Enhancing Breast Cancer Detection with Machine Learning: A Predictive Modeling Approach

J. Sherine Glory[1*], M. Bhavani[2], B. Saratha[3], A. Akila[4], Dr. B. Prathusha Laxmi[5], Dr. V. Vijayaraja[6], N. Raghavendran[7]

[1*]*Assistant Professor, Department of Computer Science and Engineering, R.M.D. Engineering College, Kavaraipettai, Chennai*

[2]*Assistant Professor, Department of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai*

[3]*Assistant Professor, Department of Artificial Intelligence and Data Science, R.M.K. Engineering College, Chennai*

[4]*Assistant Professor, Department of Artificial Intelligence and Data Science, R.M.K College of Engineering and Technology, Chennai*

[5]*Professor, Department of Artificial Intelligence and Data Science, R.M.K. College of Engineering and Technology, Chennai*

[6]*Professor, Department of Artificial Intelligence and Data Science, R.M.K. College of Engineering and Technology, Chennai*

[7]*Assistant Professor, Department of Artificial Intelligence and Data Science, R.M.K College of Engineering and Technology. Chennai*

sherinegloryj@gmail.com[1], mbhavani1811@gmail.com[2], bsa.ad@rmkec.ac.in[3], aakilaads@rmkcet.ac.in[4], hod_ads@rmkcet.ac.in[5], vijayarajaads@rmkcet.ac.in[6], ragavendrannv2001@gmail.com[7]*

KEYWORDS

Breast Cancer; Machine Learning, feature selection, Naive Bayes and Random Forest

ABSTRACT

Compared to other methods, the one now used to diagnose breast cancer is not as sensitive or specific. The development of increasingly accurate machine learning algorithms for risk assessment, prediction, and treatment planning has made personalized breast cancer data a reality. The effectiveness of data produced by machine learning algorithms in identifying and categorizing breast cancer is examined in this research. In this post, we will go over a machine-learning strategy that may improve breast cancer diagnosis. To improve the speed and accuracy of diagnoses, our approach utilizes enhanced feature selection methods, reliable classification algorithms, and top-notch model training. After the models were built, we put in a lot of time and effort with the hyperparameters to evaluate various ML approaches. A ROC score of 1.00 for Naive Bayes and a score of 98.10% for Random Forest were the two best models. This study proves that ML algorithms, including the Naive Bayes and random forest methods, can accurately forecast breast cancer outcomes. Machine learning might be used to assess the situation in the future.

.

1. Introduction:

The second-biggest killer of women is breast cancer, followed by heart disease. Additionally, it affects about 10% of females. World Health Organisation data shows that around 500,000 women get a breast cancer diagnosis every year [1]. Women in impoverished countries sometimes reach an advanced stage with few treatment options since screening programs and awareness are lacking. The likelihood of breast cancer developing in a woman may be increased by certain factors. Her reproductive history (including nulliparity, early menarche, or late menopause), hormonal variables (including oral contraceptives or hormone replacement treatment), obesity, smoking, heavy alcohol use, or early radiation exposure are all potential causes [2]. Cancer affected a large number of people during this period. Since the ailment is linked to outside factors, we can't pinpoint its origin. This provides an additional resource for assessing the cancer's aggressiveness. When diagnosing cancer, three assessment factors should be considered: cell size consistency, shape regularity, and clamp thickness. Machine learning and other computer science technologies have grown in popularity in recent years, yet even those tasked with convincing others to act find the outcome difficult. Computer diagnostic tools have saved countless lives by identifying diseases that were killing millions before. When it comes to surgical procedures, robots are indispensable. Aside from its widespread use in cancer detection, AI has also shown to be a useful tool in the new system implemented in the critical care unit [3]. One in eight American women between the ages of 14 and 40 will get cancer at some point in their lives. Sagging breasts or tumours may result from uncontrolled cell division, which is the primary cause of breast cancer. Nonetheless, the tumour often does not endanger the patient's well-being. Accurately classifying patients' illnesses as benign or malignant is a huge challenge for clinicians; yet, understanding the components may aid with survival irrespective of the diagnosis.

Extensive research into artificial intelligence (AI) technology has been devoted to the creation of cancer prediction models [4, 5]. Artificial intelligence (AI) models, like ML algorithms, may analyse medical image collections and patient data to potentially identify breast cancer or predict the chance of developing the illness. Medical images, such as ultrasounds and mammograms, may have quantitative data extracted from them using principal component analysis (PCA).

Integrating several cancer risk indicators into AI-based prediction models might lead to the development of personalised imaging and treatment regimens. Factors like heredity, way of life, and surroundings are all part of this. One emerging area of AI that has the potential to greatly enhance the accuracy and timeliness of breast cancer diagnoses is deep learning (DL) algorithms [6]. These data-driven methods have the potential to revolutionise breast imaging by learning on their own and detecting complex patterns associated with cancer using vast amounts of data.

1.1 Motivation:

We should promote the use of machine learning in breast cancer diagnostics because it may lead to better patient outcomes, lessen the global burden of illness, and foster the creation of ground-breaking medical research and technology.

1.2 Contribution

- This study builds a robust prediction framework by using machine learning techniques to extract characteristics from pre-trained models.
- In-depth performance analysis employing meticulous evaluation measures validates the efficacy of the proposed method.
- This study's findings could revolutionise patient care by providing doctors with evidence-based information to improve breast cancer diagnosis and treatment.
- When compared to similar prior methods, the proposed model performs well.
- Four feature selection strategies—the Wrapper Method, Random Forest (RF) feature importance, Correlation Matrix, and Principal Component Analysis (PCA)—improve the performance of the ideal model.

2. To find the best model, retrain these ensemble methods with the returned geometrical information. This increases their accuracy.

Related Works:

Numerous studies have made use of preliminary machine learning to monitor breast lesions. Classified breast cancer using statistical features extracted from datasets, including discrete wavelet transform and Fourier cosine transform [7]. Using entropy, we were able to identify the top attributes. The vote classification method achieved an accuracy rate of 96.06% when several classifiers were used. Using the Hough transform, the authors presented a method for feature classification [8]. Following feature extraction, the SVM classifier classified tumours as benign, malignant, or normal. They were the most accurate at 95%. The use of t-tests for feature extraction is recommended. The ideal number of attributes was chosen by using a dynamic threshold. When employing SVM as a binary classifier, our accuracy in determining whether mammograms were benign or malignant was 95.08% [9].

The extraction of features was driven by considerations of morphology and texture. Results after therapy are better when breast cancer is detected early using deep convolutional neural networks. Several classification techniques were investigated [10]. We ran ten-fold cross-validation on the classification methods on eight NCD datasets. One way to determine precision is by looking at the area under the curve. The authors claim that NCD databases include a lot of irrelevant and disorganised material. Notable noise control skills are shown by KNN, SVM, and NN. The team settled on doing accuracy-checking preprocessing and removing unnecessary features. Asthma is one medical condition that bionic inspiration computing (NIC) might detect. Using insects, five distinct methods were suggested for the detection of NIC diabetes and cancer [11]. The authors found breast, ovarian, prostate, and lung cancers. Algorithms for detecting breast cancer are enhanced by using neural networks for directed ABC. Researchers have made significant progress in the early diagnosis of diabetes and leukaemia. Using conventional classification methods with NICs may provide more reliable and promising results. The focus was on ongoing studies that aim to identify illnesses and diabetes. There is some evidence that NNs might be useful for early cancer diagnosis [12]. According to their research, a large number of NNs may be able to identify cancer cells.

3. Methods and Techniques:

The suggested approach, is seen in Figure 1. Education, risk assessment, and understanding are all aided by the proposed breast cancer detection technique, which is an efficient feature selection-based strategy. The suggested model is based on the following five points: (I). Gathering information from a variety of sources according to predetermined criteria is known as data collection. (ii) Data pre-processing. Take outliers from the dataset as well. (iii) Split the data such that it may be used for training and testing. (iv) To verify the accuracy of the classification results, the model employs Random Forest and Naive Bayes classifiers. (v) The Precision, F1-score, recall, and accuracy of the performance evaluation.

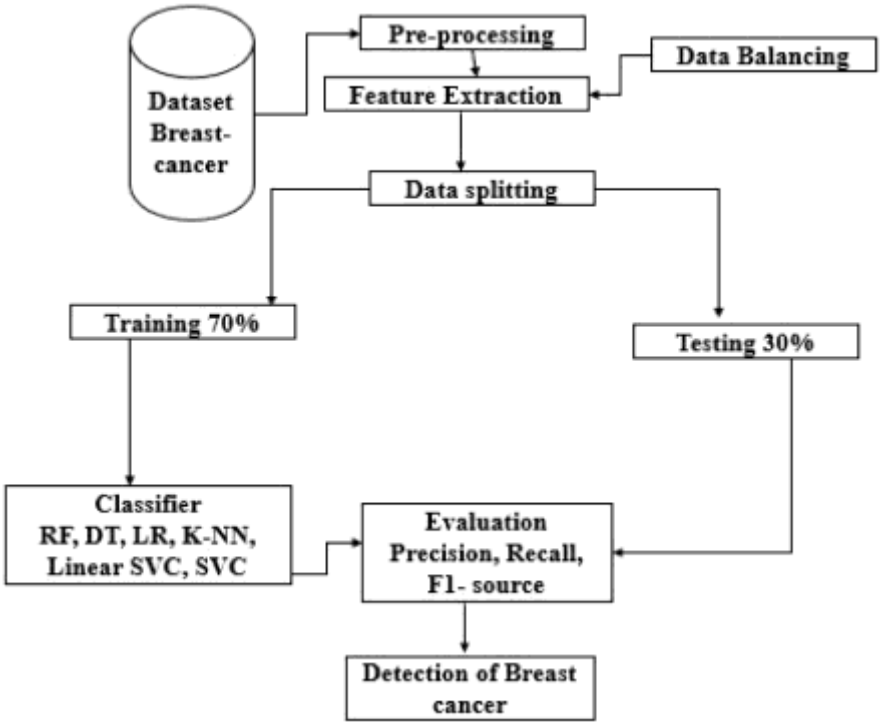

Figure 1: Proposed Architecture

3.1 Dataset:

In the fields of medicine and artificial intelligence, the Breast Cancer Wisconsin (Diagnostic) dataset is a famous and widely used resource. This dataset, derived from digital images of fine needle aspirates (FNA) of breast masses, may help in breast cancer identification by analysing the characteristics of cell nuclei. In this piece, we go into the metadata, numbers, and significance of this dataset. There are 569 occurrences, 30 numerical attributes utilised for prediction with class labels, and 212 malignant and 357 benign datasets.

3.2 Data Preprocessing:

Incomplete and improperly formatted data might result from preprocessing procedures that transform real-world data into unstructured, null, or error data. During the pre-processing phase, we looked at data normalisation, encoding, missing values, and feature selection. Optional characteristics: Sequential forward selection is used because feature selection is important. We used this strategy to get rid of dataset characteristics that weren't statistically significant. From the sets we used for training and testing, we removed the following: "id," "perimeter," "area-mean," "concavity," "radius," "area," "radius worst," "texture worst," in that order. Exploratory data analysis is the first step in finding and fixing missing values. Missing data was filled up using iterative imputers. Features are shaped via iterative imputation using other features as a basis. Encoding is a prerequisite for pre-processing ML models and other forms of data analysis. The goal is to make data more amenable to algorithms. The pre-processing phase makes use of these encodings. There can be a lot of data fields in the dataset. To classify things effectively, you need to decode data values using the same object type. To change the numerical column values without removing numbers or changing the range of potential values, the parameter values are scaled between zero and one before being put on a standard scale.

3.3 Data split up:

Steps include data collection, data partitioning (80% training, 20% testing), feature selection, method selection (supervised machine learning), sample classification (benign or malignant tumours), and model evaluation. Acquiring the dataset required for model training is the first stage. We clean and turn the raw data into machine learning after the dataset has been gathered. This is called preprocessing. Two subsets of the dataset were created during the preprocessing phase: one for training the model and another for testing it.

3.4 Feature Extraction:

The Breast Cancer Wisconsin dataset is one example of a high-dimensional dataset that Principal Component Analysis (PCA) use to reduce its dimensionality without losing any of the essential information. Thirty numerical characteristics, culled from digital photographs of benign and cancerous breast mass nuclei, make up this dataset. Using the covariance matrix to compute eigenvectors and eigenvalues, principal component analysis (PCA) transforms datasets into orthogonal principle components, which it then uses to detect patterns in the dataset. Since most of the variation is included in the first few components, we choose to look at these to summarise the dataset's variability. Using principal component analysis (PCA) to identify important features like mean radius, texture, perimeter, and compactness could improve visualisation and reduce processing complexity in machine learning algorithms used to diagnose breast cancer. By reducing the number of dimensions, predictive models become more accurate and overfitting becomes less probable. Crucial for clinical decision-making, the classification findings are now more robust and transparent.

3.5 ML Model Classification:

Automated learning for precise disease prognosis in breast cancer patients. We have achieved machine learning algorithms. It is generally agreed upon that deep learning is a subset of machine learning. This new dataset was subjected to two independent tests to determine the efficacy of the machine learning approach. These classes are used to classify methods such as Random Forest and Naive Bayes.

3.5.1 Random Forest Classification:

Due to its accuracy and robustness, Random Forest classification is a well-liked machine-learning approach for breast cancer classification, as shown in Table 1. To improve generalisability and decrease overfitting, train a cluster of decision trees and then aggregate their predictions. To identify benign or malignant tumours, the algorithm examines the Breast Cancer Wisconsin dataset for features including cell radius, texture, and compactness. To identify the most important traits, Random Forest databases rank feature values and use them to categorise high-dimensional data. Ideal for medical datasets because of its inherent noise and absence of data management capabilities. It is possible to improve results by adjusting hyperparameters such as maximum depth and tree count. The model is a great tool for helping doctors detect breast cancer early and accurately because of its validity and interpretability, which improves patient outcomes.

Table 1: Pseudocode for Random Forest

Step 1: Build the Random Forest Model
  Randomly divide the dataset into subsets (bootstrap sampling).
  For each subset:
    Randomly select a subset of features.
    Train a decision tree using the selected features and subset.
  Combine all the trained decision trees to form the Random Forest.
Step 2: Classification Using the Random Forest
  Pass the input sample through all decision trees.
  Collect the classification results (votes) from each tree.
  Use majority voting to determine the final class label.

3.5.2 Naïve Bayes Classification:

A common probabilistic method for breast cancer classification, Naïve Bayes classification is based on Bayes' Theorem and is known for its simplicity and efficacy. It streamlines computations and produces correct answers for several medical datasets assuming feature independence. Using Naïve Bayes, the Breast Cancer Wisconsin dataset determines whether a tumour is benign or malignant by analysing characteristics such as mean radius, texture, and symmetry. Combining the prior probabilities with the likelihood of observable qualities dependent on the class, this

technique yields the posterior probability for every class. Despite the need for independence, it is computationally efficient, performs well with high-dimensional datasets, and can be utilised in real time. Naïve Bayes successfully handles small datasets with missing data, as shown in Table 2, and delivers a rapid and intelligible response to classification difficulties. They are useful for clinical decision support systems and early diagnosis of breast cancer since they may achieve satisfactory accuracy with little training data.

Table 2: Pseudocode for Naïve Bayes
Step 1: Model Training Compute the prior probabilities of each class by dividing the number of samples in each class by the total number of samples. Calculate the likelihood for each feature given each class using a probability distribution (e.g., Gaussian for continuous features), storing these probabilities. Step 2: Classification For a given input sample, compute the posterior probability for each class by multiplying the class prior with the likelihoods of the sample's features given the class. Assign the class label corresponding to the highest posterior probability.

4 Results and Discussion:

To conduct the studies, Anaconda's Python code was used. Pandas is used to load the data collection, and Pilots is used to see the Python packages. Machine learning techniques may also be written in Python. To conduct the trials, a Windows 10 PC with the following specifications was used: Core i7, 8GB RAM, a 620 GPU, a 2.9 GHz CPU, and 5GB HDD. The proposed model's correct operation was ensured by this. Breast cancer was assessed using a conventional machine learning algorithm that used RF and NB. The hyperparameters shown in Table 3 are used to test all methods.

Table 3: Algorithm Parameters value

Algorithm	**Parameters**
Random Forest	max_depth = 15, min_samples_split = 5, min_samples_leaf = 4, splitter = ‘random’, criterion = ‘entropy’
Naïve Bayes	learning_rate = 0.01, n_estimators = 180, max_depth = 5

Table 4: Algorithm Performances

Algorithm	**Accuracy**	**F1-Score**	**Precision**	**Recall**
Random Forest	99.5	98.45	98.10	97.52
Naïve Bayes	99.2	98.30	98.12	97.54

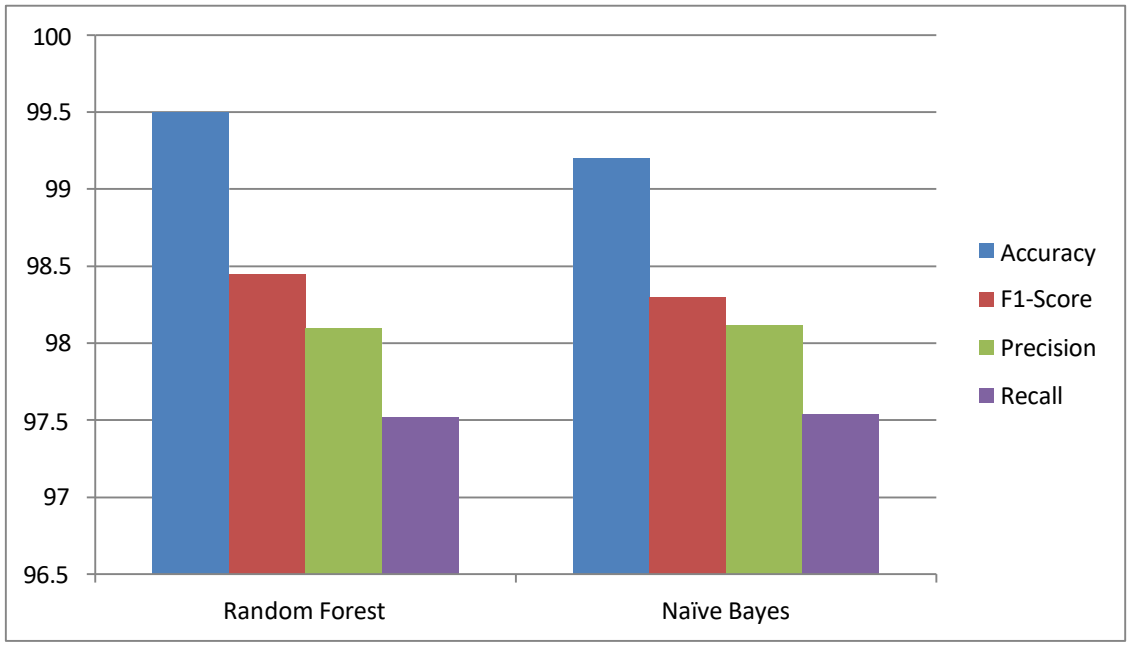

Figure 2: Algorithm Performance Metrics

The performance metrics of Random Forest and Naïve Bayes, two machine learning algorithms, in a classification task that is probably associated with breast cancer detection are shown in Table 4 and Figure 2. It primarily employs recall, accuracy, precision, and F1-Score as assessment measures.

Compared to Naïve Bayes, which has an accuracy rate of 99.2%, Random Forest performs better. Additionally, the F1-Score of Random Forest, which is a harmonic mean of Precision and Recall, is at 98.45, in contrast to Naïve Bayes's 98.30. The accuracy scores, which indicate how well the algorithm can detect favourable conditions, are almost equal for both models; Random Forest has a score of 98.10 and Naïve Bayes has a score of 98.12. Naïve Bayes somewhat surpasses Random Forest in the recall, a metric that gauges the capacity to identify all genuine positive instances, with 97.54 compared to 97.52. These findings show that both algorithms are successful, although Random Forest slightly exceeds it.

5 Conclusion:

The results show that supervised machine learning methods might be useful for forecasting when breast cancer will be detected in its early stages. Two supervised machine-learning algorithms were used to examine primary data from 500 patients at Dhaka Medical College Hospital in this study. Random Forest obtained a maximum accuracy of 99.5% in this study's assessment. When comparing Random Forest algorithms to Naive Bayes algorithms, the former performed better across the board (0.97 recall, 0.98 accuracy, and 0.98 F1 score). But some restrictions must be considered. To test the model's generalisability to bigger populations, more real-world data is needed. Another issue was that the dataset did not include raw scans but rather inferred picture characteristics, which prevented it from being used for direct image-based analysis. Notwithstanding these limitations, this study provides a strong proof-of-

concept for enhancing AI-based breast cancer diagnosis. This technique has the potential to speed up patient screenings and help pinpoint situations requiring more investigation due to high risk. The implementation of tailored treatment plans and prompt intervention would have a profound effect on survival rates in such a scenario. A potential next step in making the model more resilient would be to validate its performance on bigger multi-centre datasets that include medical imaging. The development of a more accurate cancer detection system may benefit from testing out combinations of human oncologists' knowledge and computer forecasts.

References:

1) Siegel Mph, R. L. et al. Cancer statistics, 2023. pathologyinnovationcc.org Siegel, KD Miller, NS Wagle, A JemalCa Cancer J Clin, 2023, pathologyinnovationcc.org 73, 17–48 (2023).
2) Akter, S. et al. Recent advances in ovarian cancer: Therapeutic Strategies, potential biomarkers, and technological improvements. Cells 11, 650 (2022).
3) Tsochatzidis, L., Costaridou, L. & Pratikakis, I. Deep learning for breast cancer diagnosis from mammograms—A comparative study. J. Imaging 5, 37 (2019).
4) He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med (2019) 25:30–6. doi: 10.1038/s41591- 018-0307-0
5) Kim KH, Lee SH. Applications of artificial intelligence in mammography from a development and validation perspective. J Korean Soc Radiol (2021) 82(1):12. doi: 10.3348/jksr.2020.0205
6) Hamed G, Marey MAER, Amin SES, Tolba MF. (2020). Deep learning in breast cancer detection and classification. In: Proceedings of the international conference on artificial intelligence and computer vision, advances in intelligent systems and computing. Springer, Cham (2020) 1153:322–33. doi: 10.1007/978-3-030-44289-7_30
7) Tang, X.; Zhang, L.; Zhang, W.; Huang, X.; Iosifidis, V.; Liu, Z.; Zhang, M.; Messina, E.; Zhang, J. Using Machine Learning to Automate Mammogram Images Analysis. In Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Korea, 16–19 December 2020; pp. 757–764.
8) Vijayarajeswari, R.; Parthasarathy, P.; Vivekanandan, S.; Basha, A.A. Classification of Mammogram for Early Detection of Breast Cancer Using SVM Classifier and Hough Transform. Meas. J. Int. Meas. Confed. 2019, 146, 800–805.
9) Meselhy Eltoukhy, M.; Faye, I.; Belhaouari Samir, B. A Statistical Based Feature Extraction Method for Breast Cancer Diagnosis in Digital Mammogram Using Multiresolution Representation. Comput. Biol. Med. 2012, 42, 123–128.
10) Fatima, N.; Liu, L.; Hong, S.; Ahmed, H. Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis. IEEE Access 2020, 8, 150360–150376.

11) Al-Antari, M.A.; Al-Masni, M.A.; Park, S.-U.; Park, J.; Metwally, M.K.; Kadah, Y.M.; Han, S.-M.; Kim, T.-S. An Automatic Computer-Aided Diagnosis System for Breast Cancer in Digital Mammograms via Deep Belief Network. J. Med. Biol. Eng. 2017, 38, 443–456.
12) Khan, F.; Khan, M.A.; Abbas, S.; Athar, A.; Siddiqui, S.Y.; Khan, A.H.; Saeed, M.A.; Hussain, M. Cloud-Based Breast Cancer Prediction Empowered with Soft Computing Approaches. J. Healthc. Eng. 2020, 2020, 8017496.