Journal of Computer Science

Research Article

Cuckoo Search Optimized Random Forest for Breast Cancer Prognosis

Prasad S. Sase[1], Debabrata Swain[2] and Shailesh Kumar[1]

[1]*Department of Computer Science and Engineering, Shri JJT University, Jhunjhunu, India*
[2]*Department of Computer Science and Engineering, Pandit Deendayal Energy University, Gandhinagar, India*

Article history
Received: 27-03-2025
Revised: 28-07-2025
Accepted: 08-08-2025

Corresponding Author:
Debabrata Swain
Department of Computer Science and Engineering, Pandit Deendayal Energy University, Gandhinagar, India
Email: debabrata.swain7@yahoo.com

Abstract: Breast cancer remains a major global health concern due to its high mortality rate, particularly when diagnosis occurs at an advanced stage. Accurate and early differentiation between benign and malignant tumors is therefore critical for improving patient outcomes. Conventional diagnostic practices largely depend on manual assessment and clinical expertise, which may lead to subjective variability in decision-making. To overcome this limitation, this study presents an automated machine learning–based screening framework for breast cancer prognosis. The proposed approach employs a Random Forest classifier for tumor classification, with feature space transformation performed using Principal Component Analysis to reduce dimensionality and enhance discriminative capability. To further improve predictive performance, the hyperparameters of the classifier are optimized using the Cuckoo Search algorithm. The model is trained and assessed using the benchmark breast cancer dataset from the UCI Repository. Experimental results demonstrate that the optimized framework achieves an accuracy of 98% on the test dataset, indicating strong classification capability. The proposed method offers a reliable and efficient computational tool that can assist clinicians in early-stage breast cancer diagnosis.

Keywords: Random Forest, Hyperparameter Tuning, Cuckoo Search, Principal Component Analysis, Standard Scaling

Introduction

Cancer is considered as one of the deadliest diseases found in the society. It exists in many forms, affecting different organs of the body. Every year, many women lose their lives due to breast cancer. According to a 2022 WHO report, about 670,000 women died from the disease (World Health Organization, 2021). Breast cancer often shows no symptoms in its early stages, so detecting it late makes it much harder to save the patient's life. Most breast cancer cases develop in the milk production tissues such as ducts and lobules of the breast. Approximately 85% originate in the ducts, while around 15% begin in the lobules (Feng *et al*., 2018). Ducts consist of group of tubes, responsible for circulating the milk. Lobules are composed of small sized sacs used for producing milk (Afaq and Singh, 2024). In earlier days lungs cancer was found to be as one of common type cancer found in the society. But in present days the cases of breast cancer have exceeded the count of lungs cancer. The disease mainly spreads with in a female in different stages starting from 0 to 4. In stage-0 abnormal cells formation starts in the ducts. Whereas during stage- 1, 2, and 3 the tumor gets created and grows in to larger size (Kathale and Thorat, 2020). In the last stage the cancerous cell affects the nearby organs like bones, liver and lungs. For the detection of this disease generally physicians suggest different imaging tests such as X-ray, MRI, and Ultrasound. Using different imaging techniques, physicians generally try to find the presence of any abnormal grown muscle. For this task Doctors use their experience to locate the presence of the abnormal tissue. However, the clinical decision made by the physician based on their prior knowledge applied on the medical image input is always not accurate. The diagnosis of malignant cells becomes difficult for those women who have undergone any surgery in recent past or who have dense breast muscle. Other than medical imaging physicians many times suggest other pathological tests to detect the disease. These all complexities make the

diagnosis process more expensive and time consuming. Now a days different Artificial Intelligence based methods like Machine Learning (ML) helps to solve these kinds of problems to large extent using complex pattern matching. Different ML algorithms generally help to visualize the data and discover any hidden factors to segregate the different instance with more accuracy. ML can perform these tasks after analyzing and capturing knowledge from the input data. By taking motivation from this here Ensemble based boot strap aggregation algorithm such as Random Forest (RF) is used to identify the presence of Breast Cancer disease with more accuracy using a patient health data. RF is formed by the collection of different individual classifier such as decision trees. Hence by having more than one classifier the task of detection becomes more accurate. In this work to avoid the model over fitting issue, different precautionary steps like feature elimination using Principal component analysis and hyperparameter (H-param) tuning using Cuckoo search method is performed.

Literature Survey

Delen *et al*. (2005) used a data mining method for assessing the risk of breast cancer. Two different approaches such as Artificial Neural network and Decision Tree (DT) were used for predicting the risk of the disease. The dataset consists of 200,000 records. 17-features were used for the model training and validation. The highest accuracy of 93% was report by the DT model.

Tripathy *et al*. (2014) used 4 different AI models for the prior screening of Breast cancer disease. There were a total of 11 features used for the classification. During the data pre-processing data scaling was performed. Out of all different models SVM has shown the best accuracy of 95%.

Chaurasia *et al*. (2018) has used 3 different algorithms such as Naïve bayes, RBF network and tree-based classifier for the classification of breast cancer. The data imbalance issue is not handled. The Naïve bayes algorithm has given the highest accuracy of 97%.

Esmaeili *et al*. (2020) used different ML based clinical decision-making system for the better identification using mammography reports. Out of all different models used the highest accuracy of 84% was shown by K-NN model.

Khatun *et al*. (2021) used Multilayer perceptron for the accurate prediction of breast cancer disease. There were 10 features and 116 records available in the dataset. During pre-processing the alphabet values were converted in to numeric. The highest accuracy obtained by the model is 85%.

Pal *et al*. (2023) developed a ML tool for correct prediction of breast cancer. There were 11 different features used in the dataset representing different characteristics of the breast. Out of all the models used in the tool, K nearest neighbor has shown the better performance. The accuracy reported by the system was 95%.

Ara *et al*. (2021) applied different ML algorithms for the detection of malignant and benign cases. Out of all applied algorithms Support vector machine and RF have shown the highest accuracy of 96.5%.

Yarabarla *et al*. (2019) developed a breast cancer detection system using Gradient Boosting (GB) algorithm. The dataset balancing is not performed in this work. The total number of features used in this are 10. The highest accuracy shown by the model is 71%.

Kaur and Gupta (2024) developed a breast cancer classification system using RF and DT algorithm. There are30-features used from the dataset. The RF algorithm has shown the highest accuracy of 93%.

After investigating the above discussed literatures, different research gaps are identified. In several studies data balancing was absent, that mostly leads towards a biased classifier. Additionally, the absence of systematic feature reduction leads to models being trained on redundant or irrelevant attributes, increasing computational complexity and potentially degrading predictive performance. Furthermore, hyperparameter optimization is frequently neglected, and the use of default parameter settings can significantly contribute to model overfitting and poor generalization.

Proposed System

The detailed phases of the proposed system are given in Fig. 1.

Dataset Collection

The dataset is collected from the Breast Cancer Wisconsin available in the UCI repository. The dataset contains 30 features representing different characteristics of breast muscles (UCI Machine Learning Repository, 2018). All the features are having real values. There is total 569 records available in the dataset. The target column is having 2 values such as benign and malignant. The dataset is having 357 benign and 212 malignant records. The feature set details of the dataset are shown in the following Table 1.

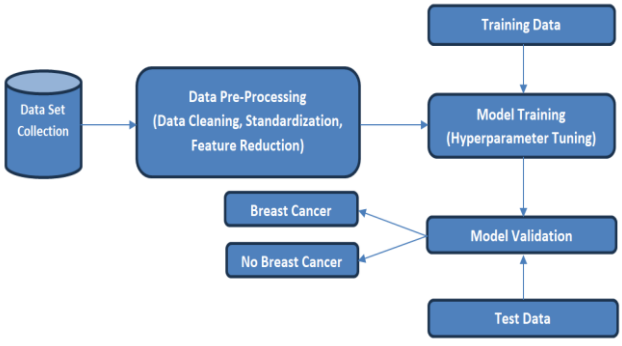

Fig. 1: Proposed System Architecture

Table 1: Feature set

Feature Name	Feature Type
mean radius	Continuous
mean texture	Continuous
mean perimeter	Continuous
mean area	Continuous
mean smoothness	Continuous
mean compactness	Continuous
mean concavity	Continuous
mean concave points	Continuous
mean symmetry	Continuous
mean fractal dimension	Continuous
radius error	Continuous
texture error	Continuous
perimeter error	Continuous
area error	Continuous
smoothness error	Continuous
compactness error	Continuous
concavity error	Continuous
concave points error	Continuous
symmetry error	Continuous
fractal dimension error	Continuous
worst radius	Continuous
worst texture	Continuous
worst perimeter	Continuous
worst area	Continuous
worst smoothness	Continuous
worst compactness	Continuous
worst concavity	Continuous
worst concave points	Continuous
worst symmetry	Continuous
worst fractal dimension	Continuous
Target	Categorical

Data Pre-Processing

During this phase some additional operations are being performed on the dataset like handling missing values, scaling the values to a specific range, finding duplicate values etc. At beginning of this phase missing values present in each feature is examined. But it is found that none of the features containing any missing values as shown in Figure 2.

In the next step, presence of duplicate records is checked and it is found that the dataset is not containing any duplicate values. In the next step of pre-processing data scaling is performed.

Standard Scaling

Here the standard scaling method is used to transform the values between 0 and 1 (Swain and Pani, 2022). During this process it keeps the mean of 0 and standard deviation as 1. The detailed equation for standard scaling is given below in Eq. 1:

$$dscal = \frac{d-\mu}{\sigma} \quad (1)$$

Where d = actual feature value, μ = mean of feature, σ = standard deviation of the feature.

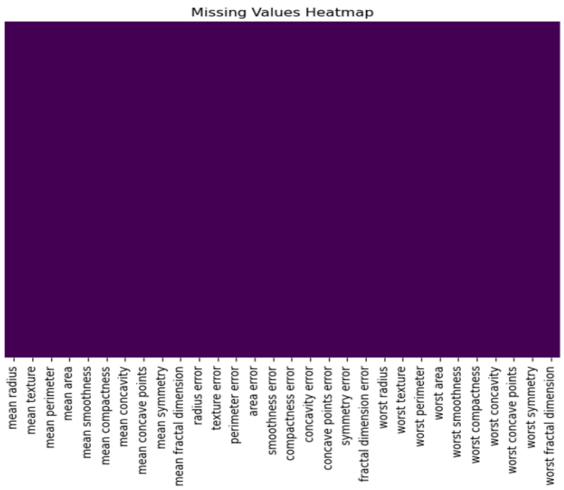

Fig. 2: Heat Map for missing value visualization

Data Balancing

The dataset is having unbalancing issue as shown in Figure 3. The target column contains unequal number of records for both classes of the target feature. The benign class contains 357 benign records (62%) and 212 malignant records (38%). An imbalanced datas*et al*ways creates a biased classifier towards the class containing more records. To avoid this issue SMOTE (Synthetic Minority Over-sampling Technique), method is used here. SMOTE is an advanced oversampling technique that generates synthetic examples of the minority class by interpolating between existing samples. Unlike random oversampling, which duplicates instances, SMOTE helps to reduce overfitting by creating new, realistic data points in the feature space. After applying SMOTE for each class 357 numbers of records are generated which is shown in Fig. 4.

SMOTE Algorithm-

1. Input:
   Minority class samples, oversampling ratio N, number of nearest neighbors k
2. For each sample x_i in the minority class:
   - Find its k nearest neighbors from the same class using Euclidean distance.
3. For each synthetic sample to generate:
   - Randomly select one of the k nearest neighbors x_{nn}
4. Generate synthetic sample by interpolation:

   $x_{new}=x_i+\delta * (x_{nn} - x_i)$

   where $\delta \in [0,1]$ is a random number.
5. Repeat steps 3–4 until the desired number of synthetic samples is created.
6. Output:
   Augmented dataset with original and synthetic minority samples.

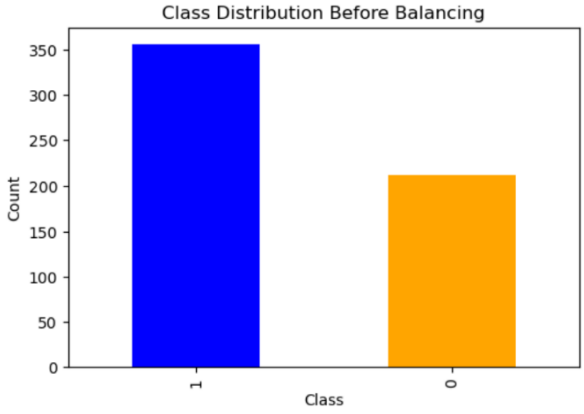

Fig. 3: Before balancing dataset

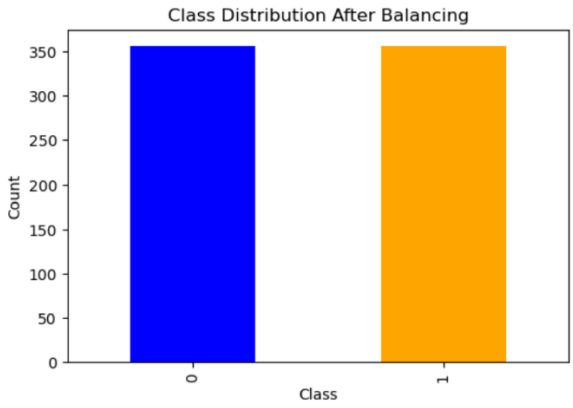

Fig. 4: After balancing dataset

Feature Reduction Using Principal Component Analysis

In this step the features reduction operation was performed using Principal Component Analysis. The detailed steps for PCA are given below. The principal components are identified based on their eigen values (Sehgal *et al*., 2014). PCA is mainly used to generate orthogonal axes that captures the maximum variance in different features. Each Principal component is a linear combination of all the features present in the dataset. If the total number of features is 'n' then each principal component is represented by vectors having length-n.

PCA Algorithm

Step 1: Compute the covariance matrix of the dataset.
Step 2: Find the eigenvectors and eigenvalues of this covariance matrix.
Step 3: Eigenvectors become the principal components, representing new axes.
Step 4: Rank the principal components on the basis of their Eigen values. (higher eigenvalues = more variance)
Keep only the top eigenvectors to reduce the number of dimensions while preserving information

The Eigen values are calculated using the following Equation 2:

$$C.v=\lambda.v \quad (2)$$

Where:

C = Covariance matrix of the dataset
v = Eigen vector
λ = Eigen value

After applying the principal component analysis, the number of features is reduced to 5. The eigen value of the Principal Components are shown in the following Table 2. Variance measures how much each feature varies on its own, while covariance captures how features vary together. PCA leverages the covariance matrix to identify directions (principal components) where the data shows the most combined variance, ensuring maximal information retention during dimensionality reduction.

Table 2: Principal Component

Principal Component	Eigen Value
PC-1	0.435
PC-2	0.200
PC-3	0.100
PC-4	0.063
PC-5	0.050

Model Training

Random Forest Classifier

For classifying the different data points in to two classes (Benign and Malignant) here RF classifier is used. The classifier is working on the principle of boot strap aggregation (Swain *et al*., 2024). Initially it creates different subsets D_i from the main dataset D. After that it creates different DT classifiers those have combinedly formed this ensemble classifier. During the formation of each classifier, it utilises 'n' number of features from the total number of features 'N' (where $n<N$). The final decision is found after finding the mode of the decision given by each DT (Swain *et al*., 2022). This process is known as Majority voting algorithm. The detailed steps for this algorithm are depicted in the following:

Random Forest Algorithm

Step 1: **Bagging (Bootstrap Aggregating):** Given a dataset D with N samples, multiple subsets D_i are created using **random sampling with replacement.**

Each subset D_i is used to train an independent **DT.**

Step 2: Random Feature Selection:

2.1. At each node split, instead of using all features, only a **random subset** of features is considered.

2.2. If there are M total features, only m (where m<M) features are randomly selected at each split.

Step 3: Decision Making (Voting):

For classification: The final prediction is made using **majority voting** from all trees.

$$\hat{y} = mode\ (T_1\ (x), T_2\ (x), \ldots\ldots\ldots\ldots, T_k\ (x)$$

Where $T_i\ (x)$ is the prediction made by the i^{th} tree.

Hyperparameter Tuning Using Cuckoo Search

Cuckoo search is an optimization algorithm used to find the best H-param value at which the algorithm outperforms. This algorithm is mainly based on the way a cuckoo bird lays its eggs in other bird's nest. It first prepares different H-param combinations to create sample solutions. After that it improves those solutions by making random jumps called Levy flights (Majumdar and Mallick, 2016). In this way the worst solutions are further replaced with the best solutions with step wise improvements. Here the different H-param and their values considered for tuning the RF model are given below:

n_estimators = {10 (min) to 200 (max)}
max_depth = {2 (min), 50 (max)}
min_sample_split = {2 (min), 20 (max)}
max_feature = {1 (min), 5 (number of principal component- max)}
criterion = {gini, entropy}

Cuckoo Search Algorithm

Step 1: Start

Step 2: Select H-param for model optimization that will be the part of each nest.
Nest= [n_estimators, max_depth, min_samples_split, max_features, criterion]

Step 3: Generate n-nests (initial solutions) by randomly selecting the value of H-param from their specified range. Each nest is evaluated to obtain its objective function score (Accuracy).
Ex- $nest_1$= [100, 20, 2, 5, "entropy"]

Step 4: Create new solution by making changes in the initial solution using Levy flights.
New nest = Old nest + α * Step size
Where α = Scaling factor, Step size = (max value-min value)

Step 5: Evaluate the new solution by using the accuracy of the Model.
If $new\ solution_{Accuracy} > old\ solution_{Accuracy}$ then
Accept new solution else discard it.

Step 6: Replace p_a abandoned solution with random values.
Where p_a= Abandonment Probability

Step 7: Repeat Step 3 to 5 for multiple iterations till best value of H-param are identified that gives best accuracy score.

Step 8: End.

The parameter setting for Cuckoo search implementation is shown as per the following:

i. Abandonment probability (Pa) was set to 0.25
ii. Step size (α) was set to 1.5
iii. The number of nests was 25
iv. The algorithm was run for 100 iterations

Results

The performance of the optimized model is assessed using different performance factors such as accuracy, precision, recall, ROC curve, F1- Score, PR curve and Cohen Kappa score. Accuracy is the ratio of correctly predicted cases with total number of cases. Precision measures the correctness of the model during only positive predicted cases. Recall finds the robustness of the model while dealing with the total number of positive cases. F1-score finds the harmonic mean of precision and recall (Swain *et al*., 2021). The formula for accuracy, precision, recall and F1-score are shown below in the Equations 3, 4, 5, 6. The model has reported accuracy, precision, recall and F1 score as 98, 98, 98 and 98%. The confusion matrix of the classification task is shown in Table 3. The untuned RF has shown an overall accuracy of 96% while classify the records.

Table 3: Confusion matrix

True positive	73
False Positive	1
True negative	67
False negative	2

$$Accuracy = \frac{I+K}{I+J+K+L} \quad (3)$$

$$Precision = \frac{I}{I+J} \quad (4)$$

$$Recall = \frac{I}{I+L} \quad (5)$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (6)$$

Where:

I = True Positive

J = False Positive
K = True Negative
L = False Negative

ROC Curve

It is a plot drawn between true positive rate and false positive rate. It gives a graphical presentation of the model performance while doing binary prediction at different thresholds (Swain *et al*., 2023a). The ROC curve is shown below in the Figure 5. The AUC score shown by the model is 0.99. The closer the score is to 1.0, the better the model is at predicting true positives while avoiding false positives. A score of 0.99 indicates that 99% of the time, the model correctly ranks a randomly chosen positive instance higher than a randomly chosen negative one.

PR Curve

It is curve drawn between Precision and Recall at different thresholds. It is a string indicator about how the classifier differentiates between the positive and negative cases (Swain *et al*., 2023b). The PR-curve of the model is shown in the following Figure 6. The average precision score reported by the model is 0.97. The AP score is the area under the precision-recall curve, and a score of 0.97 suggests excellent performance, with very few false positives across various threshold settings.

Stratified Cross-Validation

This operation is useful to test the generalizability of the model by maintaining the same class distribution as like original dataset while performing the training and testing in every fold of the cross-validation. The original dataset is having 357 benign records (62%) and 212 malignant records (38%). The following Table 4 shows the classification accuracy in different folds.

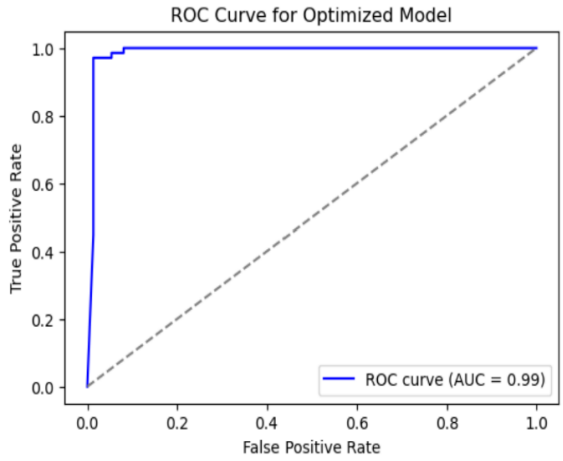

Fig. 5: ROC curve

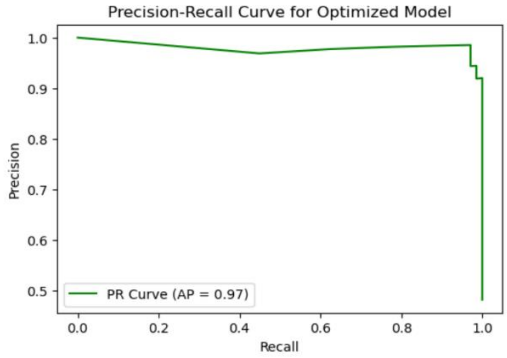

Fig. 6: PR curve

Performance Analysis

A detailed comparative analysis between the proposed system and different discussed literatures are given in the following Table 5. Data balancing is considered as one of the most crucial factors for getting better performance. An unbalanced data training always gives a biased classifier (Delen *et al*., 2005; Tripathy *et al*., 2014; Chaurasia *et al*., 2018; Esmaeili *et al*., 2020; Khatun *et al*., 2021). Initially the referred data was containing unbalancing issue with 357 (62%) data for malignant cases and 212 (38%) data for benign cases. To balance the data, Synthetic Minority Over-sampling method is used to balance the data count 357 records for each of the classes. Principal component analysis was performed in this work to reduce the number of features from 30 to 5. Feature elimination always helps a model to handle overfitting, faster the model training, and focuses on the most relevant features only. In all discussed works more than 5 number of features were used [5-13]. In the proposed system, the RF algorithm has combined the capability of a number of classifiers to create a powerful classifier. Hence the result obtained in the proposed model is more stable than others (Delen *et al*., 2005; Tripathy *et al*., 2014; Chaurasia *et al*., 2018; Esmaeili *et al*., 2020; Khatun *et al*., 2021; Pal *et al*., 2023). For improving the performance of the system cuckoo search optimization algorithm is used for H-param tuning. The benefit of using cuckoo search algorithm is that it uses Levy flights to walk around different feature space to avoid the local optima and explore global optima. Additionally, the algorithm helps to do faster convergence by adopting a perfect balance between exploration and exploitation strategies.

Table 4: Stratified Cross-validation Result

Fold No.	Accuracy
1	96.49%
2	92.11%
3	95.61%
4	95.61%
5	96.46%
Average Accuracy	95.26%

Table 5: Comparative analysis

Reference No.	Method Used	Accuracy
5	DT	93%
6	SVM	95%
7	Naïve Bayes	97%
8	K- NN	84%
9	Multi Layer Perceptron	85%
10	K-NN	95%
11	RF	96.5%
12	GB	71%
13	RF	93%
Proposed Method (Optimized Random Forest)		98%

Limitation of Existing Study

The proposed system addresses key limitations identified in the literature, specifically data imbalance, feature reduction, and hyperparameter optimization.

One of the general problems found with most of the existing studies is the absence of handling data balancing (Delen *et al*., 2005; Tripathy *et al*., 2014; Chaurasia *et al*., 2018; Esmaeili *et al*., 2020; Khatun *et al*., 2021). Data imbalance degrades model performance by biasing the classification process toward the majority class. At the initial level the dataset was having a distribution of 62% and 38% between the two target classes.

Another limitation relates with the usage of a greater number of features without performing any feature reduction method. The presence of a greater number of features not only enhances the computational cost but also it introduces some noise, that majorly impact the training process in the negative direction (Delen *et al*., 2005; Tripathy *et al*., 2014; Chaurasia *et al*., 2018; Esmaeili *et al*., 2020; Khatun *et al*., 2021; Pal *et al*., 2023). All the mentioned studies have involved more than 5 number of features which is greater than the number of features used in the proposed system.

The hyperparameter optimization another crucial aspect that has not considered in major of the studies. The unavailability of this method always runs the model with default values of the hyperparameters (Delen *et al*., 2005; Tripathy *et al*., 2014; Chaurasia *et al*., 2018; Esmaeili *et al*., 2020; Khatun *et al*., 2021; Pal *et al*., 2023, Ara *et al*., 2021; Yarabarla *et al*., 2019, Kaur and Gupta, 2024). Due to this problem most of the time the models suffer with the issue of overfitting. An overfitted model always shows a poor result during the validation of the model.

Conclusion

In this work, a ML-based screening system for breast cancer detection has been developed using the RF algorithm, with Principal Component Analysis (PCA) for feature extraction and Cuckoo Search optimization for H-param tuning. The use of the UCI breast cancer dataset for training and validation has demonstrated the effectiveness of the proposed approach. With an accuracy of 98%, the model provides a reliable and efficient alternative to traditional clinical decision-making, reducing dependency on subjective judgment. By enabling early and accurate detection of breast cancer, this system has the potential to significantly improve patient outcomes and aid physicians in making more informed diagnoses. This work can also be extended further to include ablation study to understand the benefit of different phases. The model can be trained and validated using some large dataset to have more generalizability.

Acknowledgment

The authors gratefully acknowledge Shri JJT University, Jhunjhunu and Pandit Deendayal Energy University, Gandhinagar for providing the necessary research facilities and academic support that enabled the successful completion of this research work. The encouragement and resources extended by both institutions played a vital role in shaping the progress of this study. The authors also appreciate the conducive academic environment and collaborative spirit that significantly contributed to achieving the research outcomes.

Funding Information

This research work is not funded by any organization.

Author's Contributions

Prasad S. Sase: Model development, model optimization, model validation.

Debabrata Swain: Data Collection, Pre-processing.

Shailesh Kumar: Data validation, visualization.

Ethics

This research did not involve any studies with human participants or animals performed by the authors. The study is based on publicly available secondary datasets, and no personally identifiable information was used at any stage of the analysis. All methods were carried out in accordance with relevant guidelines and regulations. The authors affirm that the work adheres to ethical standards of research integrity and transparency.

Conflict of Interest

The authors declare that there are no conflicts of interest.

References

Afaq, S., & Singh, N. (2024). Breast Cancer Detection using Deep Learning. *Proceedings of the 2024 International Conference on IoT, Communication and Automation Technology (ICICAT)*, 802–807.

Ara, S., Das, A., & Dey, A. (2021). Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms. *Proceedings of the IEEE International Conference*, 97–101. https://doi.org/10.1109/icai52203.2021.9445249

Chaurasia, V., Pal, S., & Tiwari, B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms & Computational Technology*, *12*(2), 119–126. https://doi.org/10.1177/1748301818756225

Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial Intelligence in Medicine*, *34*(2), 113–127. https://doi.org/10.1016/j.artmed.2004.07.002

Esmaeili, M., Ayyoubzadeh, S. M., Ahmadinejad, N., Ghazisaeedi, M., Nahvijou, A., & Maghooli, K. (2020). A decision support system for mammography reports interpretation. *Health Information Science and Systems*, *8*(1). https://doi.org/10.1007/s13755-020-00109-5

Feng, Y., Spezia, M., Huang, S., Yuan, C., Zeng, Z., Zhang, L., Ji, X., Liu, W., Huang, B., Luo, W., Liu, B., Lei, Y., Du, S., Vuppalapati, A., Luu, H. H., Haydon, R. C., He, T.-C., & Ren, G. (2018). Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. *Genes & Diseases*, *5*(2), 77–106. https://doi.org/10.1016/j.gendis.2018.05.001

Kathale, P., & Thorat, S. (2020). Breast Cancer Detection and Classification. *Proceedings of the 2020 International Conference on Emerging Trends in Information Technology and Engineering (Ic-ETITE)*, 1–5. https://doi.org/10.1109/ic-etite47903.2020.367

Kaur, A., & Gupta, S. (2024). Unveiling Precision in Breast Cancer Prediction with Random Forest and Decision Trees. *Proceedings of the IEEE International Conference*, 1232–1236. https://doi.org/10.1109/icosec61587.2024.10722493

Khatun, T., Utsho, Md. M. R., Islam, Md. A., Zohura, Mst. F., Hossen, Md. S., Rimi, R. A., & Anni, S. J. (2021). Performance Analysis of Breast Cancer: A Machine Learning Approach. *Proceedings of the 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, 1426–1434. https://doi.org/10.1109/ICIRCA51532.2021.9544879

Majumdar, D., & Mallick, S. (2016). Cuckoo search algorithm for constraint satisfaction and optimization. *Proceedings of the IEEE International Conference*, 235–240. https://doi.org/10.1109/icrcicn.2016.7813662

Pal, M., Parija, S., & Panda, G. (2023). Prediction of breast cancer using tools of machine learning techniques. *Onkologia i Radioterapia*, *17*(4), 1–6. https://doi.org/10.52793/oncoradiotherapy/2023174

Sehgal, S., Singh, H., Agarwal, M., Bhasker, V., & Shantanu. (2014). Data analysis using principal component analysis. *Proceedings of the IEEE International Conference*, 45–48. https://doi.org/10.1109/medcom.2014.7005973

Swain, D., & Pani, S. K. (2022). A support system for coronary artery disease detection using a deep dense neural network. *International Journal of Computing Science and Mathematics*, *16*(3), 292–305. https://doi.org/10.1504/ijcsm.2022.128187

Swain, D., Bijawe, S. S., Akolkar, P. P., Shinde, A., & Mahajani, M. V. (2021). Diabetic retinopathy using image processing and deep learning. *International Journal of Computing Science and Mathematics*, *14*(4), 397–409. https://doi.org/10.1504/ijcsm.2021.120686

Swain, D., Kumar, M., Nour, A., Patel, K., Bhatt, A., Acharya, B., & Bostani, A. (2024). Remaining Useful Life Predictor for EV Batteries Using Machine Learning. *IEEE Access*, *12*, 134418–134426. https://doi.org/10.1109/access.2024.3461802

Swain, D., Mehta, U., Bhatt, A., Patel, H., Patel, K., Mehta, D., Acharya, B., Gerogiannis, V. C., Kanavos, A., & Manika, S. (2023a). A Robust Chronic Kidney Disease Classifier Using Machine Learning. *Electronics*, *12*(1), 212. https://doi.org/10.3390/electronics12010212

Swain, D., Parmar, B., Shah, H., Gandhi, A., Acharya, B., & Hu, Y.-C. (2023b). Enhanced handwritten digit recognition using optimally selected optimizer for an ANN. *Multimedia Tools and Applications*, *82*(28), 44021–44036. https://doi.org/10.1007/s11042-023-15402-0

Swain, D., Parmar, B., Shah, H., Gandhi, A., Pradhan, M. R., Kaur, H., & Acharya, B. (2022). Cardiovascular Disease Prediction using Various Machine Learning Algorithms. *Journal of Computer Science*, *18*(10), 993–1004. https://doi.org/10.3844/jcssp.2022.993.1004

Tripathy, R. K., Mahanta, S., & Paul, S. (2014). Artificial intelligence-based classification of breast cancer using cellular images. *RSC Advances*, *4*(18), 9349. https://doi.org/10.1039/c3ra47489e

UCI Machine Learning Repository. (2018). Breast Cancer Wisconsin. https://doi.org/10.24432/C5DW2B

World Health Organization. (2021). Breast cancer. *WHO Fact Sheets*. https://www.who.int/news-room/fact-sheets/detail/breast-cancer

Yarabarla, M. S., Ravi, L. K., & Sivasangari, A. (2019). Breast Cancer Prediction via Machine Learning. *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, 121–124. https://doi.org/10.1109/icoei.2019.8862533