP).

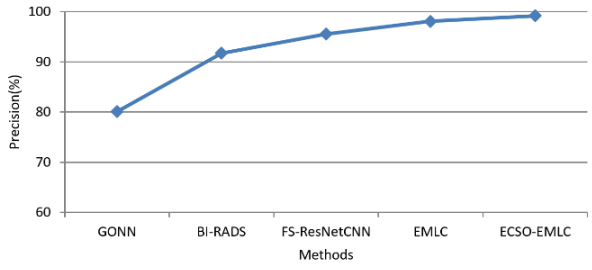

Fig. 7: Precision

As illustrated in Figure (7) the comparison of classification accuracy across various methodologies is presented, where the x-axis represents the different algorithms and the y-axis indicates the corresponding accuracy values. The proposed ECSO-EMLC algorithm demonstrates superior performance on the breast cancer dataset compared to existing techniques, including GONN, BI-RADS, FS-ResNetCNN, and the standard EMLC approach, all of which exhibit lower precision levels. The enhanced accuracy of ECSO-EMLC is attributed to its effective feature selection mechanism, which optimally identifies the most relevant attributes for classification. This improvement underscores the robustness of the proposed method in accurately classifying breast cancer cases.

Recall

The recall value is determined in the manner described below:

$$Recall = \frac{True\ positive}{True\ positive + False\ negative} \quad (28)$$

The following representation of the comparison graph:

Recall and precision are calculated as the total number of relevant documents divided by the number of relevant documents returned by a search.

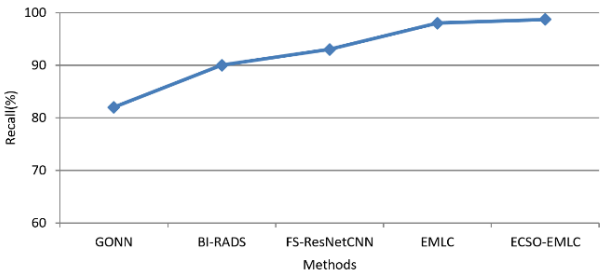

Fig. 8: Recall

In Figure (8) both existing and recommended methodologies are used to evaluate the comparison measure recall. Approaches are on the x-axis and recall is on the y-axis. Our ECSO-EMLC algorithm provides greater recall for the breast cancer dataset than current approaches such as GONN, BI-RADS, FS-ResNet CNN, and EMLC algorithms. The results conclude that the proposed ECSO-EMLC improves breast cancer classification accuracy by carefully selecting characteristics for the best subset.

F-Measure

$F1$-score is defined as:

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall} \quad (29)$$

Figure (9) illustrates a comparative analysis of F-measure values obtained from both existing and the proposed methodologies. In the plot, the y-axis represents the F-measure scores, while the x-axis denotes the different algorithms evaluated. The proposed ECSO-EMLC algorithm achieves a significantly higher F-measure on the breast cancer dataset compared to existing methods, including GONN, BI-RADS, FS-ResNetCNN, and the baseline EMLC, all of which demonstrate relatively lower F-measure performance. These results confirm that the ECSO-EMLC approach enhances classification performance by effectively

selecting the most discriminative features, thereby improving the overall predictive capability of the model.

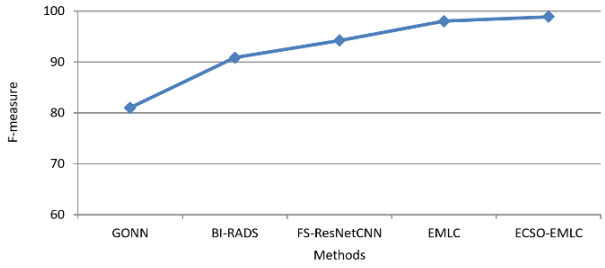

Fig. 9: F-measure

Accuracy

Accuracy, a key performance metric, is computed by dividing the sum of correctly classified instances—true positives (T_p) and true negatives (T_n)—by the total number of evaluated instances, which includes true positives, true negatives, false positives (F_p), and false negatives (F_n). Mathematically, accuracy = (T_p + T_n) / (T_p + T_n + F_p + F_n) and it quantifies the model's overall classification effectiveness. This is how the accuracy is calculated:

$$Accuracy = \frac{T_p+T_n}{(T_p+T_n+F_p+F_n)} \quad (30)$$

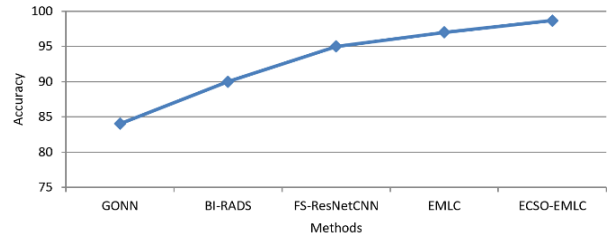

Fig. 10: Accuracy

As shown in Figure (10) both existing and recommended techniques evaluate the comparison metric's accuracy. The y-axis shows accuracy, while the x-axis shows methods. GONN, BI-RADS, and FS-Res Net are a few of the current techniques. CNN and EMLC algorithms perform less accurately for the provided breast cancer dataset, but the suggested ECSO-EMLC approach does. The suggested approach focuses on choosing the most relevant data using IPCA, which successfully increases the accuracy of features across the breast cancer dataset. Consequently, according to the results, the suggested ECSO-EMLC boosts the accuracy of breast cancer classification by choosing the right characteristics.

AUC

A binary classification model's overall performance is measured by the Receiver Operating Characteristic (ROC) curve's Area Under the Curve (AUC). The True Positive Rate (TPR) versus False Positive Rate (FPR) trade-off across different classification thresholds is depicted graphically by the ROC curve. Better model discrimination between the positive and negative classes is indicated by a higher AUC score.

As shown in Figure (11) AUC compares the present and recommended approaches. The y-axis shows AUC, while the x-axis shows methods. Compared to existing methods like GONN, BI-RADS, FS-ResNet CNN, and EMLC, the proposed ECSO-EMLC algorithm provides a higher AUC for the presented datasets. Pre-processing is performed to improve the classification accuracy. The findings conclude that using optimized features, the ECSO-EMLC technique effectively handles dataset performance.

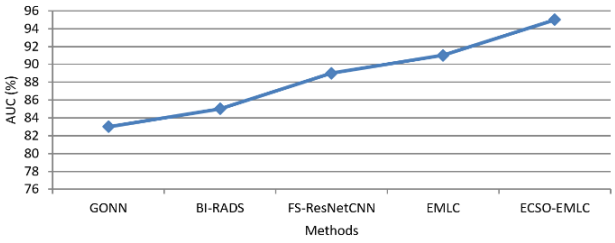

Fig. 11: AUC

ROC

A graphical tool for assessing a classification model's diagnostic performance across different threshold settings is the Receiver Operating Characteristic (ROC) curve. Plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) gives information on how well the model can differentiate between classes at various decision boundaries.

Fig. 12: ROC

In Figure (12) the existing and recommended techniques were used to evaluate the comparison measure using ROC. For the x-axis, false positive rates are determined, while the y-axis shows real positive rates. The findings show that the suggested ECSO-EMLC method improves dataset performance via optimized features.

Time Complexity

The algorithm is better when the proposed method provides lower time complexity.

As shown in Figure (13) current and recommended approaches evaluate the comparison measure's execution

time. The x-axis shows methods and the y-axis shows execution time. The suggested ECSO-EMLC technique has a reduced execution time for the provided breast cancer dataset compared to the current methods, such as GONN, BI-RADS, FS-ResNetCNN, and EMLC. Consequently, the analysis's findings indicate that the suggested ECSO-EMLC could increase classification accuracy by carefully choosing characteristics.

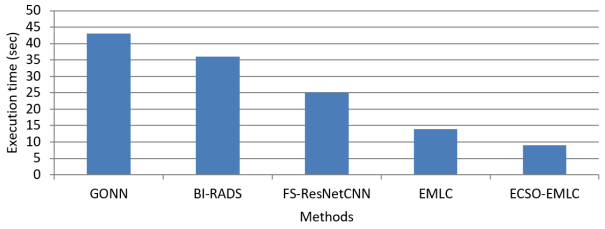

Fig. 13: Execution time

Conclusion

This study proposes the Enhanced Cuckoo Search Optimization with Ensemble Multi-Level Classification (ECSO-EMLC) approach to enhance breast cancer classification accuracy. Initially, statistical correlation-based preprocessing is applied to improve classifier efficiency. Features are extracted from mammographic Regions of Interest (ROIs) using Incremental Principal Component Analysis (IPCA). However, IPCA alone leads to high computational complexity due to the large number of extracted wavelet coefficients. To address this, the ECSO algorithm, inspired by swarm intelligence, is employed for optimal feature selection and to reduce computational overhead. In this method, the update coefficient for each cuckoo is determined by its distance from the global best solution—cuckoos farther away receive higher update coefficients to promote faster convergence, while those closer conduct finer local searches to enhance solution accuracy.

To further boost predictive performance, an ensemble learning strategy is implemented, integrating three classifiers: Enhanced Granular Neural Network (E-GNN), Adaptive Neuro-Fuzzy Inference System (ANFIS), and Weighted Support Vector Machine (WSVM). Experimental results indicate that the inclusion of discriminative features such as nodule shape, intensity, population size, margin, dimension, granularity, and texture significantly improves classification performance in terms of both accuracy and model training effectiveness. However, this study has limitations such as the breast cancer classification model may not work as well on other datasets with different features, even if it was successful on the chosen dataset. The model's prediction precision and applicability are susceptible to variations in the quality of the input data, the distribution of features, and any imbalances between classes from various sources. Further, the study's main focus is on a controlled experimental environment, which could not adequately represent the intricacies and unpredictability of real-life applications. For example, model performance might be impacted by noise or missing values in clinical data. To overcome these obstacles, the model's resilience and flexibility would have to be evaluated with more testing on other datasets, including real-world clinical data.

Acknowledgment

The authors would like to thank the open-source data platforms and repositories such as Kaggle for providing accessible breast cancer datasets. The authors also acknowledge their respective institutions for their continuous support during the course of this research.

Funding Information

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author's Contributions

Satyabrata Patro: Conceptualization, methodology design, algorithm implementation, and manuscript writing.

Jyotirmaya Mishra: Supervision, critical review of methods and results, and manuscript editing.

Bhavani Sankar Panda: Experimental analysis, result validation, and final proofreading of the manuscript.

All authors read and approved the final manuscript.

Ethics

This study used publicly available, anonymized datasets and did not involve any human or animal participants directly. Hence, ethical approval was not required. All procedures followed are in accordance with institutional guidelines and data usage policies.

References

Alia, A., & Taweel, A. (2021). Enhanced Binary Cuckoo Search With Frequent Values and Rough Set Theory for Feature Selection. *IEEE Access*, *9*, 119430-119453. https://doi.org/10.1109/access.2021.3107901

Aljuaid, H., Alturki, N., Alsubaie, N., Cavallaro, L., & Liotta, A. (2022). Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning. *Computer Methods and Programs in Biomedicine*, *223*, 106951. https://doi.org/10.1016/j.cmpb.2022.106951

Asri, H., Mousannif, H., Moatassime, H. A., & Noel, T. (2016). Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. *Procedia Computer Science*, *83*, 1064-1069. https://doi.org/10.1016/j.procs.2016.04.224

Baskaran, V., Guergachi, A., Bali, R. K., & Naguib, R. N. G. (2011). Predicting Breast Screening Attendance Using Machine Learning Techniques. *IEEE Transactions on Information Technology in Biomedicine*, *15*(2), 251-259. https://doi.org/10.1109/titb.2010.2103954

Dietterich, T. G. (2000). *Ensemble Methods in Machine Learning*. *1857*. https://doi.org/https://doi.org/10.1007/3-540-45014-9_1

Elkorany, A. S., Marey, M., Almustafa, K. M., & Elsharkawy, Z. F. (2022). Breast Cancer Diagnosis Using Support Vector Machines Optimized by Whale Optimization and Dragonfly Algorithms. *IEEE Access*, *10*, 69688-69699. https://doi.org/10.1109/access.2022.3186021

Ganesan, K., Acharya, U. R., Chua, C. K., Lim, C. M., & Abraham, K. T. (2014). One-Class Classification of Mammograms Using Trace Transform Functionals. *IEEE Transactions on Instrumentation and Measurement*, *63*(2), 304-311. https://doi.org/10.1109/tim.2013.2278562

Gao, S., Gao, Y., Zhang, Y., & Xu, L. (2019). Multi-Strategy Adaptive Cuckoo Search Algorithm. *IEEE Access*, *7*, 137642-137655. https://doi.org/10.1109/access.2019.2916568

Hamed, G., Marey, M., Amin, S. E., & Tolba, M. F. (2021). Automated Breast Cancer Detection and Classification in Full Field Digital Mammograms Using Two Full and Cropped Detection Paths Approach. *IEEE Access*, *9*, 116898-116913. https://doi.org/10.1109/access.2021.3105924

Haznedar, B., Arslan, M. T., & Kalinli, A. (2021). Optimizing ANFIS using simulated annealing algorithm for classification of microarray gene expression cancer data. *Medical & Biological Engineering & Computing*, *59*(3), 497-509. https://doi.org/10.1007/s11517-021-02331-z

Huang, Z., & Chen, D. (2022). A Breast Cancer Diagnosis Method Based on VIM Feature Selection and Hierarchical Clustering Random Forest Algorithm. *IEEE Access*, *10*, 3284-3293. https://doi.org/10.1109/access.2021.3139595

Jiang, Y., Hu, T., Huang, C., & Wu, X. (2007). An improved particle swarm optimization algorithm. *Applied Mathematics and Computation*, *193*(1), 231-239. https://doi.org/10.1016/j.amc.2007.03.047

Kaggle. (2021). CBIS-DDSM: Breast Cancer Image Dataset. *Curated Breast Imaging Subset DDSM Dataset (Mammography)*. https://www.kaggle.com/datasets/awsaf49/cbis-ddsm-breast-cancer-image-dataset

Lai, Z., Xu, Y., Chen, Q., Yang, J., & Zhang, D. (2014). Multilinear Sparse Principal Component Analysis. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(10), 1942-1950. https://doi.org/10.1109/tnnls.2013.2297381

Li, G., Li, C., Wu, G., Xu, G., Zhou, Y., & Zhang, H. (2022). MF-OMKT: Model fusion based on online mutual knowledge transfer for breast cancer histopathological image classification. *Artificial Intelligence in Medicine*, *134*, 102433. https://doi.org/10.1016/j.artmed.2022.102433

Liu, X., & Tang, J. (2014). Mass Classification in Mammograms Using Selected Geometry and Texture Features, and a New SVM-Based Feature Selection Method. *IEEE Systems Journal*, *8*(3), 910-920. https://doi.org/10.1109/jsyst.2013.2286539

Loizidou, K., Skouroumouni, G., Nikolaou, C., & Pitris, C. (2022). Automatic Breast Mass Segmentation and Classification Using Subtraction of Temporally Sequential Digital Mammograms. *IEEE Journal of Translational Engineering in Health and Medicine*, *10*, 1-11. https://doi.org/10.1109/jtehm.2022.3219891

Ma, P., Ren, J., Zhao, H., Sun, G., Murray, P., & Zheng, J. (2021). Multiscale 2-D Singular Spectrum Analysis and Principal Component Analysis for Spatial-Spectral Noise-Robust Feature Extraction and Classification of Hyperspectral Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 1233-1245. https://doi.org/10.1109/jstars.2020.3040699

Malebary, S. J., & Hashmi, A. (2021). Automated Breast Mass Classification System Using Deep Learning and Ensemble Learning in Digital Mammogram. *IEEE Access*, *9*, 55312-55328. https://doi.org/10.1109/access.2021.3071297

Mohsin, M., Li, H., & Abdalla, H. B. (2020). Optimization Driven Adam-Cuckoo Search-Based Deep Belief Network Classifier for Data Classification. *IEEE Access*, *8*, 105542-105560. https://doi.org/10.1109/access.2020.2999865

Naseem, U., Rashid, J., Ali, L., Kim, J., Haq, Q. E. U., Awan, M. J., & Imran, M. (2022). An Automatic Detection of Breast Cancer Diagnosis and Prognosis Based on Machine Learning Using Ensemble of Classifiers. *IEEE Access*, *10*, 78242-78252. https://doi.org/10.1109/access.2022.3174599

Nasir Khan, H., Shahid, A. R., Raza, B., Dar, A. H., & Alquhayz, H. (2019). Multi-View Feature Fusion Based Four Views Model for Mammogram Classification Using Convolutional Neural Network. *IEEE Access*, *7*, 165724-165733. https://doi.org/10.1109/access.2019.2953318

Padmanabhan, S., & Sundararajan, R. (2012). Enhanced accuracy of breast cancer detection in digital mammograms using wavelet analysis. *2012 International Conference on Machine Vision and Image Processing (MVIP)*, 153-156. https://doi.org/10.1109/mvip.2012.6428783

Sameti, M., Ward, R. K., Morgan-Parkes, J., & Palcic, B. (2009). Image Feature Extraction in the Last Screening Mammograms Prior to Detection of Breast Cancer. *IEEE Journal of Selected Topics in Signal Processing*, *3*(1), 46-52. https://doi.org/10.1109/jstsp.2008.2011163

Singh, L. K., Khanna, M., & Singh, R. (2023). Artificial intelligence based medical decision support system for early and accurate breast cancer prediction. *Advances in Engineering Software*, *175*, 103338. https://doi.org/10.1016/j.advengsoft.2022.103338

Song, M., & Pedrycz, W. (2013). Granular Neural Networks: Concepts and Development Schemes. *IEEE Transactions on Neural Networks and Learning Systems*, *24*(4), 542-553. https://doi.org/10.1109/tnnls.2013.2237787

Subasree, S., Sakthivel, N. K., Tripathi, K., Agarwal, D., & Tyagi, A. K. (2022). Combining the advantages of radiomic features based feature extraction and hyper parameters tuned RERNN using LOA for breast cancer classification. *Biomedical Signal Processing and Control*, *72*, 103354. https://doi.org/10.1016/j.bspc.2021.103354

Wu, J.-X., Liu, H.-C., Chen, P.-Y., Lin, C.-H., Chou, Y.-H., & Shung, K. K. (2020). Enhancement of ARFI-VTI Elastography Images in Order to Preliminary Rapid Screening of Benign and Malignant Breast Tumors Using Multilayer Fractional-Order Machine Vision Classifier. *IEEE Access*, *8*, 164222-164237. https://doi.org/10.1109/access.2020.3022388

Wuniri, Q., Huangfu, W., Liu, Y., Lin, X., Liu, L., & Yu, Z. (2019). A Generic-Driven Wrapper Embedded With Feature-Type-Aware Hybrid Bayesian Classifier for Breast Cancer Classification. *IEEE Access*, *7*, 119931-119942. https://doi.org/10.1109/access.2019.2932505

Xie, L., Li, G., Xiao, M., Peng, L., & Chen, Q. (2017). Hyperspectral Image Classification Using Discrete Space Model and Support Vector Machines. *IEEE Geoscience and Remote Sensing Letters*, *14*(3), 374-378. https://doi.org/10.1109/lgrs.2016.2643686

Xiong, X., Kim, Y., Baek, Y., Rhee, D. W., & Kim, S.-H. (2005). Analysis of Breast Cancer Using Data Mining and Statistical Techniques. *Sixth International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing and First ACIS International Workshop on Self-Assembling Wireless Networks (SNPD/SAWN'05)*, 82-87. https://doi.org/10.1109/snpd-sawn.2005.19

Zhou, Z., Li, S., Qin, G., Folkert, M., Jiang, S., & Wang, J. (2020). Multi-Objective-Based Radiomic Feature Selection for Lesion Malignancy Classification. *IEEE Journal of Biomedical and Health Informatics*, *24*(1), 194-204. https://doi.org/10.1109/jbhi.2019.2902298