ISSN No:-2456-2165

Mammographic Mass Detection Using Machine Learning Classifiers

[1]Vaishnavi M S
Computer Science & Engineering Department,
Dayananda Sagar College of Engineering, Affiliated to VTU
Bangalore, India

[3]Deekshitha S
Computer Science & Engineering Department,
Dayananda Sagar College of Engineering, Affiliated to VTU
Bangalore, India

[2]Ujjawal Choudhary
Computer Science & Engineering Department,
Dayananda Sagar College of Engineering, Affiliated to VTU
Bangalore, India

[4]Supriya D R
Computer Science & Engineering Department,
Dayananda Sagar College of Engineering, Affiliated to VTU
Bangalore, India

[5]Dr. J Vimala Devi, (Associate Professor)
Computer Science & Engineering Department,
Dayananda Sagar College of Engineering, Affiliated to VTU Bangalore, India

Abstract:- The most serious sort of cancer that affects women most frequently in modern times is breast cancer (BC). If it is not detected in the beginning stages, the death rate is significant. Breast cancer affects around 12% of woman, and the percentage is continually rising. The invention of a ML-based breast cancer classification system that can identify the disease from a patient's digital scan is artificial intelligence's greatest contribution to healthcare. Breast cancer is found using the mammography method however, radiologists' interpretations vary greatly. Fine needle aspiration cytology is commonly used in the diagnosis of breast cancer (FNAC). Uniform cell size, if the tumour has a consistent cell shape and other factors are taken into account, the prediction will determine if the tumour is benign or malignant. We have gathered both the characteristics of breast cancer cells and cells from healthy individuals. We were able to differentiate between malignant and benign employing a supervised machine learning classifier system to identify tumours. However, by taking the right medications, needless therapy can be avoided if patients are correctly recognised early on employing ML approaches. Though computer vision, ML technologies have demonstrated a high level of accuracy in healthcare applications, Physical examinations shouldn't solely be conducted using these systems. These are meant to support doctors, not replace them. Machine learning has a distinct advantage in that it can find relevant breast cancer features in large datasets. In predictive modelling and pattern recognition, the technique is extensively used. Assessing each classifier's effectiveness in terms of accuracy, precision, and recall is the project's main objective. Examining the effectiveness, accuracy, and early identification of breast cancer using different machine learning classifiers is the goal.

Keywords:- *Breast Cancer (BC), Mammography, Fine Needle Aspiration Cytology(FNAC).*

I. INTRODUCTION

Breast cancer, which affects 10% of all women at some point in their life, is the most prevalent kind of cancer in females. The largest leading cause of death for women is breast cancer and its prevalence is rising 25% of American women will be diagnosed with BC at some time in their life, according to a survey. Women are diagnosed with BC 43% of the time in the United Arab Emirates. It is still challenging for many medical professionals to accurately forecast the growth of a cancerous tumor. The introduction of new medical technologies and the vast amount of patient data have allowed for the creation of new methods for the prognosis and detection of cancer. DNA modifications, alterations play an important play in the growth of breast cancer. Common breast cancers include invasive carcinoma, ductal carcinoma in situ (DCIS). Angiosarcoma and phyllodes tumors are less frequent varieties. Fatigue, headaches, discomfort and numbness (peripheral neuropathy), bone loss, and osteoporosis are some of the adverse effects of breast cancer symptoms. The impact of the sickness depend on the patient's age, risk level, and cancer status. Mammogram screening or categorical data can be used to examine breast lumps and determine whether they are breast cancerous. Tumors or the aberrant cell development that causes them, can be benign or malignant, depending on how they form early detection and prevention of the disease's transmission will spare people from death. The only method for spotting breast cancer was an X-ray. Although FNAC is frequently used to detect breast cancer, only 90% of cases are accurately diagnosed on average. Artificial intelligence and deep learning are two methods that have been created and proposed for finding a procedure that is more efficient than X-ray procedures.

ISSN No:-2456-2165

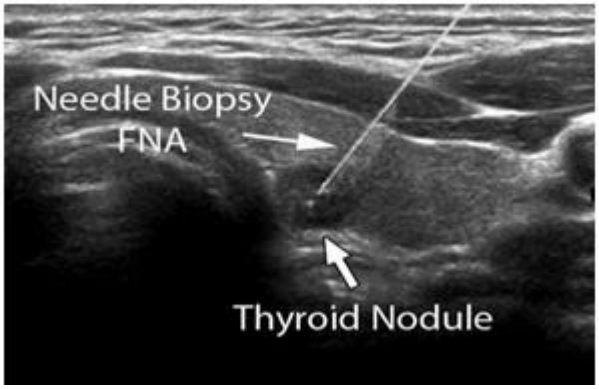

Fig 1 Thyroid Nodule

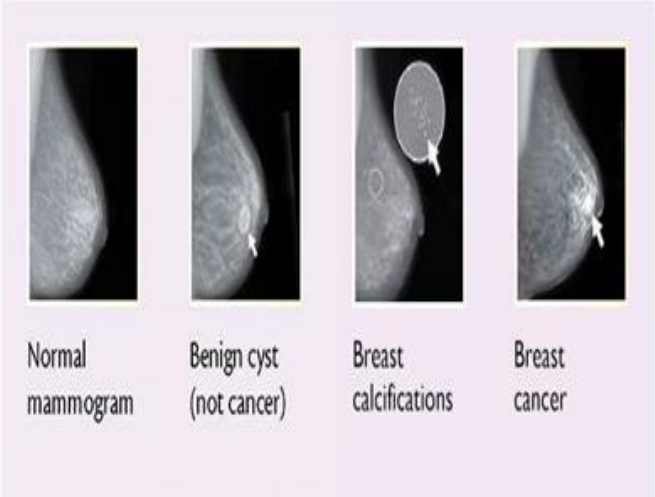

Fig 2 Classification Techniques for Breast Cancer

A branch of artificial intelligence called machine learning studies the creation of algorithms from data. That will identify, treat, and forecast dangerous, prevalent diseases like cancer, hepatitis, and heart disorders, machine learning is commonly utilized. Supervised and unsupervised machine learning algorithms are the two categories. There are numerous classification techniques for breast cancer outcomes. The performance of four classifiers is compared in the following paper: The most important data mining methods are SVM, LR, Random Forest algorithm, and KNN. Medically, it can be found early during a screening exam using mammography or a portable cancer detection device.

Patients with breast cancer must have surgery, chemo, radio, and endocrine treatments to stop the spread of the disease. The objective of the study is to distinguish between patients who are malignant and benign and to categorize them accurately. In order to describe breast cancer, we are investigating a variety of datasets and machine learning techniques. We want to reduce error rates as exactly as feasible. JUPYTER, analyses data and evaluates its effectiveness and efficiency utilizing a machine learning approach called the 10-fold cross validation test.

II. MACHINE LEARNING ALGORITHMS

The process of integrating learning capabilities into a system based on training data sets is referred to as "Machine Learning," and it is a subset of artificial intelligence. Contrary to the typical method, there are several approaches and strategies available to create systems that can learn. Neural networks, decision trees, and clustering are a few of them. Three categories are used to classify AL-ML. 1. Reward- based learning 2.Supervised learning 3.Unsupervised learning.

- *An iterative learning method based on the environment is reinforcement learning. The system gradually learns all possible system states over an extended period of time. Random Forest, a supervised learning algorithm, generates a group of decision trees. The system is trained using the bagging strategy, and the main technique is recursion. Throughout each iteration, a random sample of size N is picked from the entire data set.*
- *The function created from the training data directs the system to provide insightful results for fresh data sets that are put into it. Based on observations made during the learning process, supervised learning makes predictions about results.*
- *Unsupervised learning is the process of teaching a computer from an unlabelled dataset, then separate it based on a set of characters, and then use that information to make decisions on its own without the assistance of a human trainer.*

Results can be categorised effectively and efficiently using classification and data mining attributes. We use a variety of ML techniques to forecast the occurrence of BC and then use machine learning models to assess the early prediction performance metrics, the area under the receiver operating characteristic curve, and the prediction accuracy.

After analysis, the support vector algorithm achieved accuracy with ROC; their mathematical models can alter this strategy to enhance breast cancer prediction. Based on their documented collection of parameters, this model aids physicians in making accurate predictions about a patient. This is used to diagnose, prognosticate, and treat serious, frequent diseases. We divide these algorithms into supervised and unsupervised categories. Supervised learning is a function that converts input into output pairs. By analysing the labelled training data set, an inferred function is produced that may be used to map new samples. Unsupervised learning is a subset of these methods that uses the least amount of human interaction and no pre-defined labels to find patterns in data that were previously undetected. Cluster analysis and main components are two techniques used in unsupervised learning. It represents the architecture for putting these methodologies' related research into practise. It discusses various ML algorithms that are applied to the analysis processes. We demonstrate an experimental setup and ML algorithm results. The SVM model performs better when categorising data using linear or nonlinear kernel functions and nonlinear data. For

prediction, we've employed a linear kernel function. Although more features have been included, the KNN is still simple to develop. However, prior to applying methods that are assessed for accuracy, precision, recall, f1-score, and ROC, features must first be scaled. The measurement of a prediction made by an algorithm is accuracy. The ROC is utilised in the framework for medical diagnostics, which provides a common mechanism for evaluation based on the mean of each ROC curve point. Accuracy and ROC performance metrics for Machine Learning algorithms. In order to determine the class, we will employ Random Forest, Confusion Matrix, and K-Nearest-Neighbour (KNN) representations of the training data points close to the test data points. The K-nearest-neighbour techniques are used to determine a data set's membership in a group based on the neighbouring data sets. This method, which uses supervised learning, is employed in the classification and regression processes. KNN gathers all the data points near a new data point before processing it. A high degree of variance in the qualities has a significant impact on the distance. The KNN algorithm locates the k-nearest-neighbours of N training vectors regardless of labels.

Using the Bayes theorem as a foundation, classifiers are described as being non-deterministic in nature. Naive Bayes appears to work well for a variety of ML problems, despite the fact that it is naive to assume that each feature is independent of the others—which is not necessarily the case in real-world circumstances. There is a comparison of the proposed algorithms using the Naive Bayes and Confusion Matrix. The only topic discussed is the comparison of these algorithms—KNN, Naive Bayes, and Random Forest—in terms of their capabilities and performance.

The Naive Bayes scenario is one where K is the number of randomly chosen variables for each node, d is the feature dimension, and N is the number of training examples. The training set's pre-processing is not done before the KNN test period. The KNN, Random Forest algorithms will address both classification and regression concerns, while the Naive Bayes technique exclusively addresses issues with classification. Both KNN, Random Forest can generate highly accurate results in terms of accuracy, but the Naive Bayes method necessitates an abundance of records to produce a more precise result. Parametric and nonparametric models are algorithms that decompose the function to be known from. KNN, Naive Bayes, and Random Forest are contrasted. When this model is trained, accuracy, which is defined as the measurement of the correct prediction in accordance with the equation used to calculate accuracy's value, is a good indicator of the degree of correctness.

$$Accuracy = \frac{TrueNegatives + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

Recall is the ratio of correctly identified positive observations to all observations, also known as sensitivity in general. Recall can be used as a gauge for how well a system predicts good outcomes and calculates expenses.

True Positive = False Negative if recall is True Positive.

The degree of accuracy in forecasting the favourable outcomes will serve as the definition of precision. It essentially measures the proportion of real positive outcomes to all good outcomes. This demonstrates the system's ability to handle positive values, but it makes no mention of how it would respond to a negative value.

$$\text{Precision} = \frac{TP}{(TP + FP).}$$

Below is a representation of the precision and recall weight average.

F1's rating is

$$F1 = \frac{2TP}{(2TP + FP + FN)}$$

III. RELATED WORK

1. Manav Mangukiya paper used different ML techniques for finding breast cancer. Project goals were to evaluate ML predictions by visualizing and analyzing the Wisconsin dataset. With the help of this, we demonstrate the superiority of Naive Bayes, Support Vector Machines, Adaboost, Random Forest Classifiers, KNN, Decision Trees, XGboost, and other methods over SVM. Thus came to the conclusion that XGboost, with an efficiency of 98.24%, is the finest algorithm for identifying breast cancer that produces the most accurate results. However, the approach cannot be applied without first processing the dataset. In the future, he would like to expand the dataset and evaluate the algorithm's effectiveness and scalability.
2. In this study, SVM, KNN, random forests, ANN, and LR were compared to each other as ML techniques to detect breast cancer. Each of this five machine learning techniques' fundamental characteristics and operation were demonstrated. The lowest accuracy obtained from RFs and LR, which is 95.7% and maximum accuracy attained by ANNs is 98.57%. In the medical field, the diagnosing process is very time- and money-consuming. In the event of a misdiagnosis, the system predicted that ML approaches would be extremely helpful for newly qualified medical professionals or doctors. They might also act as a clinical aid for the detection of breast cancer. The model created using ANNs is more trustworthy than any other method mentioned and has the potential to completely change the way breast cancer is predicted. The study claims that the disease may be accurately and automatically diagnosed using ML techniques.
3. Anoy Chowdhury project, developed graphs and results for the same breast cancer tumour predictor that they learned to build using the Wisconsin dataset. A solid dataset has been found to offer greater accuracy. The creation of prediction systems will result from the

selection of appropriate algorithms with a strong home dataset. When a patient has been diagnosed with breast cancer, these systems can help determine the best course of treatment. Based on the stage of a patient's breast cancer, there are a variety of treatments available; By extracting knowledge from these pertinent databases, data mining and machine learning (ML) can assist in choosing the course of treatment to be taken.

4. This paper's primary goal is to analyze various ML algorithm that helps in early detection of breast cancer. The Wisconsin (Diagnostic) dataset, which include 569 records, 32 features, was used for this. Eleven different machine learning techniques were used to analyse the dataset: Logistic Regression, SVM, Extra Trees Classifier, Ada Boost Classifier, Light Gradient Boosting Machine, KNN classifier, Ridge Classifier, Random Forest Classifier, Naive Bayes, Gradient Boosting Classifier, and DT Classifier. The experiment's findings unmistakably demonstrate that of all regression models, logistic regression has the highest accuracy (97.89%), F1-score (97.35%), specificity (90.69%), and sensitivity (100%) of any regression model. In the future, more data will be used to improve the algorithm's effectiveness and scalability.
5. Tanmay Padhi, Aditi Kumari paper says In order to treat, breast cancer it must be found in its early stages. Despite the growing pandemic, many people are currently battling breast cancer, and many have already passed away in recent years. We came to the conclusion that SVM provides a greater precision and accuracy 97.14% than any other method and can be used for medical services after running various algorithms against the dataset. With the increase in data quantity and other factors, this data will become better in the next years.
6. M Sumanth provide a verified ensemble machine learning method in order to diagnose breast cancer. This technique's accuracy rate is 98.50 percent. This study used only 16 participant-reported variables to identify a cancer diagnosis.
7. In, different ML methods are utilized for identifying breast cancer. They conducted side-by-side comparison of the CNN, K-Nearest Neighbor, Support Vector Method, Logistic Regression, Naive Bayes, and Random Forest algorithms. It was discovered that CNN outperforms the present methods in terms of precision, accuracy, and data set size.
8. Samer Hamed, Abdelwadood Mesleh, Abdullah Arabiyyat paper uses On the Wisconsin BCs data set, five machine learning techniques were used to detect BCs. As a result of achieving the highest F-measure and accuracy scores, random forest produced the best outcomes, additionally, random forest completed the testing process in under a second. This suggests that random forest will be able to help clinicians automatically and confidently determine if their patients have benign or malignant tumors with its 99% accuracy.
9. Seeyog Kapadne, Arsh Patne, Jahan Chaware and Prof. Priyanka Shahane study has led us to the conclusion that a variety of machine learning techniques and algorithms, including CNN, SVM, RNN, Linear regression, BCC, KNN, Random Forest, and others, can be used to solve the problem of early cancer cell detection. CNN and linear regression, out of all these algorithms, have the highest accuracy and can be utilized to extract an effective and precise result. 99.7% of the time it is accurate. To enhance throughput even further, we can employ deep learning and softmax regression.
10. We can see that SVM builds its model in roughly 0.07 seconds as opposed to k-0.01 NN's seconds. As opposed to C4.5, Naive Bayes, and k-Nearest Neighbor, whose accuracy ranges from 95.12% to 95.28%, Support Vector Method's accuracy (97.13%) is higher. It is also clear that SVM surpasses the competition in terms of the proportion of correctly categorized cases versus the proportion of wrongly classified cases. We may now assess the outcomes to discover how well our algorithms performed after developing the predicted model. SVM and C4.5 achieved the greatest TP value for the benign class (97%), despite the fact that k-NN correctly predicts 97% of cases that belong to the malignant class. Support Vector Method classifiers have a lower FP rate (0.03 for benign class and 0.02 for malignant class), and then the k-NN, C4.5, and NB methods are used. These findings demonstrate why SVM outperformed other classifiers. Finally, using recall and accuracy, SVM was able to show the effectiveness and efficiency of its methodologies.
11. The type of cancer that develops most frequently is BC. A woman selected at random has a 12% chance of developing the illness. So that, many precious lives can be saved through early identification of breast cancer. This paper provides a model that evaluates various machine learning methods for the detection of breast cancer. Using the Wisconsin Diagnosis dataset, performance comparison of ML methodologies are conducted on Dataset of cancer, each method used to identify benign or malignant tumours was found to have an accuracy rate of greater than 94%. KNN beat the competition in terms of accuracy, precision, and F1 score, making it the most effective algorithm for detecting breast cancer. Therefore, supervised machine learning algorithms will be highly beneficial for early detection and cancer type prognosis in cancer research.
12. Medical areas have used machine learning extensively, and they were an effective diagnostic tool for doctors in the development and interpretation of the data provided to expert systems in medicine. The SVM, a popular ML method is commonly used in the detection and diagnosis of breast cancer, the Bayesian networks with Random Forest (RF), and (BN). The three ML techniques' key traits and methodology were described. Evaluation of the performance under examination.
13. Paper have covered two well-liked ML approaches on Wisconsin BC categorization in this research. For the classification of the WBC (Original) dataset in the

WEKA tool, ANN and SVM are employed as ML approaches. Comparing the performance of various applied ML methods employ crucial performance parameters like recall, accuracy, and ROC area. SVM (Sequential Minimal Optimization Algorithm) has shown the highest level of accuracy performance for the WBC dataset's diagnosis and forecast, based on the performance criteria of the ML techniques employed.

14. Table 1 shows that SVM builds its model in roughly 0.08 seconds as opposed to K-0 NN's seconds. This might be because K-NN learns slowly and doesn't produce models while training, different from other classifiers, this phenomenon can be explained. However, SVM's accuracy is 97.9%, higher than that of RF, Naive Bayes, and k-NN, whose accuracy levels are 96%, 92.6%, and 96.1% respectively. As demonstrated in table 1, SVM has the highest percentage of correctly graded cases and the lowest percentage of incorrectly rated occurrences, which is another clear distinction between it and other classifiers. Table 1 makes it clearer that SVM produces the best classification probability of with the least warning error rate, 0.95% of 0.021. In addition, after building the projected model, we can now assess the outcomes to gauge the efficiency of our algorithms. In actuality, First table demonstrates that Support Vector Machine achieved the greatest 997.7% TP value for the innocuous class, however only 94.6 for the malignant class. These findings help us to understand why SVM performed better than other classifiers. The ROC curve can be used to better understand the potential of a ML algorithm. Figure 2 makes it clear that SVM is the best classifier since it is (99% sensitive and 99% specific), beginning in the lower left corner and moving up to the upper left corner, upper left corner, and then up to the upper right corner. It should be highlighted that SVM has the finest agreement between the authenticity and dependability of the data collected. The high number of examples improperly because RF and NB have the highest mistake rates (23 incorrect instances for RF and 42 incorrect instances for NB).
15. This paper's goal is to categories the two types of tumors, benign and malignant, and it does so by using multiple machine learning methods to select the best model. It has the ability to accurately classify the tumour. The ML repository at UC Irvine provided the data set for the Wisconsin Breast Cancer Diagnosis (WDBC) project. Prior to doing Principal Component Analysis (PCA) with 17 components, the dataset is pre-processed. Confusion matrices are used to evaluate the results after the application of a number of machine learning algorithms, such as K-NN, LR, and Ensemble Learning. There are 569 cases in the collection, and more information will be included to the database in the future to help machine learning models learn more effectively and work more accurately while also informing us of the relationships between various attributes.
16. One of the dreadful diseases that strike women today is BC. The Wisconsin Breast Cancer Coalition's work with Multiple machine learning (ML) Several techniques were used on a dataset to evaluate their efficacy and utility in identifying the breast cancer types that are malignant and benign with the highest degree of accuracy. For feature selection, the association between the dataset's various features has been examined. The results will be used to choose the best machine learning algorithm to create an automated system for diagnosing breast cancer. Our study points us in this direction that Support Vector Machine plus RF provides the greatest accuracy rate of 96.5%. Future work will involve managing a sizable dataset and including other variables, such as breast cancer phase identification. Our research is hoped to help with the clinical application of breast cancer treatment.
17. The ANN, k-NN, Binary SVM, and AdaBoosting data mining algorithms are briefly discussed in this study. The four CAD methods combine 32 statistical variables across inputs and output in the BCWD. The CF matrix is utilized as the foundation for the data analysis and categorization to determine how many FN, TP, TN, and FP there are. The findings demonstrate that, after features are chosen, k-NN is more trustworthy to support decision-making regarding the severity of breast cancer in this type of dataset. CAD systems are continually subject to new developments in algorithm development and optimization to enhance the system and make it more precise, focused, and extremely sensitive. Convolution Neural Network CNN applications are frequently used in studies on breast cancer pictures. To accomplish the objective to get a more exact and accurate classification, further research along this line should concentrate on integrating classifiers to enhance their performance.
18. In two ML algorithms—DT Classifier and Logistic Regression—are used to determine the likelihood of breast cancer. Each method's accuracy was assessed to decide which was most appropriate for the task. The best method for prediction is the Decision Tree Classifier since it has the [On the "Breast Cancer Wisconsin (Diagnostic) Data Set"], best prediction accuracy. Therefore, utilizing this dataset's features along with our Decision Tree Classifier approach, Breast cancer can be predicted with almost perfect accuracy.
19. Through image analysis, breast cancer can be accurately identified. The set of training data and the features considered for evaluating BC place restrictions on how accurate machine learning models can be. The Wisconsin Breast Cancer Dataset provides a reliable dataset for creating machine learning models. These data are frequently used since it contains a sizable number of nearly noise-free examples. Given that nearly all of the machine learning algorithms that have been put to use have yielded prediction accuracy of greater than 95%, it is obvious that they may be used to predict breast cancer.
20. In, we introduce and compare three unsupervised learning models for BC detection: t-SNE, AEs, and SOMs. Three UL models nonlinearly dimensionally

reduce the space of input instances while maintaining the information from the initial data. In the trials, SERS data that has previously been used in the literature for BC identification two openly accessible BC data sets from the UCI repository were also utilized. Comparing our top-performing UL model to supervised classifiers that have been utilized in the past and produced output on data sets utilized in this study. In terms of the supervised learning evaluation criteria, the comparison revealed that the UL models fared well (AUC, sensitivity, specificity). In the future, the experimental evaluation will be expanded to include additional real-world BC data sets and will switch from sparse to variational AEs. We'll also keep looking at the SERS data set to see which SER spectrum elements are best at distinguishing between benign and malignant classes.

21. In this study, we used the Wisconsin Dataset and a number of machine learning (ML) algorithms to analyze their efficacy and usage in determining the highest level of accuracy in detecting malignant and benign breast cancer. The relationship between the different characteristics in the dataset has been investigated for feature selection. Results will assist in choosing the best machine learning algorithm to create an automated system for diagnosing breast cancer. According to our research, SVM and Random Forest have an accuracy rate of 96.5%, which is the highest. We'll attempt to advance our work in the future by managing a sizable dataset and adding further features like breast cancer phase detection. It is envisaged that our research may assist in the clinical application of breast cancer treatment.
22. Through image analysis, breast cancer can be accurately identified. The training data set and features employed in the analysis of breast cancer limit the accuracy of machine learning models. The Wisconsin Dataset provides a reliable dataset for creating machine learning models. Since it has a substantial number of almost noise-free instances, this dataset is often used. Since nearly all of the used machine learning algorithms have yielded prediction accuracy of 95%, it is evident that these algorithms can be utilized to identify breast cancer.
23. Based on histopathological scans, a Pa-DBN-BC technique for classifying breast cancer is provided. By generating patches of photos of the same size, the suggested model automatically learns the features. Pre-training is done without supervision, and fine-tuning is done under supervision. After learning the features, a probability estimate matrix is built using the patch matching model. The findings demonstrate that the deep learning approach increased the classification precision for breast cancer cases. Our findings show that the patch-based model outperformed models that extract feature information from whole images. This will reduce the cost of calculation and provide extremely high accuracy for the binary classification problem. If more hardware resources, like as a GPU, were made available in the future to support the utilization of a large number of input patches, our model might operate more precisely and effectively. Additionally, our study uses binary classification since we only distinguish between regions associated with cancer and regions that are unrelated to it. We can use this model in the future to work on categorizing other cancer types.
24. Two antenna topologies, BAVA and XETS, are evaluated for their ability to detect breast cancers using ML algorithms and MWI. In order to do this, we put together a diverse collection of experimental measures using five anthropomorphic breasts that each had two tumors in a different volume (measurements without the tumors were also taken for classification purposes). We evaluated three classifiers for ML processing (KNN, LDA, and SVM) and radar-based imaging methods for MWI. Though the Machine Learning algorithms only achieved a maximum accuracy performance of 60% for the XETS when using this antenna for MW imaging, it showed strong detection. The ML algorithms for the Vivaldi antenna are up to 85% accurate, but because of the radiative properties of this type of antenna, the imaging is more crowded than with XETS. ML and MWI's performances have been complementary. Thus, we infer that combining information from the two methodologies improves the precision of tumor detection outcomes. Now, we want to relate the inverse trend of the MWI and ML results for both antennas to their radiating characteristics parameters, such as antenna cross-section, radiation pattern, and phase center instability.
25. The numerous classifiers and machine learning techniques have been explored in this study. The datasets for breast cancer and diabetes were collected from the Kaggle website. For the classifiers, the performance evaluation metrics accuracy, precision, recall, f1score, and MSE have been taken into account. According to the findings, the Logistic Regression provided the highest accuracy for the diabetes dataset, at 75.32%. The Gaussian classifier has the best recall, whereas logistic regression offers the best precision and MSE. The best F1-score was produced with GaussianNB and Gradient Boosting. It was found that the SVC provided the highest accuracy for the breast cancer dataset, i.e. 96.67% and the lowest MSE, F1 score, and best precision. In this instance, the linear regression classifier has the highest recall.
26. This paper, expanded an ensemble learning-based model to diagnose breast cancer, by the synthesis of statistical estimates from many machine learning classifiers. The XGBoost classifier, an ensemble learning technique, improved the performance of the classifiers by combining the individual predicted probabilities from many classifiers. We did statistical data preparation, which comprised feature selection, data augmentation, and data scaling. Our thorough analysis of a reference dataset amply demonstrates the immense potential of this integrated method. For the purpose of enhancing our model, we intend to use CNN in the future in an optimum manner.
27. This paper suggested an AdaBoost and Random Forest integration model for breast cancer diagnosis. The

model used Grid Search CV to automatically look for hyper parameters while choosing parameters, preventing over- and under-fitting. Nevertheless, multiple studies have shown that in terms of BC detection, this integrated model performed better than typical single algorithm models. The combined model exhibits good prediction performance by achieving 98.6% accuracy, 99% precision, 98% f1-score, and 99% recall rate. Even though machine learning has made some progress in disease prediction, accuracy still has to be increased. However, in terms of training time, interpretability, the improved RF model based on grid search still has better research potential enabling clinicians to diagnose patients more quickly and accurately.

28. By removing high-level characteristics from breast histopathology pictures, the efficacy of the diagnostic technique can be improved. This project's primary goal is to employ VGG16, a pre-trained model from CCN deep learning, to extract high-level features from breast picture. To do that, we deleted the final layer in VGG16 that was fully connected. Following that, a group of heterogeneity classifiers were used to categorize the obtained features. Using the publicly available Break His dataset for extensive testing, a set of performance metrics for performance assessment were created. The experimental results outperformed a number of state-of-the-art procedures. Future research will concentrate on creating a group of diverse classifiers and trained models that can perform well in this difficult environment.
29. In this paper, an innovative method for locating biomarker genes that uses deep learning. Using the suggested framework, a set of 54 differentially expressed and physiologically important genes that enable the identification of breast cancer subtypes have been discovered. Within the setting of multi-omics data, we want to investigate the coherence and/or variation in gene finding for the classification of breast cancer subtypes in future study. In addition, we want to identify new subtypes of cancer by segmenting based on whole-transcriptome sequencing data, cancer heterogeneity. In the future, we also intend to investigate whether the suggested framework can be used to treat other cancer types. Additionally, future investigation may look into the possibility of identified biomarkers for developing Pharmacological therapies.
30. The model in this research was validated using the suggested approach's k folds cross validation method, and the optimal hyper parameters were selected. Model performance has been assessed using performance evaluation metrix. The BC data sets have been used to test the suggested methodology. The experimental results demonstrate that the choice of characteristics is crucial for the timely and effective detection of BC. The proposed technique achieved 99.91% accuracy on the Relief FS feature selection algorithm, resulting in outstanding accuracy results. Additionally, compared to how well SVM performs on features chosen by the Relief algorithm, it performs poorly on features chosen by auto encoder and PCA. Hence, the suggested approach TABLE 1. Relief- Support vector in this essay, mathematical symbols and notations are utilized. Using a system to diagnose BC is highly recommended. The proposed method outperforms the current state-of-the-art method in terms of accuracy. Additionally, we utilized McNemar's statistical test to evaluate the effectiveness of our models. The innovative aspect of the proposed work is the development of a BC diagnosis method using machine learning classification and feature selection approaches. First, a suitable FS algorithm was used to select the key features, and the SVM classifier produced results with high accuracy. Second, data sets with little bearing on BC prediction have successfully been cleaned of the weak features. Thirdly, when compared to the WDBC data set, the WBC data set is more suited, and the classifier SVM worked well Last but not least, the SVM- based approach to detecting BC is more effective. The suggested method for diagnosing BC could also be easily implemented into the healthcare system. Future research will concentrate on enhancing BC detection through the application of extra features, selection methods and extra BC data sets. BC detection will also make advantage of deep learning techniques.
31. One of the cancer that causes the most fatalities each year is breast cancer. Only a small number of reliable prognostic and predictive variables are currently employed in clinical practice to manage breast cancer patients. In this case, the Clustered with Level Set technique can be used to find affected cell shapes with high accuracy and precise labelling of the observed outlines. By choosing the best attributes, the suggested system improves mammography retrieval performance. For image segmentation, the Fuzzy- C-means (FCM) clustering has been employed. Each data point is determined by the aim function is a member of numerous clusters into variable degrees. Principal Component Analysis (PCA) and Multi-level Discrete Wavelet Use of transformation to fully analyze the segmented region. Future enhancements to the system could include other capabilities like the suggestion of drugs or therapies based on the nature of the illness. Doctors can more effectively detect and treat the disease with the help of this prediction and decision process.
32. This study uses a deep learning technique to distinguish between normal and atypical breast cancer using the VGG16 and ResNet50 networks. Precision, accuracy rate, and recall were the three performance metrics used to assess the classification algorithms. VGG16 had the highest classification accuracy score with 94%. The aberrant photos can be divided into benign and malignant tumors for future research. That is really beneficial for carrying out the patients' subsequent procedure.
33. Recently, ML has started to gain popularity in the oncology sector for the creation of prognostic classification models of cancer progression and survival. We think that a combination approach employing machine learning techniques and RO

models would have the ability to increase model precision by evaluating the relative importance of attributes. The proposed approach actually seeks not only decision but also model understandability, which—along with the use of a real-world BC dataset— represents the novel aspect of our research. This is consistent with the current trend. Prior to implementing any machine learning strategy in clinical practice, multicenter prospective studies are necessary for validating and effectively implementing privacy concerns with regard to digital EHR data..

34. In conclusion, during the course of their medication, cancer patients accumulate numbers of clinical notes, and human evaluation of each file is expensive, error-prone, and impractical. Because of this, clinical note data—including patient experience and symptoms—rarely undergoes analysis that might be useful for both patient treatment and research. It is possible to construct NLP and machine learning techniques, like the CRF model we propose here, for use in scenarios for which data are not already available.
35. Our logistic regression and linear discriminant analysis perform better when inputs are more varied, and neural network models performed better at predicting the five-year risk of breast cancer than the BCRAT. These findings suggest that additional easy-to-obtain inputs can also improve the Gail model's ability to predict breast cancer risk, in contrast to other studies that suggested only inputs produced by expensive and/or medical procedures could improve the Gail model's performance. Our models could serve as the basis for innovative, low-cost, non-invasive tools to direct and prompt screening as well as short- and long-term preventative actions with the potential to improve early detection and decrease the incidence of breast cancer.

Table 1 Summary of Performance Metrics Calculated for Different Classifiers Used.

References	Classifiers	Accuracy	Sensitivity	Precision	F1- Score
Rawal, Ramik [10]	SVM	97.13%	97%	96.90%	97.08%
Chhaya Gupta [4]	KNN	97.61%	97.31%	95.82%	97.7%
V. Chaurasia [7]	CNN	98.40%	98%	97.9%	97.80%
Samer Hamed [8]	Random Forest	99.22%	97.98%	96.98%	95.8%
D.Yifan [27]	Adaboost	98.6%	99%	99%	98%
Shilpa M [10]	Naïve Bayes	99.81%	97.80%	98.8%	97.6%
Manav Mangukiya [1]	XGboost	98.24%	98.6%	98.76%	96.21%
Chhaya Gupta [4]	Logistic Regression	97.89%	100%	90.69	97.35
Islam MM [2]	ANN	98.57%	96.89%	91.67	98.05

IV. CONCLUSION

Among all cancers, one of the most prevalent types is breast cancer. Nearly 12% of those affected by this malignancy were women. Breast cancer currently affects women of all ages because of the tiny cells' rapid growth and unchecked progression. Medical professionals frequently employ machine learning (ML) techniques, which can be useful as a diagnostic tool and in helping them analyze already- available data. In order to save a precious life, it is therefore preferable to do early breast cancer diagnostics. Using algorithms, it has been found that the accuracy is greater than 94% for identifying malignant tumours and benign tumours. In comparison to other algorithms, KNN has been found to have the highest accuracy, F1 score, and precision, making it the best at spotting breast cancer. Therefore, ML techniques that employ supervised learning are very helpful in cases of early prognosis and diagnosis in the study of cancer.

REFERENCES

[1]. Mangukiya, Manav. “Breast Cancer Detection with Machine Learning.” International Journal for Research in Applied Science and Engineering Technology 10, no. 2 (n.d.): 141–45. doi:10.22214/IJRASET.2022.40204.
[2]. Islam MM, Iqbal H, Haque MR, Hasan MK. Prediction of breast cancer using support vector machine and K- Nearest neighbors. In: Proc. IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Dhaka, 2017, pp. 226–229.
[3]. Chowdhury, Anoy. (2020). Breast Cancer Detection and Prediction using Machine Learning. 10.13140/RG.2.2.23969.84320.
[4]. Chhaya Gupta , Kirti Sharma, 2022, Early Breast Cancer Detection using Various Machine Learning Techniques, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 11, Issue 06 (June 2022)
[5]. A.reddy, Support Vector Machine Classifier For Prediction Of Breast Malignancy Using Wisconsin Breast Cancer Dataset, J. Artif. Intell. Mach. Learn. Neural Netw., vol. VII, no. 21, pp. 18, 2022, doi: 10.55529/jaimlnn.21.1.8.
[6]. L. Liu, Research on logistic regression algorithm of breast cancer diagnose data by machine learning,

Proc. – 2018 Int. Conf. Robot. Intell. Syst. ICRIS 2018, pp. 157160, 2018, doi: 10.1109/ICRIS.2018.00049.

[7]. V. Chaurasia and S. Pal, Applications of Machine Learning Techniques to Predict Diagnostic Breast Cancer, SN Comput. Sci., vol. 1, no. 5, 2020, doi: 10.1007/s42979-020-00296-8.

[8]. Silpa, P. & Reddy, A. & Durga, Ch & Sri, V. & Bhavani, U. & Priya, Ch & Mounika, J.. (2022). Designing of Augmented Breast Cancer Data using Enhanced Firefly Algorithm. 759-767. 10.1109/ICOSEC54921.2022.9951883.

[9]. Kapadne, Seeyog & Patne, Arsh & Chaware, Jahan & Priyanka, Prof. (2022). Comparative Analysis of Early Stage Cancer Detection Methods in Machine Learning.

[10]. Rawal, Ramik. (2020). BREAST CANCER PREDICTION USING MACHINE LEARNING. 7.

[11]. S. Sharma, A. Aggarwal and T. Choudhury, "Breast Cancer Detection Using Machine Learning Algorithms," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 2018, pp. 114-118, doi: 10.1109/CTEMS.2018.8769187.

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

[13]. E. A. Bayrak, P. Kırcı and T. Ensari, "Comparison of Machine Learning Methods for Breast Cancer Diagnosis," 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), 2019, pp. 1-3, doi: 10.1109/EBBT.2019.8741990.

[14]. R. MurtiRawat, S. Panchal, V. K. Singh and Y. Panchal, "Breast Cancer Detection Using K-Nearest Neighbors, Logistic Regression and Ensemble Learning," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 534-540, doi: 10.1109/ICESC48915.2020.9155783.

[15]. S. Ara, A. Das and A. Dey, "Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms," 2021 International Conference on Artificial Intelligence (ICAI), 2021, pp. 97-101, doi: 10.1109/ICAI52203.2021.9445249.

[16]. S. Gc, R. Kasaudhan, T. K. Heo, and H.D. Choi, "Variability Measurement for Breast Cancer Classification Mammographic adaptive and convergent systems (RACS), Prague, Czech Republic, 2015, pp. 177–182.

[17]. P. P. Sengar, M. J. Gaikwad and A. S. Nagdive, "Comparative Study of Machine Learning Algorithms for Breast Cancer Prediction," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 2020, pp. 796-801, doi: 10.1109/ICSSIT48917.2020.9214267.

[18]. S. Hafizah, S. Ahmad, R. Sallehuddin, and N. Azizah, "Cancer Detection Using Artificial Neural Network and Support Vector Machine: A Comparative Study," J. Teknol, vol. 65, pp. 73–81, 2013.

[19]. Shilpa M, C. Nandini "Breast Cancer Diagnosis and Prediction Using Machine Learning Algorithm" International Journal of Science and Research (IJSR) Volume 9 Issue 4, April 2020.

[20]. Prerita, N. Sindhwani, A. Rana and A. Chaudhary, "Breast Cancer Detection using Machine Learning Algorithms," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2021, pp. 1- 5, doi: 10.1109/ICRITO51393.2021.9596295.

[21]. S. Ara, A. Das and A. Dey, "Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms," 2021 International Conference on Artificial Intelligence (ICAI), 2021, pp. 97-101, doi: 10.1109/ICAI52203.2021.9445249.

[22]. M. S. Harinishree, C. R. Aditya and D. N. Sachin, "Detection of Breast Cancer using Machine Learning Algorithms – A Survey," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 1598-1601, doi: 10.1109/ICCMC51019.2021.9418488.

[23]. Hirra et al., "Breast Cancer Classification From Histopathological Images Using Patch-Based Deep Learning Modeling." in IEEE Access, vol. 9, pp. 24273-24287,2021, doi: 10.1109/ACCESS 20213056516.

[24]. R. A. Martins, J. M. Felício, J. R. Costa and C. A. Hemandes. "Comparison of Slot-based and Vivaldi Antennas for Breast Tumor Detection using Machine Learning and Microwave Imaging Algorithms," 2021 15th European Conference on Propagation (EuCAP), 2021, pp. 10.23919/EuCAP51087.2021.9411128. Antennas and 1-5, doi:

[25]. Bansal and A. Singhrova. "Performance Analysis of Supervised Machine Learning Algorithms for Diabetes and Breast Cancer Dataset," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 10.1109/ICAIS50930.2021.9396043. 137-143, doi:

[26]. M. S. K. Inan, R. Hasan and F. I. Alam. "A Hybrid Probabilistic Ensemble based Extreme Gradient Boosting Approach For Breast Cancer Diagnosis," 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), 2021, pp. 1029- 1035, doi: 10.1109/CCWC51732.2021.9376007.

[27]. D. Yifan, L. Jialin and F. Boxi. "Forecast Model of Pareign Based on RF-AdaBoost."

[28]. D. Albashish. R. Al-Sayyed. A. Abdullah, M. H. Ryalat and N. Ahmad Almansour. "Deep CNN Model based on VGG16 for Breast Cancer Classification," 2021 International Conference on Information Technology (ICIT). 2021, PP. 805-810, doi: 10.1109/ICIT52682.2021.9491631.

[29]. S. Rajpal, M. Agarwal, V. Kumar, A. Gupta and N. Kumar, Triphasic DeepBRCA-A Deep Learning-

Based Framework for Identification of Biomarkers for Breast Cancer Stratification," in IEEE Access, vol. 9, 103347-103364, 1109/ACCESS 20213093616. 2021, doi:

[30]. A.U. Haq et al., "Detection of Breast Cancer Through Clinical Data Using Supervised and Unsupervised Feature Selection Techniques," in IEEE Access, vol. 9, pp. 22090-22105, 10.1109/ACCESS 2021.3055806. 2021, doi:

[31]. Sivapriya J, Aravind Kumar V, Siddarth Sai S, Sriram S. Breast Cancer Prediction using Machine Learning. ISSN: 2277-3878, Volume-8 Issue-4, November 2019

[32]. N. S. Ismail and C. Soyuthy, "Breast Cancer Detection Based on Deep Learning Technique," 2019 International UNIMAS STEM 12th Engineering Conference (EnCon). 2019, pp. 89-92, doi: 10.1109/EnCon 2019.8861256.

[33]. M. Shahbaz, S. Faruq, M. Shahen, and S. A. Masood,Cancer detection using data mining technology, Life Sci. J., vol. 9, no. 1, pp. 308-313, 2012.

[34]. Tüba Kiyanand Tülay Yildirim (2004). Breast cancer diagnosis using statistical neural networks, Journal of electrical & electronics engineering, vol.4, pp.1149-1153.

[35]. V. Sugumaran V Muralidharan and K.I. Kamachandran (2007), Feature selection using Decision Tree and classification through Proximal Support.