Early Detection of Breast Cancer: Comparative Analysis of Machine Learning and Deep Learning Algorithms

Jashanpreet Singh, Syed Raahat, Aditya Sharma, Sourabh Dellu, Saksham Arora

***Abstract:** Breast cancer classification remains a critical challenge in medical diagnostics due to the imbalanced nature of available datasets, where the minority (cancerous malignant) class is often overshadowed by the majority (benign) class. This study proposes a hybrid model based on logistic regression, enhanced with class balancing techniques and ant search optimization, to improve the identification of the malignant class. The model is compared with SVM, Random Forest, and K-Nearest Neighbors (KNN) across three stages: prediction before diagnosis, at diagnosis and therapy, and post-treatment outcomes. The experiments, conducted on the Jupyter platform using the Wisconsin breast cancer dataset, demonstrate that the hybrid model achieves a high accuracy of 92.98%, significantly reducing false negatives. The study highlights the strengths of logistic regression in providing interpretable results, crucial for clinical decision-making, especially when compared to more complex models like Artificial Neural Networks (ANN). This research offers a reliable and accurate tool for early breast cancer detection and prognosis, contributing to ongoing efforts to enhance patient outcomes through the integration of hybrid machine learning models in medical diagnostics.*

***Keywords:** Breast Cancer, Wisconsin Breast Cancer Diagnosis (WBCD) Dataset, Machine Learning, Naïve Bayes Algorithm (NB), Support Vector Machine, Random Forests, Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), Logistic Regression (LR)*

I. INTRODUCTION

Accurately diagnosing certain crucial information is a major problem in the fields of bioinformatics and medical science [1]. In the field of medicine, diagnosing an illness is a demanding and complex task. Numerous diagnostic centres, hospitals, research facilities, and websites all have access to vast amounts of medical diagnosis data [2]. A variety of data mining and machine learning methods are being utilised to predict breast cancer [3]. One of the crucial tasks is determining the best and most appropriate algorithm for breast cancer prediction. Malignant tumours are the source of breast cancer when cell development becomes unchecked [4]. The three procedures that make up "The Gold Standard" approach—clinical examination, radiographic imaging, and pathology test—are the foundation of conventional cancer detection methods [5]. The latest machine learning approaches and algorithms are based on model design, whereas the traditional method, which is based on regression process, detects the existence of cancer. The model's training and testing phases yield good expected results and are designed to forecast unknown data [6]. Preprocessing, feature selection or extraction, and classification are the three primary strategies around which the machine learning process is built [7]. The primary component of machine learning is feature extraction, which aids in cancer diagnosis and prognosis by elucidating the cancer's progression from benign to malignant tumor.

A. Breasts

At every stage of life, including puberty, adolescence, adulthood, and menopause, breasts are vulnerable to numerous illnesses. The most frequent pathological alterations of the breast include inflammatory processes, traumatic injury, and benign and malignant tumours [8]. Under normal circumstances, cells divide in a typical and regulated manner; however, if this process is uncontrolled, the cells continue to divide quickly and produce swelling known as a tumour [9].

i. Benign

One significant risk factor for breast cancer, which can occur in either breast, is benign breast disease [10]. It encompasses a spectrum of histologic entities, usually categorized as atypical hyperplasias, proliferative lesions without atypia, and no proliferative lesions. Proliferative or atypical lesions are associated with an increased risk of breast cancer [11]. Since more women are being diagnosed with benign breast illness due to the increased use of mammography, it is critical to have precise risk estimations for those women [12].

Manuscript received on 07 November 2024 | First Revised Manuscript received on 14 November 2024 | Second Revised Manuscript received on 01 December 2024 | Manuscript Accepted on 15 January 2025 | Manuscript published on 30 January 2025.

*Correspondence Author(s)

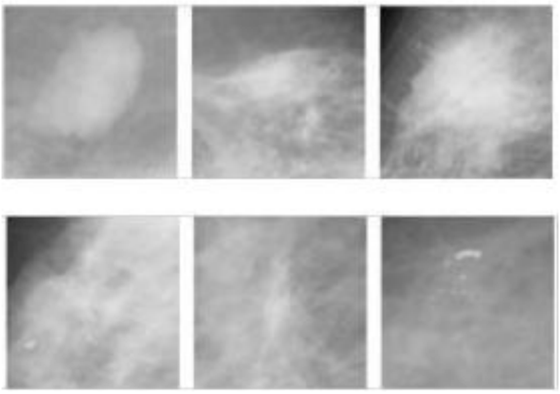
Jashanpreet Singh*, Department of Computer Science and Engineering, Chandigarh University, Chandigarh (Punjab), India. Email ID: 22BCS11724@cuchd.in

Syed Raahat, Department of Computer Science and Engineering, Chandigarh University, Chandigarh (Punjab), India. Email ID: 22BCS11708@cuchd.in

Aditya Sharma, Department of Computer Science and Engineering, Chandigarh University, Chandigarh (Punjab), India. Email ID: 22BCS11741@cuchd.in

Sourabh Dellu, Department of Computer Science and Engineering, Chandigarh University, Chandigarh (Punjab), India. Email ID: 22BCS11694@cuchd.in

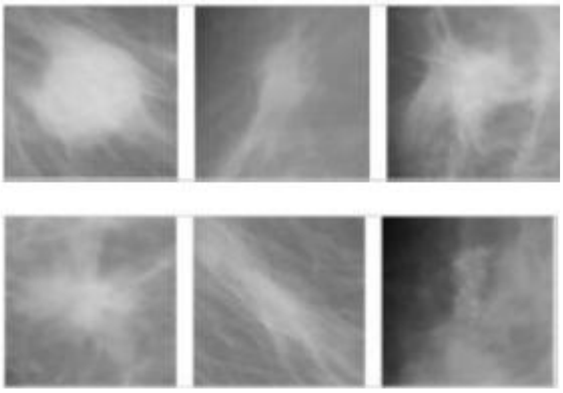
Saksham Arora, Department of Computer Science and Engineering, Chandigarh University, Chandigarh (Punjab), India. Email ID: 22BCS11870@cuchd.in



[Fig.1: Different Types of Benign Breast Tumors]

ii. Malignant

Cancer cells that have the capacity to spread outside of their original location make up malignant tumours. When cancer cells proliferate, the surrounding tissue is frequently destroyed. It is not uncommon for the cancer cells to split out from the main cancer and go to different parts of the body via blood or lymphatic channels. New tumours known as metastases are created when they reproduce in a different part of the body.



[Fig.2: Different Types of Malignant Breast Tumors]

B. Diagnosis of Breast Cancer

i. Breast Self-Examination

Self-examination of the breasts has been widely advocated and condemned. According to Haagensen, teaching women how to examine themselves was more crucial than teaching doctors how to examine their breasts [13]. Both the formal physical act and a heightened general awareness of breast anatomy and symptoms can be included in breast self-examination performance. Early clinical and pathological stages of breast cancer can be identified with breast self-examination prior to its discovery. Frequent breast self-examination was associated with a considerably higher chance of breast cancer detection in women [14].

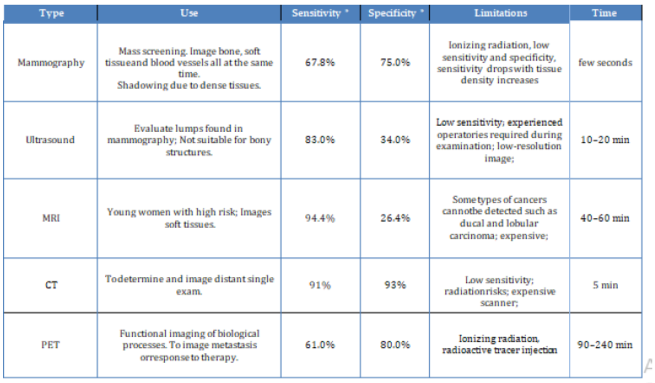
ii. Clinical Breast Examination

Over the years, a number of strategies have been employed to increase breast cancer survival by early identification [15]. To improve early detection, mammography, breast self-examination, and clinical breast examination have been combined into various screening programs worldwide. A professional in the health care delivery system often performs a clinical breast examination physically [16]. It is typically performed as part of a woman's routine medical examination, which may help detect breast cancer and other breast problems. Traditional breast screening techniques are intended to be improved upon by other techniques, such as mammography [17]. Furthermore, in today's society, using X-rays to examine different bodily sections has grown in popularity. Another type of X-ray used to examine the breast is mammography, which shows details of the tissue inside the breast using a tube voltage that ranges from 25 kVp to 32 kVp [18]. This makes it possible to identify and diagnose the illness early. Any anomaly in the breast, including the development of malignant cells, is detected by it.

Further tests are required to confirm the presence of the tumour if the mammography shows an anomaly. One or more tests are conducted, depending on a number of variables. Mammogram abnormalities include masses, microcalcifications, and macrocalcifications.

Macrocalcifications are large deposits of calcium, often present in degenerative changes in the breast. Macrocalcifications usually occur in women older than 50 years.

Table 1: Conventional Breast Screening Methods and their Limitations



Type	Use	Sensitivity *	Specificity *	Limitations	Time
Mammography	Mass screening. Image bone, soft tissueand blood vessels all at the same time. Shadowing due to dense tissues.	67.8%	75.0%	Ionizing radiation, low sensitivity and specificity, sensitivity dropswith tissue density increases	few seconds
Ultrasound	Evaluate lumps found in mammography; Not suitable for bony structures.	83.0%	34.0%	Low sensitivity; experienced operatories required during examination; low-resolution image;	10–20 min
MRI	Young women with high risk; Images soft tissues.	94.4%	26.4%	Some types of cancers cannotbe detected such as ducal and lobular carcinoma; expensive;	40–60 min
CT	Todetermine and image distant single exam.	91%	93%	Low sensitivity; radiationrisks; expensive scanner;	5 min
PET	Functional imaging of biological processes. To image metastasis orresponse to therapy.	61.0%	80.0%	Ionizing radiation, radioactive tracer injection	90–240 min

Mammography is supplemented by ultrasound, Magnetic Resonance Imaging (MRI) [19], and biopsy. High frequency ultrasonic vibrations that travel through the breast are used in the breast ultrasound procedure to examine the tissue of the breast. When the lump is discovered by mammography or palpative examination, an ultrasound examination of the breast is conducted. This diagnostic technique is particularly well-suited for identifying benign breast tumors and cysts, particularly in younger women whose glandular tissue is still developing. This diagnostic technique can identify if the lump is a solid tumor, a complicated cyst composed of tissue and liquid, or a cyst filled with liquid. Estimating the palpable mass in women under 35 and estimating the mass that cannot be palpated but is visible on a mammogram are the two most advantageous applications of ultrasonic imaging.

Magnetic Resonance Imaging (MRI) is carried out following the clinical [20], mammography, and ultrasound tests if it is not possible to accurately identify whether the breast alteration is benign or malignant. Only in extreme circumstances is this diagnostic technique employed. To create a detailed image of the body, MRI scanners use radio waves, electric field gradients, and powerful magnetic fields. Verification by cytology or pathohistopathology is necessary before a final diagnosis is adopted. If a mammogram shows a positive result or contains an anomaly that cannot be definitively identified, or if an ultrasound examination confirms that the mass is solid or complex, a biopsy is conducted. Taking a tissue sample for microscopic examination is known as biopsy. Breast biopsies can be performed in a number of ways. The simplest method, if the mass is palpable, is to use a thin needle to pierce it and then use a syringe to extract a portion of the analytical material. A needle is used for the biopsy under ultrasound guidance if the lump cannot be felt.

Positron Emission Tomography (PET) F-fluorodeoxyglucose imaging helps physicians determine where a tumor is located in the human body. Its foundation is the identification of radio labeled tracers unique to malignancy [21].

II. REVIEW OF RELATED LITREATURE

As medical research has progressed, numerous innovative technologies for detecting breast cancer have emerged. Data mining and machine learning techniques enable us to diagnose and predict different types of breast cancer [22]. By employing data mining methods such as clustering, regression, and classification, we can extract valuable insights about patients with breast cancer. These algorithms utilize training datasets, which help us assess the probability of identifying various forms of breast cancer [23]. WBCD and another breast cancer dataset that was obtained from the UCI library were used to test the modified decision tree technique that Kapil and Rana introduced an enhanced decision tree that improved weight distribution [24]. They found that by using the Chi-square test, they could rank each characteristic and keep only the most significant features for this classification task. Their proposed method achieved approximately 99% accuracy on the WBCD dataset and between 85% and 90% accuracy on the breast cancer dataset. In their study, Yue et al [25]. provided comprehensive evaluations of SVM, K- NNs, ANNs, and Decision Tree methods for breast cancer prediction, utilizing the benchmark Wisconsin Breast Cancer Diagnosis (WBCD) dataset. The authors noted that the best results came from integrating deep belief networks (DBNs) with ANN architecture (DBNs-ANNs). While a two-step clustering method combined with SVM yielded a classification accuracy of 99.10%, this architecture reached an impressive 99.68% accuracy. They also explored an ensemble technique that employed a voting mechanism to combine SVM, Naive Bayes, and J48, resulting in an accuracy of 97.13%.

In their comparative research using Tree Augmented Naïve Bayes (TAN), Boosted Augmented Naive Bayes (BAN), and Bayes Belief Network (BBN), Banu and Subramanian emphasized the application of Naive Bayes algorithms in predicting breast cancer [26]. The models were executed using SAS- EM (Statistical Analytical Software Enterprise Miner) and utilized the well-known WBCD dataset. They discovered that BBN, BAN, and TAN achieved accuracy rates of 91.7%, 91.7%, and 94.11%, respectively, when employing gradient boosting. Consequently, TAN emerged as the most effective classifier among the Naïve Bayes methods for this dataset, according to their findings. In a separate study on the WBCD dataset, Chaurasia et al [27]. utilized the RBF network, J48 Decision Tree, and Naive Bayes algorithms. They used the Waikato giving the highest average Correct Classification Rate (CCR) values for two and three predictors, respectively, of 0.897 and 0.972. Additionally, it takes a lot less time and yields the lowest average squared classification error (ASCE) and minimal description length (MDL) values.

A. System Analysis

i. Description of WBCD Dataset

The UCI machine learning repository, in partnership with Kaggle, provided the breast cancer dataset [30]. This dataset contains 569 cases, all of which are either benign or malignant. Of these, 212 (37.25%) are malignant and 357 (62.74%) are benign. The dataset's class is divided into two categories: benign cases are represented by a class of 0 and malignant cases by a class of 1.

ii. Data Preprocessing

For testing the proposed methodology, Wisconsin Diagnostic Breast Cancer was used. It comprises 32 features (ID#, 30 descriptors, and 1 decision attribute). The total record count is 569 along with 357 benign cases and 212 malignant. The features are computed on a digitized image of a fine needle aspirate (FNA) of a breast mass. The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were calculated for each image, which yields 30 features. Table 2 shows some descriptive statistics about the used features.

Table 2: Used Variables for Breast Cancer Diagnosis

Predictor	Mean	SD	Min.	Max.
Radius	14.1	3.52	6.98	28.11
Texture	19.2	4.30	9.71	39.28
Perimeter	91.9	24.30	43.79	188.50
Area	654.8	351.9	143.50	2501.0
Smoothness	0.10	0.01	0.05	0.16
Compactness	0.10	0.05	0.02	0.35
Concavity	0.09	0.08	0.00	0.43
Concave Points	0.05	0.04	0.00	0.20
Symmetry	0.18	0.03	0.11	0.30

Environment for Knowledge Analysis (WEKA) version 3.6.9 as their analytical tool. The accuracy of 97.36% for Naive Bayes surpassed the accuracy rates of 96.77% and 93.41% for the RBF network and J48 Decision Tree, respectively.

Azar et al. introduced a decision tree variant- based method for predicting breast cancer [28]. This approach utilizes decision tree forest (DTF), boosted decision tree (BDT), and single decision tree (SDT) models. After training and testing the dataset, a decision is reached. The findings indicated that during the training phase, SDT and BDT attained accuracies of 97.07% and 98.83%, respectively, highlighting BDT's superior performance over SDT. In the testing phase, the decision tree forest achieved an accuracy of 97.51%, while SDT reached 95.75%. The dataset was trained using ten-fold cross-validation. The authors of the study [29]. provide an example of a breast cancer detection method. This article discusses the research conducted to diagnose the disease utilizing local linear wavelet neural networks (LLWNN) and recursive least squares (RLS) to improve system performance. With a few computation times, the LLWNN-RLS is

iii. Theoretical Considerations

Supervised and unsupervised learning are the two primary categories of the learning process in machine learning algorithms. In supervised learning, the machine is trained on a labeled dataset, which helps it produce the correct output. In contrast, unsupervised learning does not use labeled data or predefined outcomes, making its objectives more challenging to achieve. The two most common techniques within supervised learning are regression and classification. In classification, the target variable for prediction is discrete, while in regression, it is continuous.

iv. Support Vector Machine

The support vector machine (SVM) is based on the concept of the maximal margin classifier [31], which is a fundamental type of classifier. A hyperplane in an n-dimensional space corresponds to this maximal margin classifier. The hyperplane defines a subspace of (n − 1) dimensions, which does not necessarily need to pass through the origin. Visualizing a hyperplane in higher dimensions can be difficult, so we often work with this (n − 1) dimensional subspace. If a separating hyperplane exists, creating an SVM classifier becomes straightforward. However, when the categories in the dataset cannot be separated by a hyperplane, we need to expand the feature space using functions like the Gaussian radial basis function (RBF), sigmoid function, or polynomial functions of various degrees. The following equation represents the hyperplane used in p-dimensions:

$$\beta o + \beta 4X4 + \beta ,X, + \cdots + \beta pXp \ = 0$$

where X_1, X_2,…, and X_p represent the data points in a p-dimensional sample space, and β_0, β_1, β_2,…, and β_p denote the hypothetical values.

v. K-Nearest Neighbors

Pattern recognition and grouping are two applications of the K-nearest neighbour algorithm. Predictive analysis makes extensive use of it. The K-NN method finds the closest existing data points to new data when it arrives. The time between data points may be sufficiently influenced by any attributes that can vary widely [32]. During the training phase, the feature vectors and class labels are saved. The representation of the data samples in a metric space is assumed by K- NNs. During the classification phase, the quantity is first described by the K training sample's most regular neighbours. The computation will then find the new data sample's K adjacent neighbours. The calculation

$$\text{Minkowski Distance: Dist}(x, y) = (\sum n\ |x; - y;|p)p^{1}$$

The Manhattan distance is used when p = 1, the Euclidean distance when p = 2, and the Chebyshev distance when p = ∞. Among these, the Euclidean distance is one of the most commonly used methods globally. Once we assess how much each of these K neighbors contributes, the calculation will categorize the new information point based on the most significant input.

vi. Random Forests

A highly effective supervised classification tool is the Random Forest classifier [33]. This ensemble technique can be viewed as a variation of the nearest neighbor predictor. Ensemble learning involves the deliberate development and integration of statistical techniques, like classifiers or experts, to tackle specific computational intelligence challenges. Rather than creating just one classification tree from a dataset, the Random Forest generates a forest of classification trees. Each of these trees produces a classification based on a given set of features. Below is a description of the random forest workflow.

(i) Choose K data points at random from the training set.

(ii) Create the decision trees using these K data points.

(iii) Repeat steps (i) and (ii) after selecting the number of N-trees from the created trees.

(iv) Assign a new data point to the category that seems most probable by constructing the N- tree that predicts the category relevant to the data points.

vii. Logistic Regression

An analytical modeling method called logistic regression links a set of explanatory variables to the likelihood of a level. It is employed when examining a dataset where a result is determined by one or more independent variables. A binary variable—one with only two possible outcomes—is used to measure the outcome. Given a set of independent variables, It is utilized to forecast a binary outcome (True/False, 1/0, Yes/No). The LR model can be expressed through the following equations:

$$x = co + \sum_{i=1}^{n} c;x;$$

$$P(x) = \frac{e^{x}}{1 + e^{x}}$$

of the distance is a major challenge because all of the data points are in metric space. N samples are taken into consideration using the following distance metric value if the number of neighbours in K-NNs is represented by N:

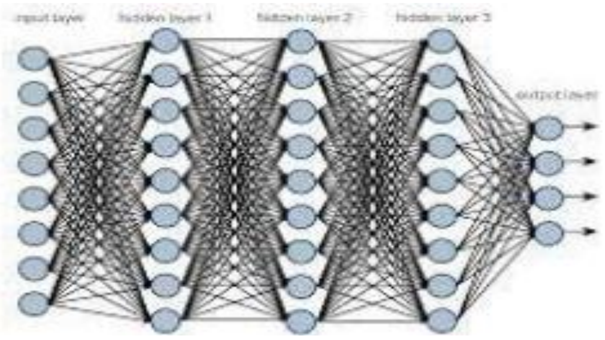
An analytical modeling method called logistic regression links a set of explanatory variables to the likelihood of a level. It is employed when examining a dataset where a result is determined by one or more independent variables. A binary variable—one with only two possible outcomes—is used to measure the outcome. Given a set of independent variables, it is used to predict a binary result (True/False, 1/0, Yes/No). The LR model is represented by the following equations:

viii. Naive Bayes Algorithm (NB)

A sizable training dataset is assumed when using this model. The algorithm uses the Bayesian approach to determine the probability. When determining the probability of noisy data used as an input, it offers the maximum accuracy. The training dataset and training tuple are compared using this analogy classifier [34].

ix. Artificial Neural Networks

The dendrite, soma, and axon processes of biological neurones are followed by artificial neural network algorithms [35], which are somewhat modelled after biological neurons [36]. An artificial neurone and a basic mathematical function make up each ANN's underlying structure [37]. A collection of interconnected neurones arranged in three distinct layers—the input, hidden, and output layers—make up the fundamental structure of an artificial neural network.



[Fig.3: Artificial Neural Network Diagram]

In general, this kind of network learns to execute tasks by taking into account a sufficient number of examples. Both classification and regression problems can be solved with a neural network. Perceptrons, the most basic type of ANN used for binary classification, and multilayer ANNs, a more advanced version of perceptron used to tackle complex classification and regression issues, are the two forms of ANNs that are available. The representation for a single neuron's forward propagation and prediction is as follows:

$$\text{Output} = b¿ + \sum^{n,} w¿;x¿$$

where w;; weight from input to output layer, b; bias value, and x; input value.

III. RESULT OF PROPOSED METHODOLOGY

A. Performance Measure Parameters

Several performance measure metrics are utilized to evaluate the effectiveness of machine learning methods. To analyze these metrics, a confusion matrix is constructed, which includes True Positives Rate (TPR), False Positives Rate (FPR), True Negatives Rate (TNR), and False Negatives Rate (FNR) based on actual and predicted data. Here are the definitions of these terms:

TRUE	Positive	Rate	=	TPR
TRUE	Negative	Rate	=	TNR
FALSE	Positive	Rate	=	FPR
FALSE	Negative	Rate	=	FNR

The following criteria are often utilized in our study to assess certain terms using the formula that corresponds to them in order to gauge their effectiveness. Numerous parameters, such as these, explain certain relationships that can be used to gauge a system's performance. The performance of the comparison study is assessed using the following formulas:

Accuracy (Acc) The ratio of correctly classified samples to total samples:

$$\textit{Accuracy Rate (Acc)} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

Sensitivity (Sen) Sensitivity is also regarded as recall. The rate of the perceived positive case with the total positive cases:

$$\textit{Sensitivity Rate (Sen)} = \frac{(TP)}{(TP+FN)}$$

Specificity (Spec) According to the rate of expected presence, including complete examples, by the existence of breast cancer, specificity is defined as the relationship between observed negative examples and all negative examples.

$$\textit{Specificity Rate (Spec)} = \frac{(TN)}{(TN+FP)}$$

Accuracy (Prec) The division of the cases that are truly positive among all of the examples that we expected to be positive is known as precision:

$$\textit{Precision Rate} = \frac{(TP)}{(TP+FP)}$$

NPV, or negative predictive value The percentage of cases with a negative classification that stayed genuinely negative is known as the NPV:

$$\textit{Negative predictive value Rate (NPV)} = \frac{(TN)}{(TN+FN)}$$

FPR, or false-positive rate The number of false- positive predictions divided by the total number of negatives is known as the false-positive rate. Valid false-positive rates range from 0.0 to 1.0, which is the maximum possible:

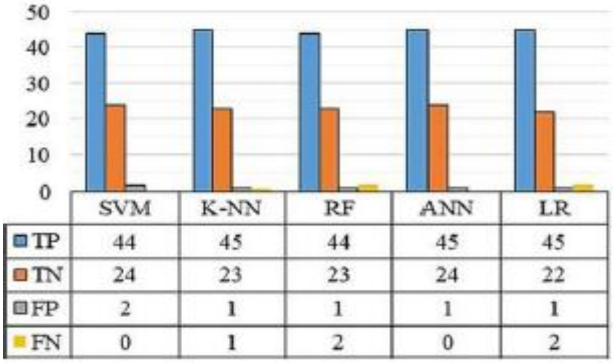
$$\textit{False-positive rate (FPR)} = \frac{(FP)}{(FP+TN)}$$

FNR, or false-negative rate People who have the condition or illness for which they are being evaluated are brought about by the rate of negative test results*:*

$$\textit{False-negative rate (FNR)} = \frac{(FN)}{(FN+TP)}$$

B. Experimental Setup

In order to determine if a cell is benign or malignant, five artificial intelligence techniques were used separately: SVM, K-NN, RF, ANN and LR. The experiment was performed on a laptop Dell Inspiron 15 with Intel core i3 processor, and also made use of Google Colaboratory to improve computation capabilities. All these models were implemented using Scikit-learn Python machine learning framework. We first and foremost used Jupyter Notebook and Google Colab, both free web applications that allow creating with interactive content such as code, images, videos and other media to effectively arrange and present our results.



	SVM	K-NN	RF	ANN	LR
■ TP	44	45	44	45	45
■ TN	24	23	23	24	22
■ FP	2	1	1	1	1
■ FN	0	1	2	0	2

[Fig.4: Confusion Matrix for the Prediction of Breast Cancer using Five Machine Learning Techniques]

The SVM confusion matrix from the ten-fold cross-validation shows that the model achieved a true positive rate of 44 instances (62.86%) for benign cases, with only 1 instance (1.43%) incorrectly classified as a false positive. For malignant cases, the model reported zero false negatives (0.00%) and successfully identified 23 instances (32.85%) as true negatives.

For the KNN confusion matrix, the model correctly classified 5 benign instances, with a true positive rate of 64.29%. However, for malignant cases, the model achieved a higher true positive count of 45, maintaining the same accuracy of 64.29%. There was 1 instance (1.43%) classified as a false negative, while 23 malignant cases (32.85%) were correctly identified as true negatives.

The RF (Random Forest) confusion matrix shows a true positive rate of 62.86% for benign cases, with 4 true positives. For malignant cases, it identified 44 instances as true positives, also at 62.86%. There were 2 false negatives (2.86%) and 23 true negatives (32.85%).

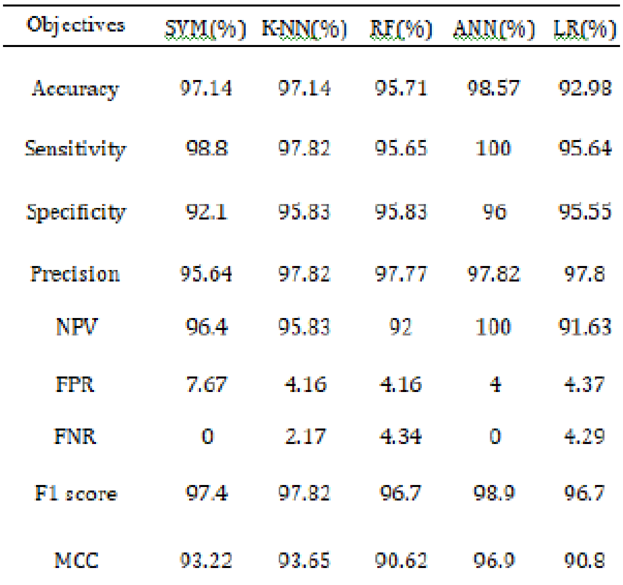
In the ANN (Artificial Neural Network) confusion matrix, the model classified benign instances with a high true positive rate of 64.29%, capturing 45 instances accurately. It also correctly identified 45 malignant cases with the same rate (64.29%), while no false negatives (0.00%) were observed. The true negatives for malignant cases were recorded as 24, resulting in an accuracy rate of 34.28%.

Lastly, the LR (Logistic Regression) confusion matrix indicates a true positive rate of 64.29% for both benign and malignant instances, with 45 cases correctly classified for each. It had 2 false negatives (2.86%) and correctly identified 22 malignant cases as true negatives (31.42%).

- Each technique's confusion matrix is computed. 499 cases, or 90% of the 569 instances in the dataset, were utilized to train each of the five methods. We tested both of our trained models on 70 occasions. The prediction results of SVM, K- NNs, RFs, ANNs, and LRs are provided by the confusion matrix of the machine learning techniques that are employed. Accuracy: ANN achieves the highest accuracy at 98.57%, followed by SVM and K-NN, both at 97.14%. RF and LR have lower accuracies at 95.71% and 92.98%, respectively.
- **Sensitivity**: ANN leads with perfect sensitivity (100%), followed by SVM at 98.8%. K-NN has 97.82%, RF 95.65%, and LR 95.64%.
- **Specificity**: ANN has the highest specificity at 96%, followed by K-NN and RF both at 95.83%. LR and SVM have slightly lower specificities, with 95.55% and 92.1%, respectively.
- **Precision**: Both ANN and K-NN report a precision of 97.82%, with LR close at 97.8%. RF also shows strong precision at 97.77%, while SVM has a precision of 95.64%.
- **Negative Predictive Value (NPV)**: ANN has the highest NPV at 100%, with SVM at 96.4%. K- NN has 95.83%, RF 92%, and LR 91.63%.
- **False Positive Rate (FPR)**: ANN has the lowest FPR at 4%, followed closely by K-NN and RF at
- 4.16%. LR and SVM have FPRs of 4.37% and 7.67%, respectively.
- **False Negative Rate (FNR)**: ANN and SVM achieve the lowest FNR at 0%, while K-NN, RF, and LR have FNRs of 2.17%, 4.34%, and 4.29%, respectively.
- **F1 Score**: ANN achieves the highest F1 score at 98.9%, followed by K-NN at 97.82% and SVM at 97.4%. RF and LR both have F1 scores of 96.7%.
- **Matthews Correlation Coefficient (MCC)**: ANN shows the highest MCC at 96.9%, followed by K-NN with 93.65% and SVM at 93.22%. RF and LR have lower MCC values of 90.62% and 90.8%, respectively.

All the techniques have an F1 score of nearly 97%, which is comparatively better.

Table 3: Performances of Breast Cancer Prediction System



Objectives	SVM(%)	K-NN(%)	RF(%)	ANN(%)	LR(%)
Accuracy	97.14	97.14	95.71	98.57	92.98
Sensitivity	98.8	97.82	95.65	100	95.64
Specificity	92.1	95.83	95.83	96	95.55
Precision	95.64	97.82	97.77	97.82	97.8
NPV	96.4	95.83	92	100	91.63
FPR	7.67	4.16	4.16	4	4.37
FNR	0	2.17	4.34	0	4.29
F1 score	97.4	97.82	96.7	98.9	96.7
MCC	93.22	93.65	90.62	96.9	90.8

IV. CONCLUSION

In this work, we have explored various algorithms with supporting deep learning and machine learning approaches for breast cancer prediction. We seek to find the most accurate algorithm that will predict breast cancer more effectively. The main objective of this review is to ascertain and document all previous efforts using machine learning as an intelligent tool for early breast cancer detection. The forms of breast cancer are the first topic covered in this paper's review. Following that, a summary of the main deep learning and machine learning approaches was given. These techniques involve extremely complex algorithms that are used to forecast breast cancer. The best result was reached by the neural networks where accuracy reached 98.57%, while the smallest accuracy, 95.7%, was reached by the RFs and LRs. In the field of medicine, the process of diagnosis takes time and is expensive. The system's output demonstrates that breast cancerML can be employed clinically in the diagnosis of breast cancer. In the event of a misdiagnosis, this technology will be very beneficial for new physicians. We can infer from the findings that machine learning methods can accurately and automatically identify the illness.

DECLARATION STATEMENT

After aggregating input from all authors, I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** This article has not been sponsored or funded by any organization or agency. The independence of this research is a crucial factor in affirming its impartiality, as it has been conducted without any external sway.
- **Ethical Approval and Consent to Participate:** The data provided in this article is exempt from the requirement for ethical approval or participant consent.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Authors Contributions:** The authorship of this article is contributed equally to all participating individuals.

REFRENCES

1. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiol Soc N Am. DOI: https://doi.org/10.1148/radiol.2017171920
2. Islam, M. M., Haque, M. R., Iqbal, H., Hasan, M. M., Hasan, M., & Kabir, M. N. (2020). Breast cancer prediction: A comparative study using machine learning techniques. DOI: https://doi.org/10.1007/s42979-020-00305-w
3. Y. Khourdifi and M. Bahaj, "Applying best machine learning algorithms for breast cancer prediction and classification," in 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), pp. 1–5. DOI: https://doi.org/10.1109/ICECOCS.2018.8610632
4. Y. Lu, J. Y. Li, Y. T. Su, and A. A. Liu, "A review of breast cancer detection in medical images," in 2018 IEEE Visual Communications and Image Processing (VCIP). DOI: https://doi.org/10.1109/VCIP.2018.8698732
5. A. Reddy, B. Soni, and S. Reddy, "Breast cancer detection by leveraging machine learning," ICT Express, 2020. DOI: https://doi.org/10.1016/j.icte.2020.04.009
6. Z. Salod and Y. Singh, "Comparison of the performance of machine learning algorithms in breast cancer screening and detection: A protocol". Journal of Public Health Research, vol. 8, no. 3, 2019. DOI: https://doi.org/10.4081/jphr.2019.1677
7. S. Eltalhi and H. Kutrani, "Breast cancer diagnosis and prediction using machine learning and data mining techniques: A review," IOSR Journal of Dental and Medical Sciences (IOSR-JDMS). https://www.researchgate.net/publication/333092560_Breast_Cancer_Diagnosis_and_Prediction_Using_Machine_Learning_and_Data_Mining_Techniques_A_Review
8. I. H. Witten and E. Frank, Data mining: practical machine learning tools and techniques with Java implementations, vol. 31 of Acm Sigmod Record. Elsevier, 2005. DOI: https://doi.org/10.1145/507338.507355
9. Milosevic, Marina; Jankovic, Dragan; Milenkovic, Aleksandar; Stojanov, Dragan . (2018). *Early diagnosis and detection of breast cancer. Technology and Health Care, (), 1–31.* DOI: https://doi.org/10.3233/THC-181277
10. Connolly JL, Schnitt SJ. Benign breast disease: resolved and unresolved issues. Cancer 1993;71:1187-9. DOI: https://doi.org/10.1002/1097-0142(19930215)71:4%3C1187::aid-cncr2820710402%3E3.0.co;2-v
11. L. C. Hartmann et al., "Benign Breast Disease and the Risk of Breast Cancer," New England Journal of Medicine, vol. 353, no. 3. Massachusetts Medical Society, pp. 229–237, Jul. 21, 2005. DOI: https://doi.org/10.1056/NEJMoa044383
12. Wang J, Costantino JP, Tan-Chiu E, Wickerham DL, Paik S, Wolmark N. Lowercategory benign breast disease and the risk of invasive breast cancer. J Natl Cancer Inst 2004;96:616-20. DOI: https://doi.org/10.1093/jnci/djhs105
13. Haagensen CD. Carcinoma of the Breast: A Monograph for the Physician. American Cancer Society, 1958; 7. DOI: https://doi.org/10.1097/00000658-194311850-00008
14. Roger S. Foster Jr; Michael C. Costanza. (1984). *Breast self-examination practices and breast cancer survival. , 53(4), 999–1005.* DOI: https://doi.org/10.1002/1097-0142(19840215)53:4<999::AID-CNCR2820530429>3.0.CO;2-N
15. F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, 'Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries', *CA Cancer J. Clin.*, vol. 68, no. 6, pp. 394–424, Nov. 2018. DOI: https://doi.org/10.3322/caac.21492
16. C. E. DeSantis, S. A. Fedewa, A. Goding Sauer, J. L. Kramer, R. A. Smith, and A. Jemal, 'Breast cancer statistics, 2015: Convergence of incidence rates between black and white women', *CA Cancer J. Clin.*, vol. 66, no. 1, pp. 31–42, Jan. 2016. DOI: https://doi.org/10.3322/caac.21320
17. Verma, B. and Zakos, J. (2001) A Computer-Aided Diagnosis System for Digital Mammograms Based on Fuzzy-Neural and Feature Extraction Techniques. IEEE Transactions on Information Technology in Biomedicine, 5, 46-54. DOI: https://doi.org/10.1109/4233.908389
18. L. Wang, 'Early diagnosis of breast cancer', *Sensors (Basel)*, vol. 17, no. 7, p. 1572, Jul. 2017. DOI: https://doi.org/10.3390/s17071572
19. M. M. Islam, M. R. Haque, H. Iqbal, M. M. Hasan, M. Hasan, and M. N. Kabir, 'Breast cancer prediction: A comparative study using machine learning techniques', *SN Comput. Sci.*, vol. 1, no. 5, Sep. 2020. DOI: https://doi.org/10.1007/s42979-020-00305-w
20. Sakri SB, Rashid NBA, Zain ZM. Particle swarm optimization feature selection for breast cancer recurrence prediction. IEEE Access. 2018;6:29637–47. DOI: https://doi.org/10.1109/ACCESS.2018.2843443
21. D. L. Olson and D. Delen, Advanced data mining techniques. Springer Science and Business Media, 2008. DOI: https://doi.org/10.1007/978-3-540-76917-0
22. L. Li et al., "Research on machine learning algorithms and feature extraction for time series," in 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), pp. 1–5, IEEE. DOI: https://doi.org/10.1109/PIMRC.2017.8292668
23. Azar AT, El-Said SA. Performance analysis of support vector machines classifiers in breast cancer mammography recognition. Neural Comput Appl. 2013;24(5):1163–77. DOI: https://doi.org/10.1007/s00521-012-1324-4
24. Yue W, et al. Machine learning with applications in breast cancer diagnosis and prognosis. Designs. 2018;2(2):13. DOI: https://doi.org/10.3390/designs2020013
25. Banu AB, Subramanian PT. Comparison of Bayes classifiers for breast cancer classification. Asian Pac J Cancer Prev (APJCP). 2018;19(10):2917–20. DOI: https://doi.org/10.22034/apjcp.2018.19.10.2917
26. Chaurasia V, Pal S, Tiwari B. Prediction of benign and malignant breast cancer using data mining techniques. J Algorithms Comput Technol. 2018;12(2):119–26. DOI: https://doi.org/10.1177/1748301818756225
27. Azar AT, El-Metwally SM. Decision tree classifiers for automated medical diagnosis. Neural Comput Appl. 2012;23(7–8):2387–403. DOI: https://doi.org/10.1007/s00521-012-1196-7
28. Senapati MR, Mohanty AK, Dash S, Dash PK. Local linear wavelet neural network for breast cancer recognition. Neural Comput Appl. 2013;22(1):125–31. DOI: https://doi.org/10.1007/s00521-011-0670-y
29. Breast Cancer Wisconsin (Original) Data Set, [Online]. Accessed 25 Oct 2024. https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data/data
30. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. 1^{st} ed. New York: Springer; 2013. DOI: https://doi.org/10.1007/978-1-0716-1418-1
31. Guido S, Müller AC. Introduction to
machine learning with python. Sebastopol: O'Reilly Media Inc.; 2016. https://www.nrigroupindia.com/e-

book/Introduction%20to%20Machine%20Learning%20with%20Python%20(%20PDFDrive.com%20)-min.pdf

32. Dong L, Wesseloo J, Potvin Y, Li X. Discrimination of mine seismic events and blasts using the fisher classifier, naive Bayesian classifier and logistic regression. Rock Mech Rock Eng. 2015;49(1):183–211. DOI: https://doi.org/10.1007/s00603-015-0733-y
33. Fatima, Noreen; Liu, Li; Sha, Hong; Ahmed, Haroon . (2020). Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques and their Analysis. IEEE Access, (), 1–1. DOI: https://doi.org/10.1109/ACCESS.2020.3016715
34. Ratner B. Statistical and machine-learning data mining: techniques for better predictive modeling and analysis of big data. Oxford: Chapman and Hall/CRC; 2017. DOI: https://doi.org/10.1201/9781315156316
35. Biswas, R., Roy, S., & Biswas, A. (2019). Mammogram Classification using Curvelet Coefficients and Gray Level Co-Occurrence Matrix for Detection of Breast Cancer. In International Journal of Innovative Technology and Exploring Engineering (Vol. 8, Issue 12, pp. 4819–4824). DOI: https://doi.org/10.35940/ijitee.l3694.1081219
36. Rani, Dr. Y. U., Kotturi, L. S., & Sudhakar, Dr. G. (2021). A Deep Learning Technique for Classification of Breast Cancer Disease. In International Journal of Engineering and Advanced Technology (Vol. 11, Issue 1, pp. 9–14). DOI: https://doi.org/10.35940/ijeat.a3119.1011121
37. Rajasekaran, G., & Ram, Dr. C. S. (2023). Breast Cancer Prediction Based on Feature Extraction using Hybrid Methodologies. In International Journal of Soft Computing and Engineering (Vol. 13, Issue 2, pp. 20–28). DOI: https://doi.org/10.35940/ijsce.b3612.0513223

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Lattice Science Publication (LSP)/ journal and/ or the editor(s). The Lattice Science Publication (LSP)/ journal and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.