International Journal on Computational Modelling Applications
Vol. 02, Iss. 02, S. No. 04, pp.45--62, June 2025
ISSN (Online): 3048-8516

Received: 12 April 2025, Accepted: 25 April 2025, Published: 08 May 2025
Digital Object Identifier: https://doi.org/10.63503/j.ijcma.2025.89

Research Article

Comparative Analysis of Machine Learning Techniques for Early Detection of Breast Cancer

Prachi Rawat[1], Rashmi Saini[1] , Anuj Kumar[2*]

[1]Computer Science and Engineering Department, G. B. Pant Institute of Engineering and Technology, Pauri Garhwal 246194, India

[2]Department of Computer Science, School of Technology, Doon University, Dehradun 248001, India

prachi.rawat03@gmail.com, 2rashmisaini@gmail.com, anujkumar.sops@doonuniversity.ac.in*

Corresponding author: Anuj Kumar, anujkumar.sops@doonubiversity.ac.in

ABSTRACT

Breast cancer is the most frequently encountered form of cancer among the populace, with women being more susceptible than men to its development. Catching it early increases the likelihood of survival, but due to the complex nature of masses and microcalcification, radiologists oftentimes fail to diagnose breast cancer properly. Radiologists use computer-aided diagnostic (CAD) systems to detect abnormalities; however, several uncertainties in breast cancer detection using mammograms make it challenging. The advancement of machine learning (ML) in the medical field for diagnosis and improving its accuracy is an inevitable futuristic step. ML techniques in breast cancer detection greatly help in early and accurate detection, thereby increasing the patient's survival rate. This paper compares various popular machine learning techniques, such as support vector machine (SVM), random forest (RF), k-nearest neighbor (k-NN), and decision tree using the Wisconsin Breast Cancer dataset. The dataset consists of 569 biopsy samples, each characterized by 30 feature sets, classified as benign or malignant. The models were examined via a 70/30 train-test split and graded based on various metrics for performance evaluation, such as accuracy, precision, recall, F1 score, specificity, false positive rate, and false negative rate. The findings reveal that SVM performed with precision (94%) and specificity (97.22%), and RF found accuracy (92.40%) and F1 score (89.43%). These results indicate that approaches for machine learning, such as RF and SVM, substantially enhance the early detection of breast cancer, benefiting patient outcomes. Future studies should focus on looking at more comprehensive datasets and other techniques to improve diagnosis accuracy.

Keywords: *Machine learning; Breast cancer; Early detection; Classification;Medical Diagnosis*

1. Introduction

The cancer that affects women worldwide with the highest prevalence is breast cancer, characterized by malignant growth that begins in breast cells. The most typical indicators of breast cancer are the emergence of a new lump or tumor, localized swelling or thickening of breast tissue, nipple retraction or spiraling inward, and nipple discharge (other than breast milk). Early detection through self-examination, mammograms, and other diagnostic tests can improve the chances of successful treatment. Surgical procedures, radiation therapy, chemotherapy, and hormone therapy are among the available treatments for breast cancer. The occurrence of breast cancer can vary depending on age, family history, and other factors. Regular breast cancer examinations and early detection can enhance the likelihood of effective therapy. Breast cancer is typically found in the ducts and lobules, with the former being the conduits for milk transport to the nipple and the latter being the milk-producing glands. If the cells lining the milk ducts of the breast have become cancerous then it is

DCIS (Ductal carcinoma in situ). Lobular carcinoma in situ (LCIS) has cancerous cells in the linings of milk-producing glands (lobules) inside the breast. Women after puberty suffer from Breast cancer in every part of the country. Breast cancer has an increasing rate of occurrence in later life. In 2020, 2.3 million women had breast cancer diagnoses globally, and the illness claimed 68,500 lives [1]. By 2020, 7.8 million women have been diagnosed with breast cancer in the past five years. [1]. The high number of diagnoses of breast cancer makes it the most widespread cancer on the planet [1]. Breast cancer results in more disability-adjusted life years (DALYs) lost among women worldwide than any other cancer form [1]. Detecting breast cancer promptly may have a pivotal function in decreasing the rising number of deaths caused by the disease.

Despite numerous increasing studies in the field of medicine and various technological developments contributing to cancer treatment, problems in cancer diagnosis still happen. Globocan 2018 statistics indicate that breast cancer's occurrence in women worldwide due to aging was 23.7 per 100,000, and the mortality rate was 6.8 per 100,000 in 2018 [2]. A range of factors can trigger alterations to DNA or RNA, resulting in the conversion of healthy cells into cancerous ones, such as an increase in entropy, the natural aging of DNA and RNA, exposure to nuclear radiation, electromagnetic radiation, chemicals, bacteria, parasites, fungi, viruses, heat, food, water, cellular injury, and free radicals [3]. Effective tumor diagnosis is very important. A large number of tumors are benign (non-cancerous) in nature. However, incorrectly diagnosing a malignant tumor as benign can significantly reduce the effectiveness of the treatment. Early intervention and treatment of small, non-metastasized breast cancer typically result in a favorable outcome. Routine screening tests are a reliable method to detect breast cancer in the initial phases [4].

Accurate and reliable diagnosis is vital in the timely detection of breast cancer, as it helps classify tumors as benign and malignant. An effective detection strategy is characterized by a low rate of both false negatives (FN) and false positives (FP). Previously, mammography was deemed the most dependable and effective modality to use in the detection and diagnosis of breast cancer [5]. A combination of approaches is used for breast cancer analysis and detection including imaging, physical examination, and biopsy [6]. Mammography and ultrasound imaging techniques are used for breast cancer detection. Mammograms are images of breasts produced by X-ray. Effective screening of mammograms relies on Radiologists to detect signs of breast cancer [7]. Patients with unmistakable breast cancer situations may undergo mammogram and sonogram examinations with both benign (or nonspecific appearance) and normal [8]. A biopsy is an invasive surgical operation performed to corroborate the presence of symptoms related to breast cancer; however, it impacts the psychological and physical well-being of patients. Moreover, according to some research evidence, the density of the breasts can affect the accuracy of breast cancer detection by radiologists, with up to 30% of cases potentially being missed [9]. Evaluation of the mammograms for breast cancer is done with the aid of two robust potent indicators which are first, masses and second, micro-calcifications. Since masses often exhibit poor contrast in images, it is more arduous to detect masses than to detect micro-calcification [10].

Despite the rise in breast cancer cases over the last decade, mortality rates for breast cancer have dropped for women of all ages [11]. The positive trend of decreased mortality rates may be attributed to advancements in the treatment of breast cancer and the extensive implementation of mammography screening. Despite their proficiency, it is widely acknowledged that many experienced radiologists may still overlook a significant number of anomalies [12]. Furthermore, a substantial number of mammographic abnormalities that undergo biopsy ultimately prove to be non-cancerous. Factors such as training, experience, and subjective criteria can impact radiologists' ability to correctly interpret mammograms. The rate of inter-observer variations among trained experts has been observed

to be as high as 65-75% [13]. The use of computer-aided diagnosis (CAD) for examining mammograms can aid radiologists in the detection and classification of whether a mass is cancerous or not According to the literature survey, biopsies conducted on possible cancer cases yield benign results in 65-90% of cases, hence it is very crucial that to distinguish between the malignant and benign lesions, by developing such technique. Detection accuracy would greatly improve if we combine expert knowledge with CAD and ML techniques [32]. Its effectiveness can be witnessed through the results - detection accuracy with CAD was obtained above 90% and without CAD below 80% [14]. For automated diagnostic systems, CAD approaches facilitate the continuous observation of a large patient population in the ICUs (Intensive Care Units) as the conventional methods rely on human observers for monitoring and diagnosing. The way these techniques operate is by converting diagnostic criteria that are primarily qualitative into a problem of feature classification that is more objective and quantitative [15]. Figure 1 shows the configuration of a classification system. No stage is independent as shown by the feedback arrows.

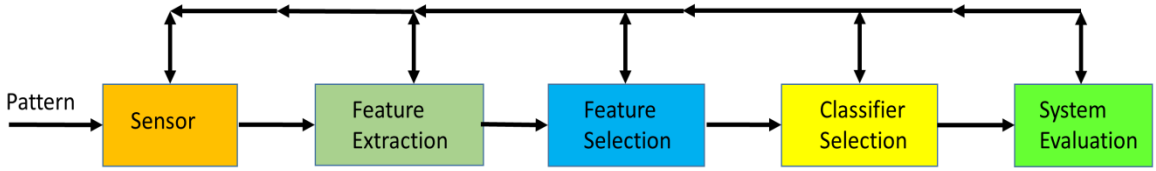

Figure 1: The various phases of a standard CAD system used in cancer detection [16]

Computers can learn on their own through ML, a subfield of Artificial Intelligence (AI), as long as they are exposed to datasets and gain knowledge through experience [17]. The widespread application of ML methods has been observed in recent decades for predictive model development supporting effective decision-making. By examining the dataset, these methods are used in cancer research to find patterns and predict whether a lesion is a benign cancer or a malignant cancer. Performance evaluation measures include classification accuracy, recall, and precision, and the area under the ROC [18] guides the assessment of how well these methods work. By analyzing and contrasting a breast cancer dataset, this research paper investigates the use of several innovative ML classifier models including SVM, k-NN, Random Forest, and Decision Tree to categorize tumors as either benign tumors or malignant tumors.

2. Related Work

Using Linear Discriminant Analysis, GLCM (Gray-Level Co-occurrence Matrix) and Optical Density Co-occurrence Matrix features were retrieved and used for breast cancer diagnosis in the work by Tai et al. [19]. Using an automated method, Nagi et al. [20] addressed the segmentation of the breast region from digital mammograms. Thresholding the mammogram at a constant grayscale level of 18 generated a binary mask. The binary mask was then filtered to remove artifacts, markers, and labels, thereby leaving only the breast region. However using a fixed grayscale intensity level for segmentation can lead to over-segmentation of the breast region, particularly in light of the presence of low-intensity pixels (below 18 grayscale) near the interface between the skin and air in mammograms.

A novel approach to feature invention proposed by Butler et al. [21] uses previous knowledge of the physical processes involved in X-ray scatter to find new features more suited for cancer detection than current ones. Using a simple naive Bayes classifier, the method extracts high-level features from low-level pixel data. One difficulty in this work is the complicated character of X-ray scatter patterns from heterogeneous tissue samples.

To segment the breast region, Qayyum and Basit [22] proposed a method combining morphological procedures and adaptive thresholding. The segmented mammography is then used to get a collection of features including statistical data, morphological features, and textural aspects. The SVM classifier classifies the mammogram as either benign or malignant using the specified feature set as input. The proposed method was evaluated using two datasets: MIAS and DDSM. The findings reveal that the proposed method may identify cancer in mammograms and achieve accurate breast segmentation. The authors found that in breast cancer diagnosis their approach might be an additional tool for radiologists. Htay and Maung [23] proposed a technique to find breast cancer in its underlying stages using K-Nearest Neighbor (k-NN) classification of mammography images and Gray Level Co-occurrence Matrix (GLCM) based feature extraction. Mammography images using the GLCM method show textural characteristics like contrast, homogeneity, energy, and entropy. Using these variables, the k-NN classifier classifies the mammogram as benign or malignant. The researchers discovered that their proposed method might be a non-invasive way to identify early-stage breast cancer by allowing earlier intervention and therapy, hence improving patient outcomes.

Based on the Random Forest (RF) algorithm, Dai et al. [24] propose a breast cancer diagnosis method. The technique classifies breast lesions as either benign or malignant using a range of clinical characteristics and imaging characteristics. The clinical attributes are age, menopausal status, and tumor size; the imaging features are texture features, shape features, and margin features. The proposed approach was evaluated using 569 breast lesion samples—357 malignant cases and 212 benign ones. The dataset used in this study originated from the Wisconsin Breast Cancer Diagnostic Center. Particularly in settings without specialist radiologists, the authors found great promise for breast cancer diagnosis in their proposed approach using the RF algorithm. Hamed et al. [25] present a CAD system for breast cancer diagnosis based on ML algorithms. Machine learning models were trained and evaluated using the Wisconsin Breast Cancer Dataset, which comprised 569 breast cancer biopsy samples. The overall performance of many ML approaches including random forest, logistic regression, AdaBoost, decision trees, naïve Bayes, and conventional neural networks (CNNs) is compared. The results show that the RF approach had the greatest accuracy of 98.9% and a better F-measure of 99% relative to the other methods. The paper emphasizes the need for further research in this domain, offers insightful analysis of the application of ML in healthcare, and finds that ML algorithms may assist in enhancing diagnostic accuracy and properly identifying breast cancer.

3. Methodology

The processes in the proposed approach are as follows: pre-processing, feature extraction, classification, and training/ testing of the input data. The system intends to correctly classify mammography images as either normal or malignant by using models—SVM, Random Forest, Decision Tree, and k-NN. The dataset showed no signs of missing values. Extracted features, in both individual and combination form, constitute the input for ML classifiers like SVM polynomials, Decision Tree, k-NN, and Random Forests. Test-train split method was utilized to divide the data into training and testing sets, which were subsequently utilized to classify the benign and malignant cancer subjects. Figure 2 illustrates the methodology diagram.

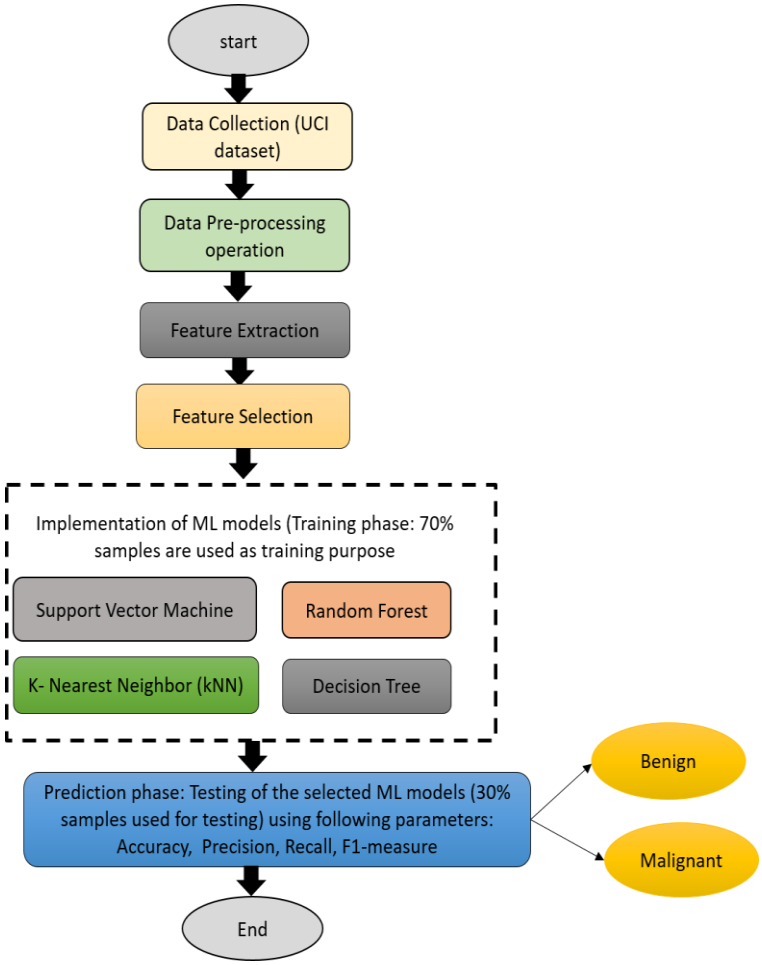

Figure 2: Methodology

3.1. Dataset Description

The work is carried out by using Wisconsin Diagnosis Breast Cancer dataset supplied by the UCI Machine Learning repository. It contains 569 samples of biopsy images of breast masses, each labeled as either benign (not harmful) or malignant (cancerous). The dataset includes 30 features computed from the digitized images of biopsy specimens, including radius, texture, perimeter, etc. By employing the features, the observations were classified as either benign (357 observations) or malignant (212 observations). Benign cases are labeled using the '0' class, whereas malignant cases are labeled using the '1' class.

Among various feature selection techniques, we made use of Ginni index as it provides a measure of impurity and is effective in identifying features that have a strong correlation with the target variable. The Ginni index aids in pinpointing the essential features that are crucial for attaining accurate breast cancer classification. The attributes selected by it for training are - 'radius_mean', 'area_mean', 'perimeter_mean', 'compactness_mean', 'concave points_mean', 'concavity_mean', 'radius_se', 'area_se', 'radius_worst', 'compactness_worst', 'perimeter_worst'. Figure 3 illustrates the distribution of different types of symptoms in our dataset.

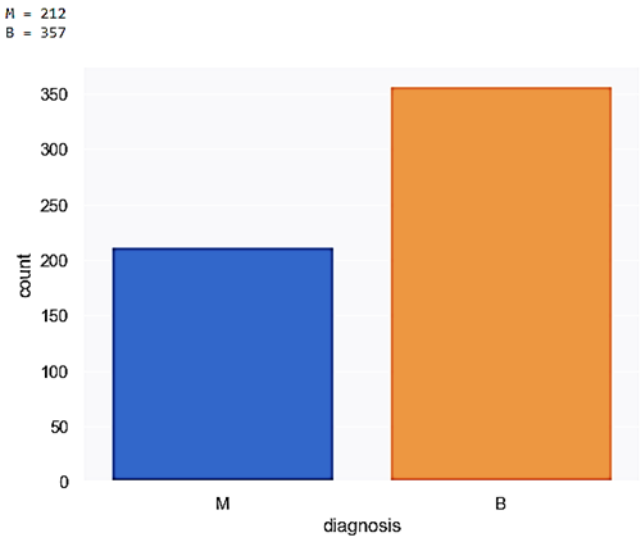

Figure 3: Different types of symptoms distribution

3.2. Training Set

Test-train split is a technique in ML to gauge the efficacy of a model when applied to a dataset. Two subsets—the training set and the test set—comprise the randomly split dataset. The latter is employed to gauge the model's performance, whilst the former is utilized for model training. With a typical split ratio of 80/20 or 70/30, it is usual to have a bigger training set than a test set. The test set lets us measure the model's performance on unknown data, therefore guaranteeing its generalization and preventing overfitting. Our model uses a 70/30 split ratio.

3.3. Simulation Model

The simulation model was developed using Python, leveraging the capabilities of Scikit-learn (sklearn). Among other things, it provides a variety of tools for model selection, clustering, dimensionality reduction, regression, classification, and data preparation. Scikit-learn's algorithms are executed easily and effectively, therefore helping both learners and professionals. It also provides a uniform interface to many ML methods and straightforward integration with NumPy and Pandas and other scientific libraries in Python. Widely employed in enterprises as well as academia for various machine learning tasks, it has a robust network of users and collaborators. Scikit-learn (sklearn) does not use CSV files or raw images as input. Data must be preprocessed and converted into sparse matrices or numerical arrays, which may then act as input for sklearn algorithms. For instance, the pandas package reads the CSV file and transforms the data into a numerical array, which might then be used as input for sklearn algorithms. Similarly, for imagine datasets, features have to be extracted from the images and converted into numerical arrays before being input into sklearn algorithms.

3.4. Machine Learning Techniques

Based on the kind of learning involved, ML methods may be roughly classified into supervised and unsupervised learning. In supervised learning, the computer is trained to provide the correct result using the labeled data instances. Since it means dealing with unlabelled data and no output expectation, unsupervised learning is a more difficult process than supervised learning.

3.4.1. Support Vector Machine (SVM)

SVM, a popular and effective supervised ML model, is often used in cancer diagnosis and prognosis. The SVM algorithm employs the selection of critical samples from each class as support vectors and creates a linear function that divides and separates the classes in the broadest possible

manner. Therefore, the SVM algorithm is employed to map an input vector to high dimensional space to discover the best hyperplane for segregating the dataset into different classes [26]. The linear classifier aims to locate the most suitable hyperplane that maximizes the marginal distance [27].

The scatter plot depicted in Figure 4 showcases two classes with two properties. ax_1 + bx_2 is used to define a linear hyperplane. The objective is to determine values for a, b, and c which meet the criteria of ax_1 + bx_2≤ c for class 1 and that ax_1 + bx_2 > c for class 2 [26], [28]. SVM sets itself apart from other ML approaches by heavily relying on support vectors. The data sets that are in the closest proximity to the decision boundary are referred to as support vectors. The reasoning is that eliminating data points located further from the decision hyperplane has a lesser effect on the boundary than that of support vectors' elimination.

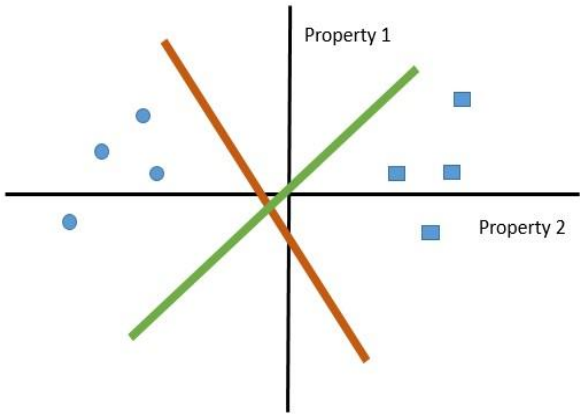

Figure 4: SVM generated hyperplanes [29]

3.4.2. Random Forest (RF)

Random Forest constructs several decision trees to create an ensemble of trees. A single decision tree yields either a very specific model or a simple model [30]. Several decision trees are combined in the Random Forest ensemble ML technique to predict a result. It uses bootstrapped samples of the data with a random subset of the features for each tree to reduce overfitting and increase overall accuracy. It is commonly used for classification and regression tasks. Random Forest builds a number of decision trees (ensemble) and combines the predictions made by each decision tree to increase the model's overall accuracy. The trees in the forest are trained using bootstrapped samples of the data with a random subset of the features at each split, which adds randomness to the model and helps reduce overfitting. The output of a Random Forest is computed by aggregating the predictions of individual decision trees, either by averaging them (for regression) or by taking a majority vote (for classification). Random Forest is a type of bagging (bootstrapped aggregating) algorithm, which reduces the variance in the prediction and improves the stability of the model. Random Forest can handle missing values, and it is not necessary to scale the features.

Finally, the classification of the observation in one category and the other is done by making a count of the trees. Classification of cases is based on a majority vote over the predictions generated by individual decision trees [31]. The Random Forest method's mechanics, as seen in Figure 5, consist of several decision trees cooperating to categorize data points.

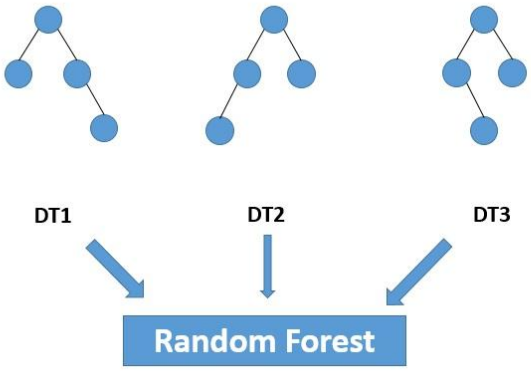

Figure 5: A graphical representation of the mechanics of the random forest technique [29]

3.4.3 Decision Tree

A decision tree is a visual depiction of potential solutions to a decision contingent upon specific variables. It is among the most prevalent supervised learning models. Decision Tree breaks down the complex decisions or problems into smaller and simpler sub-problems. Tree-like structure, it displays the order of decisions and their possible outcomes. Every leaf node is a decision or class label, while every internal node is a test of an attribute. The route that leads from the root node to a leaf node reflects a series of decisions depending on the attributes and their value. Every tree node is a decision point; branches show the possible results of the decision. Every branch's termination denotes a prediction or a final result. Classification or regression issues can be handled using the tree. Fields including data mining, ML, and predictive analytics often employ decision trees to solve issues and generate predictions.

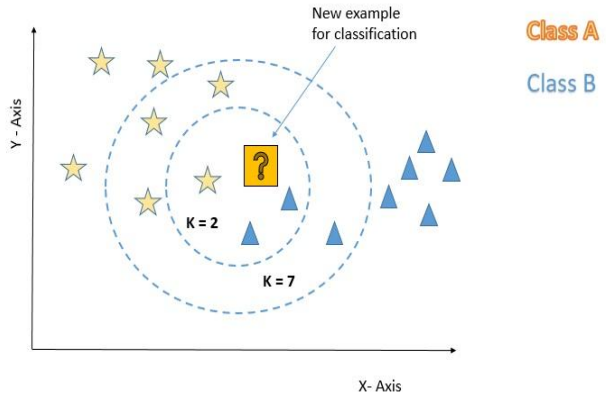

Figure 6: k-NN algorithm Illustration [4]

3.4.4 k-NN

A type of supervised learning called K-Nearest Neighbors (k-NN) is simple, non-parametric, and relies on instance-based reasoning used for running both regression and classification operations. The k-NN method predicts the label of the query instance by finding the k nearest neighbors of that instance and using their class labels (or target values for regression). Usually, Euclidean distance is employed to measure the distance between instances; nonetheless, alternative distance measurements

might also be applied. The selection of k is crucial as it may significantly affect the model's accuracy. A small k value makes the model more noise-sensitive, while a large k value can smooth the decision boundary and lower the model's data-fitting capacity. What makes k-NN a lazy learning algorithm is its lack of prior model construction. It waits instead until a prediction is needed and then constructs the model on the fly. While training is slow, k-NN is rather quick for prediction. Easy to use and with fairly low computational needs, k-NN is a viable option for small datasets or tasks with few features. For huge datasets or high-dimensional feature spaces, however, it can be memory-intensive and computationally costly. Figure 6 shows an illustration of the k-NN method, which classifies data points by finding the closest neighbors.

3.5. Performance Metrics

Performance metrics address the criteria applied for examining the efficacy of ML methodologies. Often, model evaluation in ML is based on several performance measures. These measures allow one to evaluate the model's capacity to accurately predict the target, its false positive and false negative balance, and its target distribution fit. The challenge and the measure that most closely matches the intended solution determine which metric is used. A confusion matrix showing the True Positive, True Negative, False Positive, and False Negative for the predicted and actual classes helps one to evaluate a model's performance. Other measures are specificity, false negative rate, and false positive rate.

$$\text{Accuracy} = \frac{(\text{True Positives} + \text{True Negatives})}{(\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative})} \quad (1)$$

$$\text{Precision} = \frac{\text{True Positives}}{(\text{True Positive} + \text{False Positive})} \quad (2)$$

$$\text{Recall} = \frac{\text{True Positives}}{(\text{True Positives} + \text{False Negatives})} \quad (3)$$

$$\text{F1 Score} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (4)$$

$$\text{Specificity} = \frac{\text{True Negative}}{(\text{True Negative} + \text{False Positive})} \quad (5)$$

$$\text{FPR} = \frac{\text{False Positive}}{(\text{True Negative} + \text{False Positive})} \quad (6)$$

$$\text{FNR} = \frac{\text{False Negative}}{(\text{False Negative} + \text{True Positive})} \quad (7)$$

4. Results and Discussion

This study proposes a comparison of four ML algorithms—Random Forest, k-NN, Decision Tree, and SVM—run on an Intel Core i5 computer with 8 GB RAM. Python-based open source ML libraries like numpy, pandas, and Scikit-learn have been utilized by us. Jupyter Notebook, an open-source web application, is employed to execute the program. The classifier was assessed by testing it with the Test-train split method. Here we selected the split ratio as 70/30. The split ratio of 398

training set observations to 171 testing set observations was obtained out of a total of 569 observations.

4.1. Result Analysis of SVM

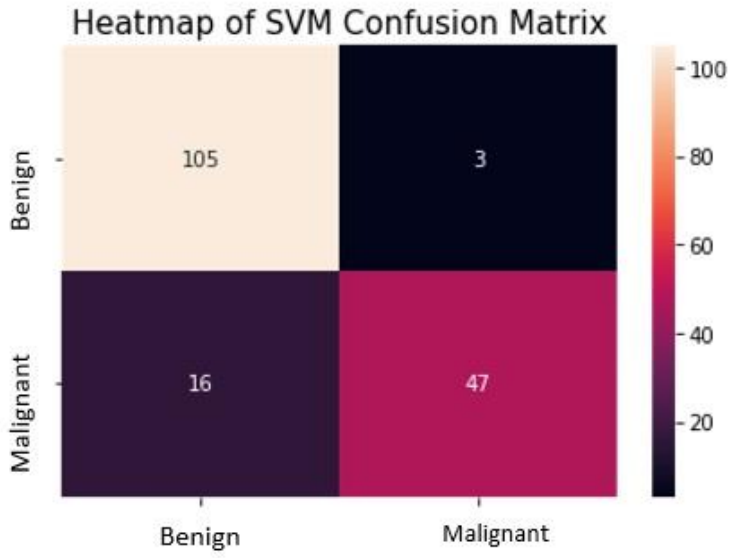

Figure 7: SVM confusion matrix (Benign indicates the 0 and Malignant indicates the 1)

The SVM model for breast cancer detection has a higher precision for detecting malignant tumors compared to benign tumors. Tables 1 and Table 2 provide summaries of the findings. With a precision of 97.22%, the SVM model showed great accuracy in identifying malignant tumors, hence demonstrating a high reliability in its positive predictions. The results from the SVM model for breast cancer detection show good precision for detecting malignant tumors at 97.22%, indicating that when the model predicts a tumor as malignant, it is likely to be correct. However, the recall rate for malignant tumors is lower at 86.777%, which suggests that the model may have missed some malignant tumors. For benign tumors, the model has lower precision at 74.603%, meaning that it may have predicted some benign tumors as malignant. However, the recall rate is higher at 94%, indicating that the model was effective at identifying most benign tumors. The overall F1-score of the model is 91.703% for malignant tumors and 83.185% for benign tumors. Furthermore, the specificity (97.22%) demonstrates that the model can accurately recognize a significant percentage of true negative cases (i.e., correctly identifying cases that are not breast cancer), and the false positive rate (6%) suggests that the model is still making some incorrect classifications of cases as breast cancer when they are not. The false negative rate (13.22%), indicates that the model missed a small but significant number of malignant cases. These metrics are detailed in Table 1 and Table 2. Figure 7 shows the generated heatmap of the SVM confusion matrix. The SVM model's Receiver Operating Characteristic (ROC) curve, displayed in Figure 8, also offers a visual depiction of the model's performance stressing its capacity to tell benign from malignant tumors.

Table 1 SVM performance measurement metrics

Tumor Type	Precision	Recall	F1-score
Malignant	97.22%	86.777%	91.703%
Benign	74.603%	94%	83.185%

Table 2 Performance Measurement Indices of SVM

Specificity	94%
False positive rate	6%
False negative rate	13.22%

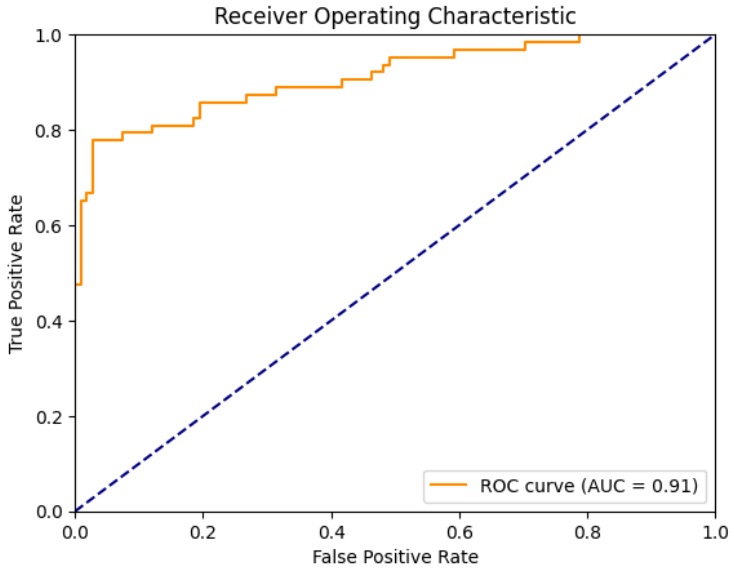

Figure 8: ROC curve of SVM

4.2. Result Analysis of Random Forest

The Random Forest (RF) model was assessed by employing many performance criteria to determine its efficacy in breast cancer detection. With benign tumors marked by '0' and malignant tumors marked by '1', the RF model's confusion matrix shown in Figure 9 offers a comprehensive breakdown of the model's predictions. The RF model's performance criteria are presented in Table 3 and Table 4. The model achieved a high precision of 95.37% for detecting malignant tumors, indicating that the model can accurately classify tumors as malignant with a low rate of false positives. The model successfully identified a large fraction of the actual malignant tumors in the dataset, as seen by the recall value of 92.793%. For detecting benign tumors, the model achieved a precision of 87.302%, indicating that it accurately identified a high proportion of benign tumors with a low rate of false positives. With a recall value of 91.667%, the model demonstrated a strong ability to identify actual benign tumors in the dataset. A satisfactory balance between precision and recall for both malignant and benign tumor classes was reflected by the F1-score values of 94.063% and 89.431%, respectively.

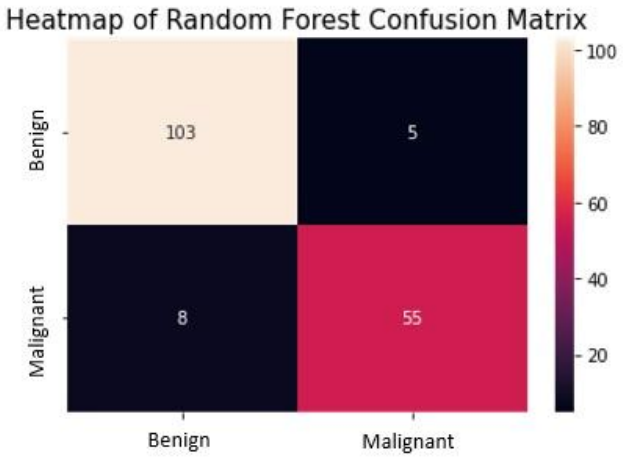

Figure 9: RF confusion matrix (Benign indicates the 0 and Malignant indicates the 1)

Table 3 Performance Measurement Indices of RF

Tumor Type	Precision	Recall	F1-score
Malignant	95.37%	92.793%	94.063%
Benign	87.302%	91.667%	89.431%

Table 4 Performance Measurement Indices of RF

Specificity	91.67%
False positive rate	8.33%
False negative rate	7.20%

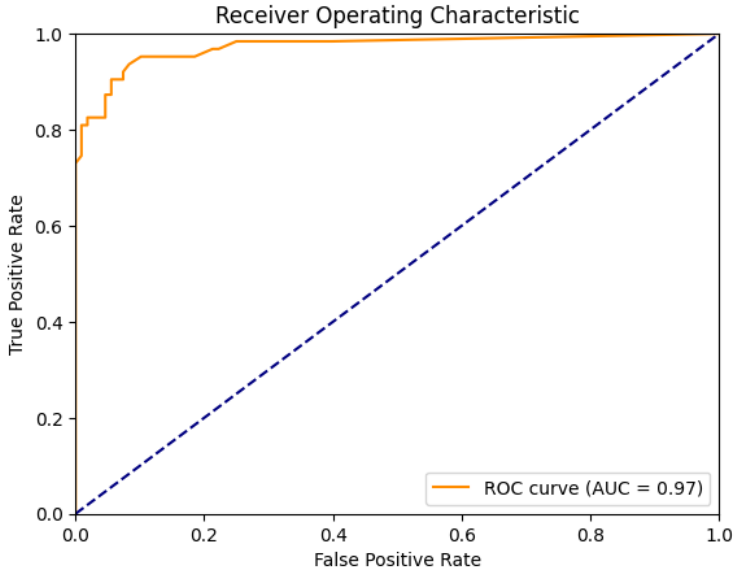

Figure 10: ROC curve of RF

The model's specificity of 91.67% is high, indicating that it correctly identified a high proportion of actual negative cases (benign tumors) in the dataset. Nonetheless, the 8.33% false positive rate suggests that the model's capacity to correctly identify negative cases requires further work. The false negative rate of 7.20% indicates that the model missed a small proportion of actual malignant cases in the dataset, indicating that the model's sensitivity has the potential to be further enhanced. Figure 10 shows the Random Forest model's ROC curve, showing its ability to identify benign and malignant tumors.

4.3. Result Analysis of Decision Tree

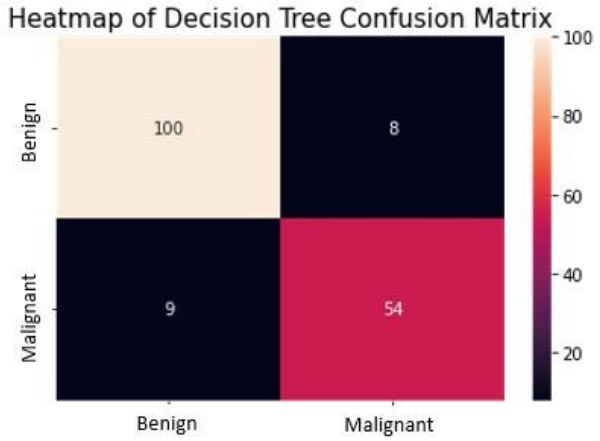

Figure 11: Decision Tree confusion matrix (Benign indicates the 0 and Malignant indicates the 1)

Table 5. Performance Measurement Indices of Decision Tree

Tumor Type	Precision	Recall	F1-score
Malignant	92.593%	91.743%	92.166%
Benign	85.714%	87.097%	86.399%

Table 6. Performance Measurement Indices of Decision Tree

Specificity	87.09%
False positive rate	12.90%
False negative rate	8.25%

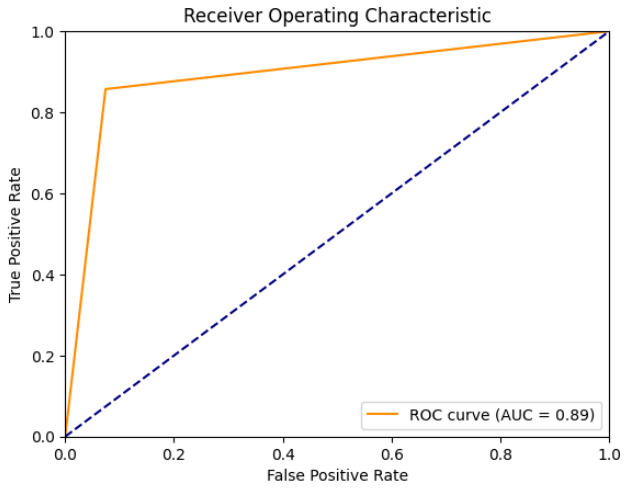

Figure 12: ROC curve of Decision Tree

The Decision Tree model's breast cancer detection ability was assessed using various criteria. Figure 11 illustrates the error matrix of the Decision Tree model, designating benign tumors as '0' and malignant tumors as '1'. Table 5 and Table 6 enumerate the performance metrics of the Decision Tree. The classifier correctly identified the majority of the malignant tumors with a high level of precision of 92.593% while scoring 91.743% recall and 92.166% F1-score. For the benign class, 85.714% precision, 87.097% recall, and 86.399% F1-score are achieved, which indicates that the classifier correctly identified most of the benign tumors with a relatively lower level of precision compared to the malignant tumors. The specificity of 87.09% and false positive rate of 12.90% indicate that the classifier correctly identified most of the benign tumors but also misclassified some of the malignant tumors as benign. The false negative rate of 8.25% suggests that the classifier missed some of the malignant tumors, which could potentially lead to serious consequences in a real-world scenario. The ROC curve of the Decision Tree model in Figure 12 reveals its ability to distinguish between malignant and benign tumors.

4.4. Result Analysis of k-NN

Various measures were used to gauge the breast cancer detection ability of the k-Nearest Neighbors (k-NN) model. With benign tumors as '0' and malignant tumors as '1', Figure 13 displays the confusion matrix indicating the predicted outcomes of the model. Table 7 and Table 8 list key performance indicators (KPI). The k-NN classifier achieved 96.296% precision and 88.889% recall for malignant tumors while scoring a precision of 79.365% and a recall of 92.593% for benign tumors. The F1-score for malignant tumors was 92.444%, and for benign tumors, it was 85.47%. The k-NN classifier achieved a specificity of 92.59%, indicating that it correctly identified 92.59% of the benign tumors. The false positive rate was 7.40%, indicating that the k-NN classifier misclassified 7.40% of the benign tumors as malignant. The false negative rate was 11.11%, which means that 11.11% of the malignant tumors were incorrectly classified as benign. Figure 14 illustrates the ROC curve, emphasizing the ability of the model to tell the difference between malignant and benign tumors.

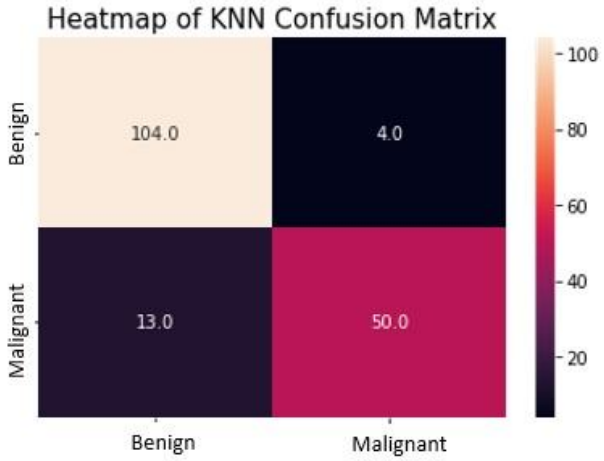

Figure 13: k-NN confusion matrix (Benign indicates the 0 and Malignant indicates the 1)

Table 7. Performance Measurement Indices of k-NN

Tumor Type	Precision	Recall	F1-score
Malignant	96.296%	88.889%	92.444%
Benign	79.365%	92.593%	85.47%

Table 8. Performance Measurement Indices of k-NN

Specificity	92.59%
False positive rate	7.40%
False negative rate	11.11%

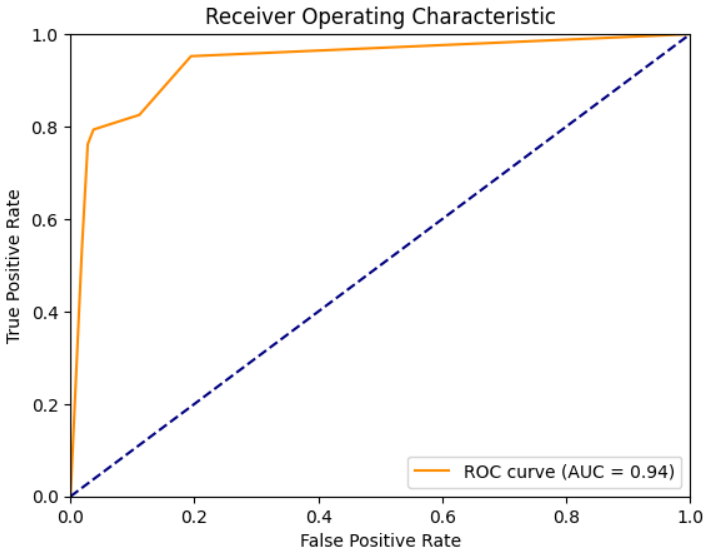

Figure 14: ROC curve of k-NN

The results presented in Table 9 show that SVM has the best Precision score and Specificity performance measure but Random Forest has the best recall, F1 score, and accuracy performance measure over k-NN, Decision Tree, and SVM. SVM showed the least False Positive Rate among the classifiers while Random Forest has the least False Negative Rate.

Table 9: Performances Measure Indices

Model Performance during Testing Phase (in %)				
	SVM	**Decision Tree**	**k-NN**	**Random Forest**
Accuracy	88.89	90.64	90.06	92.40
Precision	94	88.52	92.59	91.67
Recall	74.6	85.71	79.37	87.30
F1 score	83.19	87.10	85.47	89.43
Specificity	97.22	93.52	96.30	95.37
FPR	2.78	6.48	3.70	4.63
FNR	25.40	14.29	20.63	12.70

5. Conclusion

In this study, we evaluated various machine-learning techniques for breast cancer diagnosis utilizing the Wisconsin Breast Cancer dataset. We examined the methods for categorizing tumors as either malignant or benign depending on key diagnostic metrics. All models performed strongly, with each showing particular strengths in different evaluation criteria. Random Forest was the most effective classifier overall, with 92.40% accuracy, 87.30% recall, and an 89.43% F1 score. Results show Random Forest has the best balance between correctly identifying malignant tumors while minimizing false classifications. In reducing false positives, the SVM attained 97.22% specificity and 94% precision. Both decision tree and k-NN show above 90% accuracy, showing diagnostic reliability.

The findings have important implications for clinical practice. Automated tumor classification using ML reduces inter-observer variability in mammography analysis. The comparison results clearly show which algorithms to choose based on clinical needs; while SVM is best for confidently detecting cancer, the study demonstrates that the Random Forest has superior overall accuracy. Future studies could integrate these algorithms with ensemble or deep learning techniques to enhance performance. Diverse data and imaging techniques may improve efficacy. This research employs ML to enable the timely detection of breast cancer.

Funding source

None

Conflict of Interest

The authors declare no conflict of interest.

References

[1] World Health Organization (WHO), "Breast cancer," World Health Organization (WHO). Accessed: Dec. 21, 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[2] O. Ubrurhe, N. Houlden, and P. S. Excell, "A Review of Energy Efficiency in Wireless Body Area/Sensor Networks, With Emphasis on MAC Protocol," *Annals of Emerging Technologies in Computing*, vol. 4, no. 1, pp. 1–7, Jan. 2020, doi: 10.33166/AETiC.2020.01.001.

[3] M. Piñeros, A. Znaor, L. Mery, and F. Bray, "A Global Cancer Surveillance Framework Within Noncommunicable Disease Surveillance: Making the Case for Population-Based Cancer Registries," *Epidemiol Rev*, vol. 39, no. 1, pp. 161–169, Jan. 2017, doi: 10.1093/epirev/mxx003.

[4] S. J. Nechuta *et al.*, "The After Breast Cancer Pooling Project: rationale, methodology, and breast cancer survivor characteristics," *Cancer Causes & Control*, vol. 22, no. 9, pp. 1319–1331, Sep. 2011, doi: 10.1007/s10552-011-9805-9.

[5] M. F. Ak, "A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications," *Healthcare*, vol. 8, no. 2, p. 111, Apr. 2020, doi: 10.3390/healthcare8020111.

[6] A. Jemal *et al.*, "Cancer Statistics, 2008," *CA Cancer J Clin*, vol. 58, no. 2, pp. 71–96, Jan. 2008, doi: 10.3322/CA.2007.0010.

[7] A. A. Ardakani, A. Gharbali, and A. Mohammadi, "Classification of Breast Tumors Using Sonographic Texture Analysis," *Journal of Ultrasound in Medicine*, vol. 34, no. 2, pp. 225–231, Feb. 2015, doi: 10.7863/ultra.34.2.225.

[8] B. L. Sprague *et al.*, "Variation in Mammographic Breast Density Assessments Among Radiologists in Clinical Practice," *Ann Intern Med*, vol. 165, no. 7, p. 457, Oct. 2016, doi: 10.7326/M15-2934.

[9] P. E. Freer, "Mammographic Breast Density: Impact on Breast Cancer Risk and Implications for Screening," *RadioGraphics*, vol. 35, no. 2, pp. 302–315, Mar. 2015, doi: 10.1148/rg.352140106.

[10] T. M. Kolb, J. Lichy, and J. H. Newhouse, "Comparison of the Performance of Screening Mammography, Physical Examination, and Breast US and Evaluation of Factors that Influence Them: An Analysis of 27,825 Patient Evaluations," *Radiology*, vol. 225, no. 1, pp. 165–175, Oct. 2002, doi: 10.1148/radiol.2251011667.

[11] H. D. Cheng, X. J. Shi, R. Min, L. M. Hu, X. P. Cai, and H. N. Du, "Approaches for automated detection and classification of masses in mammograms," *Pattern Recognit*, vol. 39, no. 4, pp. 646–668, Apr. 2006, doi: 10.1016/j.patcog.2005.07.006.

[12] F. Bray, P. McCarron, and D. M. Parkin, "The changing global patterns of female breast cancer incidence and mortality," *Breast Cancer Research*, vol. 6, no. 6, p. 229, Dec. 2004, doi: 10.1186/bcr932.

[13] R. L. Birdwell, D. M. Ikeda, K. F. O'Shaughnessy, and E. A. Sickles, "Mammographic Characteristics of 115 Missed Cancers Later Detected with Screening Mammography and the Potential Utility of Computer-aided Detection," *Radiology*, vol. 219, no. 1, pp. 192–202, Apr. 2001, doi: 10.1148/radiology.219.1.r01ap16192.

[14] P. Skaane and K. Engedal, "Analysis of sonographic features in the differentiation of fibroadenoma and invasive ductal carcinoma.," *American Journal of Roentgenology*, vol. 170, no. 1, pp. 109–114, Jan. 1998, doi: 10.2214/ajr.170.1.9423610.

[15] K. Doi, "Computer-aided diagnosis: potential usefulness in diagnostic radiology and telemedicine," in *Proceedings of the National Forum: Military Telemedicine On-Line Today Research, Practice, and Opportunities*, IEEE Comput. Soc. Press, pp. 9–13. doi: 10.1109/MTOL.1995.504521.

[16] N. GULER, E. UBEYLI, and I. GULER, "Recurrent neural networks employing Lyapunov exponents for EEG signals classification," *Expert Syst Appl*, vol. 29, no. 3, pp. 506–514, Oct. 2005, doi: 10.1016/j.eswa.2005.04.011.

[17] A. Osareh and B. Shadgar, "Machine learning techniques to diagnose breast cancer," in *2010 5th International Symposium on Health Informatics and Bioinformatics*, IEEE, 2010, pp. 114–120. doi: 10.1109/HIBIT.2010.5478895.

[18] D. Michie, D. J. Spiegelhalter, and C. C. Taylor, "Machine Learning, Neural, and Statistical Classification," 1994.

[19] Shen-Chuan Tai, Zih-Siou Chen, and Wei-Ting Tsai, "An Automatic Mass Detection System in Mammograms Based on Complex Texture Features," *IEEE J Biomed Health Inform*, vol. 18, no. 2, pp. 618–627, Mar. 2014, doi: 10.1109/JBHI.2013.2279097.

[20] J. Nagi, S. Abdul Kareem, F. Nagi, and S. Khaleel Ahmed, "Automated breast profile segmentation for ROI detection using digital mammograms," in *2010 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, IEEE, Nov. 2010, pp. 87–92. doi: 10.1109/IECBES.2010.5742205.

[21] S. M. Butler, G. I. Webb, and R. A. Lewis, "A Case Study in Feature Invention for Breast Cancer Diagnosis Using X-Ray Scatter Images," in *Lecture Notes in Computer Science AI 2003: Advances in Artificial Intelligence*, vol. 2903, 2003, pp. 677–685. doi: 10.1007/978-3-540-24581-0_58.

[22] A. Qayyum and A. Basit, "Automatic breast segmentation and cancer detection via SVM in mammograms," in *2016 International Conference on Emerging Technologies (ICET)*, IEEE, Oct. 2016, pp. 1–6. doi: 10.1109/ICET.2016.7813261.

[23] T. T. Htay and S. S. Maung, "Early Stage Breast Cancer Detection System using GLCM feature extraction and K-Nearest Neighbor (k-NN) on Mammography image," in *2018 18th International Symposium on Communications and Information Technologies (ISCIT)*, IEEE, Sep. 2018, pp. 171–175. doi: 10.1109/ISCIT.2018.8587920.

[24] B. Dai, R.-C. Chen, S.-Z. Zhu, and W.-W. Zhang, "Using Random Forest Algorithm for Breast Cancer Diagnosis," in *2018 International Symposium on Computer, Consumer and Control (IS3C)*, IEEE, Dec. 2018, pp. 449–452. doi: 10.1109/IS3C.2018.00119.

[25] S. Hamed, A. Mesleh, and A. Arabiyyat, "Breast Cancer Detection Using Machine Learning Algorithms," *International Journal of Computer Science and Mobile Computing*, vol. 10, no. 11, pp. 4–11, Nov. 2021, doi: 10.47760/ijcsmc.2021.v10i11.002.

[26] G. Williams, "Descriptive and Predictive Analytics," in *Data Mining with Rattle and R*, New York, NY: Springer New York, 2011, pp. 171–177. doi: 10.1007/978-1-4419-9890-3_8.

[27] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput Struct Biotechnol J*, vol. 13, pp. 8–17, 2015, doi: 10.1016/j.csbj.2014.11.005.

[28] T. J. Cleophas and A. H. Zwinderman, *Machine Learning in Medicine*. Dordrecht: Springer Netherlands, 2013. doi: 10.1007/978-94-007-5824-7.

[29] D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," in *2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA)*, IEEE, Dec. 2016, pp. 1–4. doi: 10.1109/ICEDSA.2016.7818560.

[30] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artif Intell Med*, vol. 23, no. 1, pp. 89–109, Aug. 2001, doi: 10.1016/S0933-3657(01)00077-X.

[31] Y. Yasui and X. Wang, "Statistical Learning from a Regression Perspective by BERK, R. A.," *Biometrics*, vol. 65, no. 4, pp. 1309–1310, Dec. 2009, doi: 10.1111/j.1541-0420.2009.01343_5.x.

[32] A. Verma, S.N. Shivhare, S. P. Singh et al., "Comprehensive Review on MRI-Based Brain Tumor Segmentation: A Comparative Study from 2017 Onwards". *Arch Computat Methods Eng*, vol. 31, pp. 4805–4851, 2024. Doi: 10.1007/s11831-024-10128-0.