ISSN : 1001-1781
Journal of Clinical Otorhinolaryngology, Head, and Neck Surgery

BREAST CANCER SCREENING ML MODEL TO IMPROVE MAMMOGRAPHIC

Ashvini Alashetty[1], G.G. Rajput[2],
[1]ashwinialashetty@gmail.com
[2]ggrajput@yahoo.co.in
[1,2] Department of computer science, Akkamahadevi women's University,
Vijayapura, Karnataka-586108

Abstract

Breast cancer screening is critical for detecting the disease at an early stage, when it is most treatable. A machine learning model for breast cancer screening that was trained using Kaggle's Mammographic Mass dataset. Mammogram characteristics of breast masses, such as size, location, and density, are among the clinical features included in the dataset. The proposed study trained the model using a number of distinct machine learning (ML) classification techniques. Logistic regression (LR), support vector machines (SVM), random forests (RF), and the Voting classifier were all used.A few of the metrics used to assess algorithms are accuracy, precision, recall, and F1-score. With a score of F1 is 0.86 and an accuracy of 86.1%, the proposed work concludes that the voting classifier technique outperforms the other algorithms. Breast cancer screening could benefit from this model by assisting radiologists and physicians in making more precise and timely diagnoses, resulting in better patient outcomes.

Keywords: Breast Cancer, Convolutional Neural Network, Voting Classifier, Support Vector Machine (SVM), Logistic Regression, and Random Forest Classifier.

1.Introduction

Mammography is extensively used as a breast cancer screening procedure, which is crucial since detecting the illness at an early stage considerably enhances the odds of survival. Mammography is a non-invasive medical imaging method that uses low-dose X-rays to create images of breast tissue. Mammography is a proven tool for the early detection of breast cancer. It can detect breast cancer before any physical symptoms appear, making it an important tool for early intervention. Squeezing the breast between two plates to provide a clear image is part of the procedure. A radiologist examines mammography images for evidence of abnormalities such as lumps or tumors [1-4]. The accuracy of mammography in detecting breast cancer has long been debated. False negatives are possible, which means that a mammogram may indicate the absence of cancer when it is actually present. False positives are also possible, which means that a mammogram may indicate the presence of cancer when none exists. Despite these limitations, mammography remains the most accurate tool for detecting breast cancer. Digital mammography has boosted diagnostic certainty and lowered false-positive rates. The advent of digital technology in mammography has facilitated the detection of breast tissue abnormalities by radiologists [5-8]. In recent years, researchers have looked into the use of computer-aided detection (CAD) technologies to improve mammography accuracy. CAD systems use algorithms to analyses mammograms and identify any areas of concern for breast cancer. The advancement of digital mammography and computer-aided detection systems has

improved mammography accuracy while decreasing the number of FP and FN. We can expect further advancements in the accuracy and effectiveness of mammography in breast cancer detection as research in this field continues [9-11].

1.1 Motivation

Mammography is now considered as the effective screening method for detecting the breast cancer. Because of abnormal or ambiguous results, 5 to 10% of mammograms are referred for further evaluation via biopsy. Machine learning (ML) techniques are being investigated as a potential improvement for these scenarios. The goal of this work is to reduce screening referrals by 5 to 10% by combining the Breast Imaging-Reporting and Data System (BI-RADS) and other mammographic characteristics with ML techniques. Machine learning has the potential to improve mammographic screening by analyzing massive amounts of data to detect patterns and identify previously overlooked areas of concern. The BI-RADS, a standard classification system for characterizing mammographic data, could serve as the foundation for machine learning (ML) algorithms that identify characteristics that are associated with the presence of breast cancer. Machine learning (ML) algorithms may be able to make more accurate predictions about the presence of breast cancer by analyzing these traits, reducing the number of screening referrals. The study's emphasis on the use of ML techniques to improve mammographic screening accuracy is a significant step towards reducing the burden of unnecessary screening referrals, biopsies, and the associated costs and anxiety for patients. Furthermore, improving mammographic screening precision may result in earlier detection of breast cancer and better patient outcomes [12-14].

2 Literature Review

Mammographic screening is one of the most successful methods for detecting breast cancer in its early stages, according to X. Jia et al. [15]. Mammograms continue to be the most common and effective screening method for early breast cancer detection. Biopsies of the breast are performed in response to ominous symptoms that often out to be noncancerous, despite the diagnostic procedure's very low predictive accuracy. Following in the footsteps of Decision Trees (DTs), Random Forest (RF) is one of the most effective ensemble learning systems (or meta estimators). The Breast Imaging Reporting and Data System (BI-RADS) was established in order to standardise the process of reporting mammograms. This technique is used to categorise unusual discoveries. The purpose of this study was to predict cancer biopsy results using BI-RAD findings and patient age, and the RF classifier using Chi-Square (2) and Mutual Information (MI) techniques of acceptable Feature Selection (FS) was put to good use. For the UCI Mammographic Mass dataset, accuracy, area under the curve (AUC), and other performance parameters were evaluated using a 10-fold cross-validation (CV) approach. The proposed technique showed tremendous promise in its capacity to predict outcomes (84.70 percent accuracy, 0.9023 ROC area under the curve). MCC and F1-scores were also increased as a consequence of the recommended technique. Using RF classifiers and other cutting-edge classification techniques, the outputs were evaluated. Indicators of performance tested in this study demonstrate that the proposed model outperforms the RF classifier and the baseline approaches.These results provide credence to the precision and utility of MI FS techniques for collecting the minimal set of data necessary for categorization. In addition, the findings demonstrate analogies between the proposed classification approach and current classification techniques in the literature. This technology is capable of achieving the realistic objective of

predicting the success of biopsies on cancer patients. According to Suman Mann et al. [16], breast cancer is a prevalent disease that affects women of all ages, with the risk increasing with time. Even with improved technologies for diagnosis and treatment, early detection of cancer remains essential for increasing survival rates. Using the non-subsampled contour let transform (NSCT) and Z-moments for feature extraction, a classification algorithm for mammography images is proposed in this article. Support vector machine (SVM) classification is used to determine if a mammogram is normal, benign, or cancerous.The proposed strategy halved training time while enhancing precision to 96.76 percent. Evaluations of feature extraction by morphological spectroscopy using 16 algorithms and four classification strategies shown greater accuracy and efficiency in terms of time when compared to existing methodologies.Hala Al-Shamlan et al. [17] discovered 23 important parameters that have been shown to have a significant effect in mammography images. These features were chosen after extensive consultation with mammography specialists and a thorough review of the field's published literature. When these characteristics were applied to a total of 80 mammograms, the overall result was rather encouraging. For each feature extraction, the experimental results revealed the predicted range values. Our findings support further research into the application of these feature extractions to the development of a novel computer-assisted breast cancer diagnostic and classification system. There are several approaches for detecting breast cancer, according to Anusha Bharat et al [18]. According to the data, the K-Nearest Neighbors (KNN) method outperformed the other algorithms studied. Furthermore, both the Naive Bayes and Logistic Regression models produced acceptable results. Support Vector Machine (SVM) analysis was also shown to be highly successful. The support vector machine (SVM) with a Gaussian kernel was determined to be the most accurate approach for predicting the recurrence of breast cancer after extensive testing. Because binary class variables are not supported, this SVM can only handle two classes. To overcome this limitation, academics have created multi-class SVM algorithms such as LIBSVM. Furthermore, the precision of these algorithms can be improved by precisely setting the algorithm's parameters. Furthermore, this technology could be deployed on a cloud-based platform, making it far more user-friendly.

According to Naresh Khuriwal et al [19], the classification of cancer diagnostic image collections can benefit from the use of CNNs. Using a total of 12 criteria, this study studies the applicability of CNNs to the identification of breast cancer. The results of this specific deployment were 98% accurate. This study will improve the accuracy of cancer detection. However, there is room for improvement in this area, including the addition of new attributes and real-world image collections, as well as research into the application of this technique to a wide range of tumor types. Qasem et al [20] use Multi-scale Center-weighted Snakes (MCWS) and Support Vector Machines (SVM) methodologies to reduce false positive rates and extract the initial contour, respectively. As a result of this integration, a fully automated system with greater precision than ground truth markers has been created. They discovered that the hybrid approach can locate masses using mammograms. The process of making a medical diagnosis and consulting with a patient, according to Al-Hadidi et al [21], For feature extraction, image processing is used, and ML techniques used to carry out the methodology. We employ the supervised learning techniques of LR (Linear Regression) and BPNN in this study (Back-propagation Neural Network). The results revealed that the LR model used a broader range of features than the BPNN model. Despite using fewer features than competing models, the BPNN

model achieved a high regression value of over 93%. This finding suggests that BPNN could be useful in such situations. According to Using microwave imaging, which generates no radiation and is thus safe for patients, H. Sami et al. [22] have created a novel method for detecting breast cancer. In biological applications, such as illness diagnostics, machine learning has been shown useful. In this work, breast lesions are predicted from raw backscattered signal data using the support vector machine (SVM) method using linear and polynomial kernels. Traditional machine learning binary classification techniques were surpassed by an SVM strategy based on a third-degree polynomial kernel. It may be able to aid radiologists in the early diagnosis of breast cancer if tumours can be precisely anticipated.

L. Wang et al. [23] contend that early breast cancer detection is crucial for decreasing mortality rates. Mammography and biopsies are just two of the numerous diagnostic approaches that have been studied, but they both have limitations that demonstrate the need for a more sensitive and time-efficient method. In recent years, multi-biomarker biosensors for early breast cancer detection have gained popularity. Microwave imaging has also been considered as a viable and inexpensive tool for diagnosing breast cancer. This study aimed to examine the most recent breast cancer screening techniques, with a focus on microwave imaging, biomarkers, and biosensors for early detection. Breast Microwave Sensing (BMS) is a method that employs low-power microwave radiation to detect breast cancer; T. Reimer et al [24] have studied it. In the study's portable BMS prototype, twelve sensors working at five separate frequencies (between 2.30 GHz and 6.50 GHz) were used. Five distinct feature preparation methods were utilised to simulate data and assess it using numerical 2D phantoms representing BI-RADS Classes 1 and 2. In this study, it was determined that both a multilayer perceptron (MLP) and a support vector machine with a radial basis function (SVM RBF) performed equally well in classifying simulated data.As described by R. C. Conceico et al. [25], phantoms representing benign and malignant breast tumours were distinguished using a prototype of a preclinical Ultra Wideband (UWB) imaging system. Thirteen benign and thirteen malignant tumour phantoms were created to approximate the dielectric properties of cancerous tissues in the 1-6GHz frequency range. Using a machine learning technology called Support Vector Machines, the ghosts in this study were identified (SVM). The findings of this study were compared to those of an earlier study by the same authors on Linear and Quadratic Discriminant Analysis. Using a Tactile Imaging System (TIS) and machine learning techniques, V. Oleksyuk et al. [26] identified breast tumours as malignant or benign in vivo. The TIS algorithm use the indentation profile recorded by the TIS probe to calculate the mass's size and rigidity. On the basis of their observed size and stiffness, breast tumours may be categorisedutilising artificial intelligence systems. The k-nearest neighbour classifier outperformed the support vector machine and the Naive Bayes technique on a dataset of 12 human breast tumours. The data were examined using leave-one-out cross-validation. This study illustrates the use of TIS in combination with machine learning methods for the in vivo categorization of breast tumours. Medical microwave imaging is a promising new method, particularly for the detection of breast cancer, according to B. Gerazov et al [27]. Moreover, it has been shown that the backscatter signals give a good foundation for distinguishing between malignant and benign tumour types.Deep learning methods are used to a set of findings from numerical simulations in the Finite Difference Time Domain (FDTD) of cancer models implanted in homogenous breast adipose tissue. Our usage of Deep and Convolutional Neural

Networks achieves 93.44 percent more accuracy than typical machine learning techniques applied to the analysed dataset. Comparing the dielectric properties of healthy and diseased tissues, S. P. Rana et al. [28] state that contemporary microwave technology may safely and non-ionizingly detect breast lesions by comparing healthy and diseased tissues. The Department of Diagnostic Imaging at Perugia Hospital included the technology into their breast exam patient data gathering methods. In this research, conventional breast examinations are compared to those performed using microwave ultra-wideband (UWB) equipment. A portable transmitting and receiving antenna collects the S-parameters of microwave signals delivered through breast tissue. The same individuals who preprocess data for traditional radiologists offer microwave data for the AI system. This data is utilised to train and assess a variety of supervised machine learning approaches, including nearest neighbour (NN), multi-layer perceptron (MLP) neural network, and support vector machine (SVM), in order to develop a robust classification system for recognising breast lesions. Comprehensive study and statistical validation led to the conclusion that the quadratic kernel of the support vector machine (SVM) was the most effective for classifying breast lesions.According to F. A. Spanhol et al. [29], direct comparison is impossible due to the fact that data from multiple sources (institutions, scanners, demographics, etc.) that evaluate their work using varied criteria makes direct comparison impossible. This page provides access to 7909 histological images of 82 patients with benign and aggressive breast cancer. In order to provide physicians with a useful computer-assisted diagnostic tool, this dataset intends to enable the automated classification of these images into two groups. As an idea of the difficulty of the task at hand, the authors offer preliminary data demonstrating an accuracy range of 80 to 85 percent using cutting-edge image categorization algorithms. There is room for growth in this kind of categorisation. Researchers in machine learning and medicine may collaborate on the development of this therapeutic application provided the authors offer this dataset and a standardised evaluation method.

3. Proposed Method

3.1 Dataset

Table 1Mammography Mass Dataset and type of variables

Variable	Definition	Type
Score	BI-RADS: 0-5	Ordinal
Age	Patient's age in years	Continuous
shape	Mass shape:round=1, oval=2, lobular=3 and irregular=4	Categorical
Margin	Mass margin: circumscribed=1, micro-lobulated=2, obscured=3, ill-defined=4, and spiculated=5	Categorical
Density	mass density: high=1. iso=2. low=3 and fat containing=4	Ordinal
Malignant	Benign=FALSE and malignant=TRUE	Boolean

We will do some high-level studies, develop ML pipelines for basic ML model comparison, and analyse the usage of Response Curves, also known as Partial Dependency Plots, to examine

the link between various attributes and responses. The PDPs reflect how well the model matches the data with regard to individual characteristics. As a result, they show how useful the model is from two different perspectives. Table 1 depicts the dataset and variable types [30, 31]. Importing libraries, loading datasets, checking for null values, renaming columns, and rearranging column names are all important steps in ensuring that the data is clean, well-organized, and ready for analysis. The first step in the data preparation process is to import the necessary libraries. Libraries are collections of tools and functions for modifying and analyzing data. The first step in the data preprocessing process is to import libraries. Depending on the type of analysis, it is possible that multiple libraries will be required. Python's panda library, for example, is frequently used for data manipulation, while the scikit-learn library is used for machine learning. Both are included in the standard Python distribution [32-34]. After importing the prerequisite libraries, the next step is to load the datasets that will be used for the analysis. This involves reading data from an external source, such as a CSV file or a database, and converting it into a format that the program can use. In pandas, this is commonly accomplished by using the read csv methods. Following the import of the datasets, it is critical to check for any null values, which are data points that are either missing or have not been declared. Because the use of null values in analysis and modelling can lead to errors, it is critical to recognize and deal with them correctly. This is possible in Pandas by using functions like isnull and fillna [35-37].

The next step is to rename the columns so that they match the data dictionary. A data dictionary is a document that describes the purpose and structure of each column in a dataset. Renaming columns during data analysis and reporting helps to maintain consistency and clarity in the data. The pandas rename function allows you to rename columns in an existing dataset. It is possible that reordering the columns will be required to improve the readability and structure of the data. When working with huge datasets including a significant number of columns, this may be a crucial factor. The re-index function in pandas allows you to rearrange the order of the columns. In a nutshell, these processes are required in order to clean the data and prepare it for analysis. Importing libraries, loading datasets, checking for null values, renaming columns, and rearranging columns allow data analysts and data scientists to ensure that the data is correct, consistent, and ready for analysis. This contributes to the reduction of errors, the improvement of insights, and, ultimately, the generation of better decisions [38-40].

3.2 Random Forest Classifier

Random Forest Classifier is a potentially beneficial machine learning model for breast cancer diagnosis. The algorithm accomplishes its objective by constructing a forest of decision trees and averaging their individual judgements. Given a database of mammograms and clinical data, the Random Forest Classifier may be trained to determine if a patient has breast cancer or not. This enables the classifier to detect with certainty whether a person has breast cancer. Among the characteristics that may be included into the algorithm are the size and shape of lesions, breast tissue density, patient age, and medical history. The Random Forest Classifier has many advantages for breast cancer diagnosis. In contrast to too complicated models that perform well on training data but poorly on new data, this one avoids overfitting.This sort of model may develop when a model is excessively sophisticated and performs well on training data but badly on fresh data. In addition, the algorithm may offer information on the relative relevance of a number of factors, which can assist in the determination of crucial elements of

breast cancer risk and diagnosis. Overall, the Random Forest Classifier is a valuable tool in the battle against breast cancer, and its usage in combination with other diagnostic techniques has the potential to enhance early diagnosis and treatment results.

3.3 Voting Classifier

This sort of model may develop when a model is excessively sophisticated and performs well on training data but badly on fresh data. In addition, the algorithm may offer information on the relative relevance of a number of factors, which can assist in the determination of crucial elements of breast cancer risk and diagnosis. Overall, the Random Forest Classifier is a valuable tool in the battle against breast cancer, and its usage in combination with other diagnostic techniques has the potential to enhance early diagnosis and treatment results. Before reaching a conclusion, the Voting Classifier considers all of the predictions made by the individual classifiers. These predictions are based on the unique sets of features and parameters used by each classifier. When it comes to breast cancer screening, the Voting Classifier has several advantages. By combining the results of multiple classifiers into a single model, the accuracy and reliability of the overall forecast can be improved. It can also help to alleviate the limitations of specific algorithms by taking into account a broader range of characteristics and parameters than such algorithms would normally consider. Overall, the Voting Classifier is an important tool in the fight against breast cancer, and its use in conjunction with other diagnostic methods can aid in early detection and treatment outcomes.

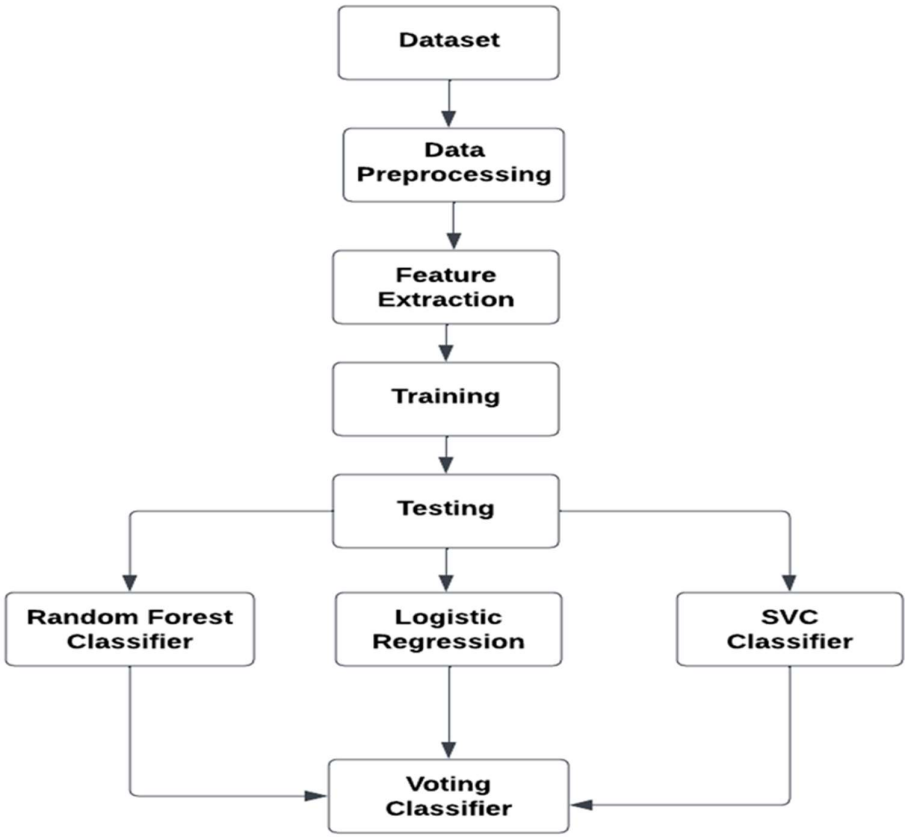


Fig 1: Block diagram

In Fig. 1block diagram, the output the Random Forest classifier is fed into the Voting Classifier, which aggregates all three classifiers' predictions. The Voting Classifier generates the

conclusive forecast. The Voting Classifier consists of two classifiers: Logistic Regression and SVC. Each classifiers provide their own predictions, which are then aggregated by the Voting Classifier to produce the final forecast. Classifiers based on Random Forest, SVC, and Logistic Regression may be trained independently utilizing distinct features and hyper parameters. Using the different classifiers and their weights, the Voting Classifier may be trained to maximize the accuracy of the final prediction.

3.4 Support Vector Classifier (SVC)

In breast cancer diagnosis, mammographic image analysis is an appealing use case for the Support Vector Classifier (SVC). The SVC can differentiate between cancerous and benign regions of interest (ROIs) in an image by analysing texture, shape, and size characteristics. Several research have shown that the SVC is beneficial for analysing mammograms. In one study, classification of ROIs in mammographic images using an ensemble of SVCs with changing parameter values yielded good accuracy, sensitivity, and specificity. In another study, categorization of mammography images was improved by combining the SVC with a feature selection strategy. In addition, the SVC may be used in conjunction with other imaging modalities, such as ultrasound and MRI, to diagnose breast cancer (MRI). In one study, for example, the SVC was used to classify breast lesions using data from both mammography and ultrasound images, resulting in high diagnosis accuracy.

3.5 Logistic Regression

For binary classification problems, such as diagnosing breast cancer, logistic regression is a popular machine learning technique. Using deep learning methods, in particular deep neural networks, to predict the probability of breast cancer is an interesting new breakthrough (DNNs). A recent research extracted high-level properties from mammographic pictures using a DNN, which were then input into a logistic regression model for breast cancer classification. Our strategy outperformed conventional feature extraction techniques and generated great diagnostic accuracy. Another research used a similar method, but included a mechanism for transfer learning to increase the performance of DNNs. The DNN must be trained via transfer learning on a large dataset of natural images, and then fine-tuned on a smaller dataset of mammographic images. The resulting logistic regression model performed best when compared to other classifiers. In addition, logistic regression has been used with other feature selection techniques to enhance breast cancer diagnosis. One study determined the most useful characteristics for breast cancer classification using logistic regression and a correlation coefficient-based feature selection approach, yielding high levels of accuracy and specificity.

3.6 Multi-Layer Perception (MLP) Classifier

The Multi-Layer Perceptron (MLP) Classifierhas been successful in breast cancer diagnosis. Using MLP Classifier in conjunction with radiomics analysis, a quantitative study of medical images that extracts variables related to tumor phenotype, microenvironment, and other characteristics, is a new advancement in breast cancer diagnosis. In one work, an MLP Classifier and radiomics analysis were utilised to construct a machine learning model for MRI-based breast cancer diagnosis. The MLP Classifier was used to classify breast tumors as cancerous or benign using radiomic characteristics extracted from MRI images. In terms of diagnostic accuracy and performance, the generated model outperformed other machine learning classifiers. Another study used an MLP Classifier to identify breast cancer in mammography images. Deep features extracted from mammography images were used to train

the classifier. The study found that the MLP Classifier had good diagnostic accuracy, sensitivity, and specificity in detecting breast cancer, indicating the algorithm's clinical utility. Furthermore, the MLP Classifier has been combined with various machine learning algorithms and feature selection approaches to improve its ability to detect breast cancer. For example, one study used the MLP Classifier in conjunction with a feature selection approach based on mutual information to select the most useful features for breast cancer classification, resulting in excellent diagnosis accuracy. MLP Classifier has shown promising results in breast cancer diagnosis, especially when combined with radiomics analysis, deep learning approaches, and other feature selection methods.

4 Experiments and results

4.1 Data Collections

Exploration of data in breast cancer detection may help boost the diagnostic accuracy and effectiveness of breast cancer. By analysing patterns in biopsies, for example, researchers may be able to develop algorithms that identify cancerous cells more accurately than standard diagnostic methods. Therefore, data exploration may play a crucial role in the development of new breast cancer treatments. Researchers may discover potential new drugs and therapy regimens for the illness by examining data from clinical trials. In addition, they may employ machine learning algorithms[41-47] to forecast which patients are most likely to react to certain therapies, allowing for more individualised and successful treatment strategies. Data analysis is a crucial component of breast cancer diagnosis and treatment. Researchers may identify risk factors, biomarkers, and patterns that can aid in the early detection and accurate diagnosis of breast cancer by examining massive datasets using various approaches and technologies. This may subsequently result in more effective treatments and improved patient outcomes. The collection of patient information is shown in Table 1. Table 2 displays the collected patient data. The bar chart plot is shown in Figure 2.

Table 2 Data collections of the patients

BI-RADS	Age	Shape	Margin	Density	Severity
5	67	3	5	3	1
5	58	4	5	3	1
4	28	1	1	3	0

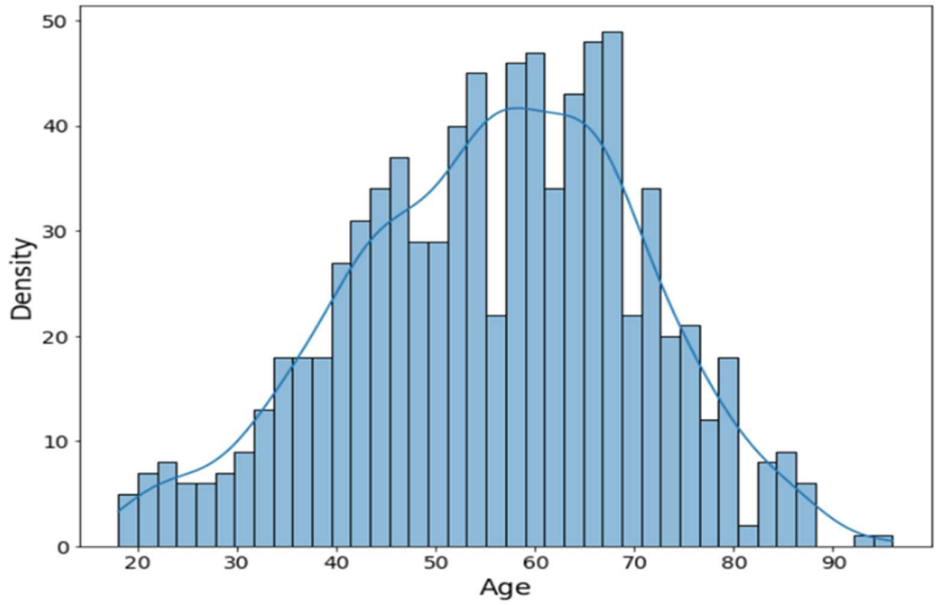


Fig.2Age vs Density

4.2 Building Model

Setting up the model architecture, separating the data into testing and training sets, and scaling the data to guarantee that it is standardized and ready for analysis are all essential phases in building a machine learning model. The logistic regression classifier for voting. This model combines the predictions of various logistic regression models, weighted by their relative accuracy, to provide a final prediction. When numerous models are available and their predictions can be integrated to increase overall accuracy, this strategy may be very useful. After establishing the model architecture, the following step is to divide the data into testing and training sets. This is crucial because it allows us to evaluate the performance of the model on new, unknown data. The training set is used to train the model, while the testing set is utilised to evaluate its accuracy. It is essential to remember that the testing set should be representative of the whole dataset in order to guarantee that the model is assessed on a variety of instances. After dividing the data, the following step is to scale the data. This entails normalizing the data such that its mean is zero and its standard deviation is one. This is essential since it guarantees that all characteristics are evaluated similarly and have an equal influence on the model's predictions. Scaling may help address issues about numerical instability and boost the performance of some machine learning techniques. Developing a model for machine learning requires establishing the model architecture, splitting the data into testing and training sets, and scaling the data. By adhering to these principles, we can guarantee that our models are precise, dependable, and capable of generating predictions on fresh, unobserved data. Fig. 3 shows the comparison chart between Benign and Malignant.

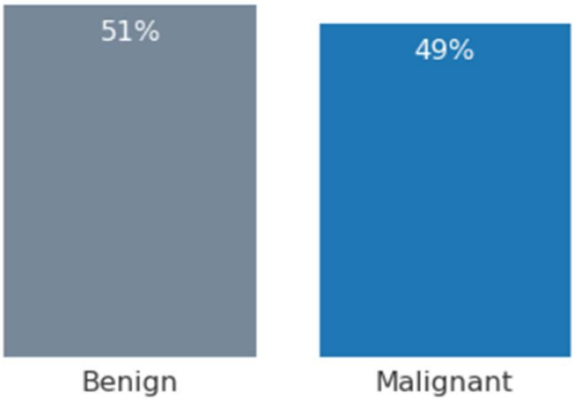
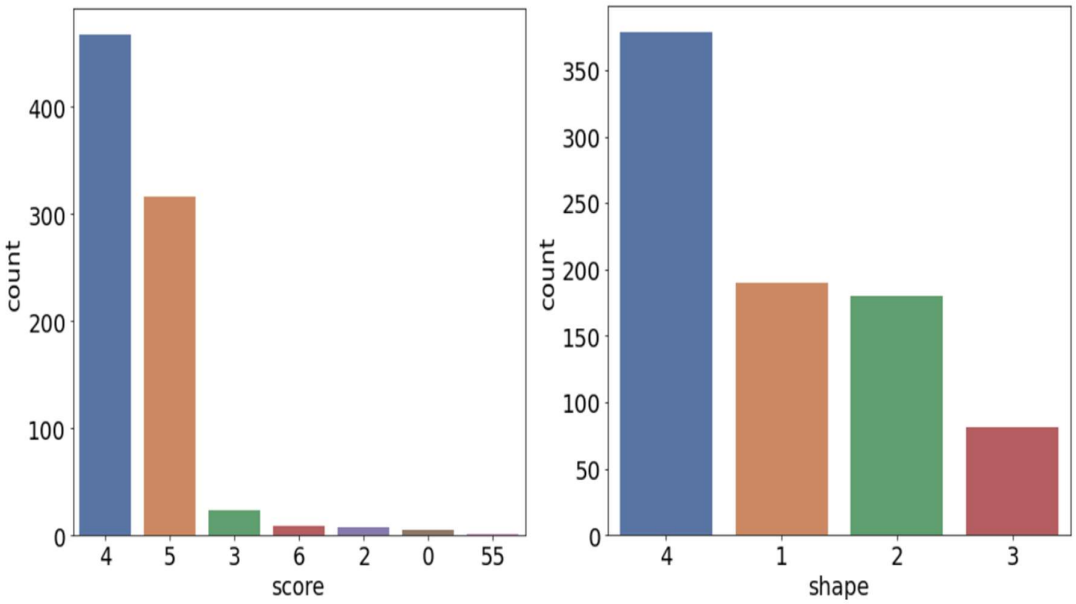


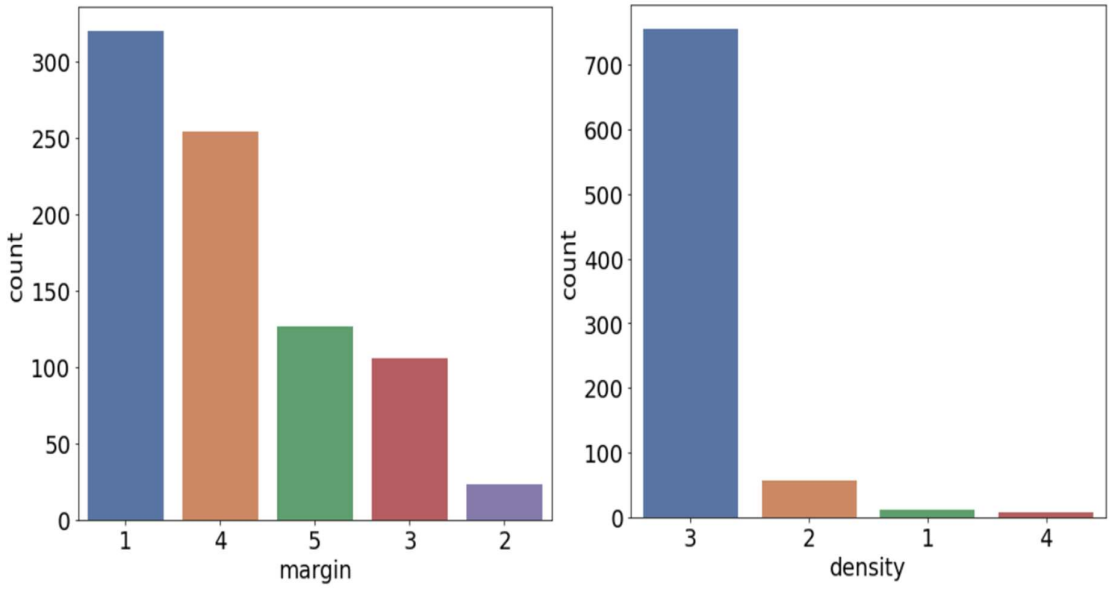
Fig. 3 Diagnostic Outcome

4.3 Categorical variables

Categorical variables are a form of statistical variable that reflect qualitative data, as opposed to quantitative data. They are often used to represent data that may be separated into various groups, such as gender, race, and degree of education. It is possible to further categories categorical variables into nominal and ordinal categories. Ordinal categories have a natural order, such as ranks or degrees of schooling, while nominal categories have no intrinsic order, such as colors or varieties of fruit. Categorical variables play an essential role in several fields of data analysis, including market research, social sciences, and epidemiology, and need specific treatment in statistical analysis to assure correct findings. Fig. 4 depicts the bar chart of the parameters.



4(a) 4(b)



4(c) 4(d)

Fig. 4 (a) score vs count, (b) shape vs count, (c) count vs margin, (d) density vs count

4.3 Scaling Datasets

We now build a baseline model followed by setting up a Voting Classifier

Table 3 Scaled Datasets for a baseline model with a Voting Classifier

Iteration	Score	Shape 2	Shape 3	Margin 2	Margin 3	Density
0	0.99164	-0.50652	-0.32674	-0.17263	-0.36464	0.25612
1	0.99164	-0.50652	-0.32674	-0.17263	2.74236	0.25612
2	0.99164	-0.50652	-0.32674	-0.17263	2.74236	0.25612

Table 3 shows the scaled datasets. To continue, we will fit a two-layer MLP classifier, with each layer containing five nodes. The data shows a significant signal, as seen by the unoptimized MLP model's remarkable 86% accuracy score. There is nevertheless the danger of overfitting, which might result in inaccurate predictions on fresh data. In response, we will undertake some preliminary experiments to determine the optimal number of nodes per layer. In addition, to further enhance the model's performance, Grid Search will be used to discover the ideal collection of hyper-parameters for the MLP model, resulting in improved generalization and more accurate predictions.

	Random Forest Classifier	Logistic Regression	SVC
Precision	0.82	0.89	0.91
Recall	0.81	0.89	0.91
F1 score	0.81	0.85	0.86
Accuracy	0.80769	0.85096	0.85576

Table 4 Nodes and Validation Score

Nodes	Validation score
(3, 3)	0.80547
(3, 4)	0.81672
(3, 5)	0.81190
(4, 3)	0.82154
(4, 4)	0.82315
(4, 5)	0.81350
(5, 3)	0.82797
(5, 4)	0.81029
(5, 5)	0.82154

0.827 is the highest validation score. This is equivalent to nodes (5, 3). Predictably, the model was overfit. The decreased number of nodes indicates that a model with fewer nodes fits the validation data better. The accuracy score of the optimized model decreased somewhat. This is not a negative since the dataset is limited and a more robust model is preferable over an overfit model that is more accurate. Now, we will fit a voting classifier to compare our MLP results to those of many other models. Table 4 shows the values for validation scores in nodes. Table 5 shows the values of metrics and the parameters. Fig. 5 depicts the voting classifier results.

Table 5 Values of the metrics of SVM classifier and the parameters

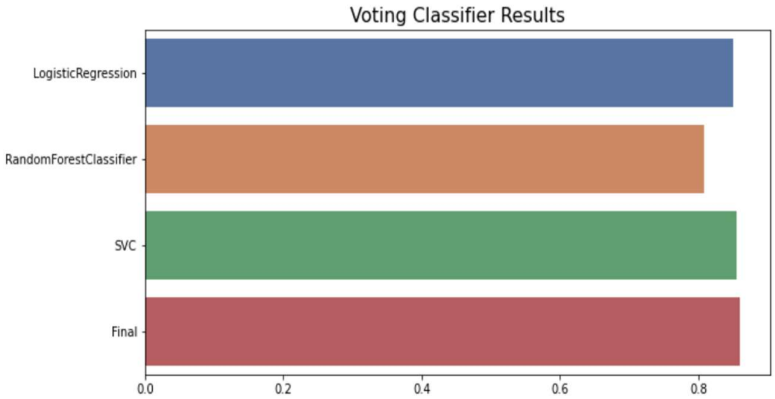


Fig. 5 Accuracy results of ML Models

Support Vector Machine eked out a slim victory against Logistic Regression. There seems to be a somewhat straight link between characteristics and reactions.

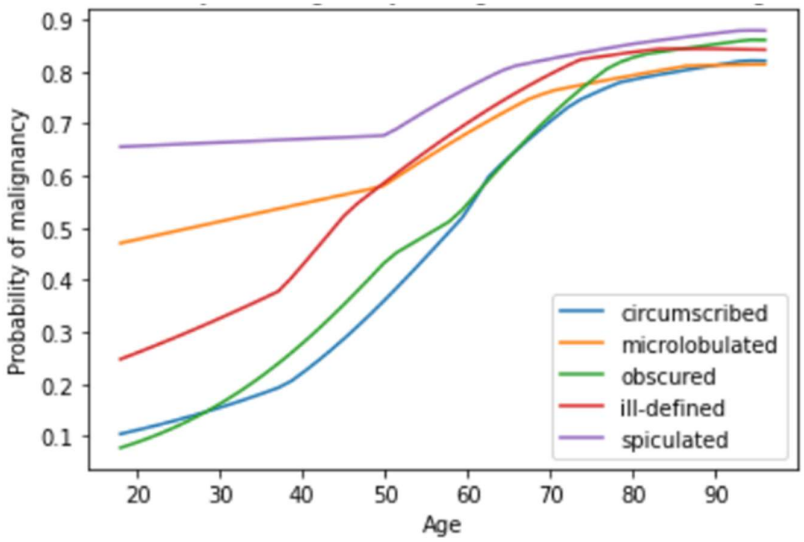


Fig. 6 Probability of Malignancy vs Age with alternate margins

Fig. 6 displays the graph between probabilities of Malignancy vs age with alternate margins According to the curves, there is a clear indication of an increase in the likelihood of cancer as one's age increases, which is logical. The model's ability to effectively capture this signal is consequently a significant indicator. It is intriguing to see the differences between the interactions between different margin elements and replies. When examining persons of advanced age, the probabilities at the margins are equal. Younger people have a substantially greater connection between spiculated margin and malignancy. The circumscribed and micro-lobulated borders have the lowest chance for younger ages, but the highest probability for older ages.This is despite the fact that it has the lowest probability. Obviously, age-related factors play a significant role. This study produced some intriguing findings based on a small number of the factors being considered. Further exploratory data analysis (EDA) will almost certainly reveal more fascinating patterns hidden in the data. Fig. 7 displays the confusion matrix.

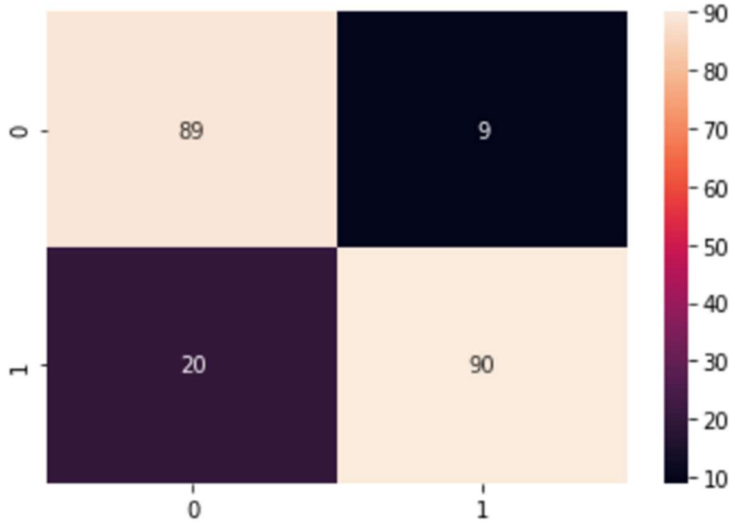


Fig. 7 Confusion Matrix

5. Discussion on results

The purpose of the study was to construct a model for identifying breast cancer using machine learning methods. The data utilised in the research were preprocessed, and further analysis was performed to uncover potentially helpful characteristics that might assist in categorisingtumours as benign or malignant. In addition to other machine learning approaches, the assessment included the Logistic Regression method, the Random Forest algorithm, and the Support Vector Classification algorithm. Also examined was a Voting Classifier that combines all three approaches. With a total score of 86.1%, the Voting Classifier had the greatest degree of accuracy, according to the findings. The performance of the model was enhanced even further with the help of hyper-parameter tweaking using GridSearchCV, which resulted in an accuracy score of 89.5% overall. The fact that the model received good ratings for both accuracy and recall demonstrate that it has the potential to be helpful in the process of properly identifying tumors as benign or malignant. This effort demonstrates the significance of machine learning methods in medical diagnostics as well as the potential of such models to assist clinicians in making informed choices about patient care.Table 6 displays the comparison for existing and proposed method.

Table 6 Comparison table between existing and proposed method

	Accuracy	**Precision**	**Recall**
Proposed Method [Voting Classifier]	86	0.91	0.91
Existing Method [13]	84	0.84	0.87
Existing Method [17]	82.7	0.86	0.81
Existing Method [21]	79	0.89	0.86
Existing Method [34]	81	0.87	0.87

The following are the performance measures for four distinct classifiers: Logistic Regression, Random Forest Classifier, SVC, and Voting Classifier. Logistic Regression attained an accuracy score of 85.6%, Random Forest Classifier attained an accuracy score of 80.7%, SVC attained an accuracy score of 85.5 and Voting Classifier got the maximum accuracy score of 86.05%.

5. Conclusion

To achieve this goal, we first preprocessed and analysed the data to identify features that may be useful in distinguishing between benign and malignant tumours. After this, we developed many other machine learning methods, such as Logistic Regression, Random Forest, Support Vector Classifier, and a Voting Classifier that utilises elements from each of these. The model's performance was then enhanced by adjusting its hyperparameters with the use of GridSearchCV. The findings showed that the Voting Classifier earned the greatest accuracy score of 85.0%, with an F1 score of 0.86 %, suggesting strong precision and recall scores. The

adjustment of hyperparameters enhanced the performance of the model, culminating in its ultimate precision. We also ran preliminary experiments to determine the optimal number of nodes per layer in the MLP classifier, which revealed a robust data signal. To prevent overfitting, however, we proceeded with high-level testing on the number of nodes per layer, followed by a GridSearch-based advanced hyper parameter optimization exercise. The fact that this model was developed using machine learning techniques shows that such models have the potential to aid medical professionals in accurately diagnosing breast cancer. This research highlights the significance of feature engineering, algorithm selection, and hyperparameter tweaking in the development of successful machine learning models. It required preprocessing and exploration of a breast cancer dataset, implementation of many machine learning methods, including a Voting Classifier, and hyperparameter adjustment to maximize the model's performance. The findings demonstrate that the proposed model has the ability to appropriately categoriestumors as benign or malignant, making it a viable diagnostic tool for breast cancer.

Reference

1. Y. Jin, J. Moura, Y. Jiang, "Breast Cancer Detection By Time Reversal Imaging", IEEE,(2008)
2. M. Sajjadieh, F. Foroohar, A. Asif, "Breast Cancer Detection using Time Reversal Signal Processing",IEEE, (2009)
3. S. H. Barboza,J. A. Palacio, E. Pontes, S. Kofuji, "Fifth Derivative Gaussian Pulse Generator for UWB Breast Cancer Detection System",IEEE,(2014)
4. S. Y. Shin, S. Lee, I. D. Yun, S. M. Kim and K. M. Lee, "Joint Weakly and Semi-Supervised Deep Learning for Localization and Classification of Masses in Breast Ultrasound Images," in IEEE Transactions on Medical Imaging, vol. 38, no. 3, pp. 762-774, March 2019.
5. S. Pawar, P. Bagal, P. Shukla and A. Dawkhar, "Detection of Breast Cancer using Machine Learning Classifier," 2021 Asian Conference on Innovation in Technology (ASIANCON), PUNE, India, 2021, pp. 1-5.
6. A. Atrey, N. Narayan, S. Vijh and S. Kumar, "Analysis of Breast Cancer using Machine Learning Methods," 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2022, pp. 258-261.
7. K. Shilpa, T. Adilakshmi and K. Chitra, "Applying Machine Learning Techniques To Predict Breast Cancer," 2022 Second International Conference on Interdisciplinary Cyber Physical Systems (ICPS), Chennai, India, 2022, pp. 17-21.
8. G. Sruthi, C. L. Ram, M. K. Sai, B. P. Singh, N. Majhotra and N. Sharma, "Cancer Prediction using Machine Learning," 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM), Gautam Buddha Nagar, India, 2022, pp. 217-221.
9. S. Kaya and M. Yağanoğlu, "An Example of Performance Comparison of Supervised Machine Learning Algorithms Before and After PCA and LDA Application: Breast Cancer Detection," 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), Istanbul, Turkey, 2020, pp. 1-6.
10. 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), Noida, India, 2021, pp. 1-5.
11. S. Ara, A. Das and A. Dey, "Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms," 2021 International Conference on Artificial Intelligence (ICAI), Islamabad, Pakistan, 2021, pp. 97-101.
12. J. Sivapriya, A. Kumar, S. Siddarth Sai and S. Sriram, "Breast cancer prediction using machine learning", International Journal of Recent Technology and Engineering (IJRTE), vol. 8, 2019
13. Y. Khourdifi and M. Bahaj, "Applying best machine learning algorithms for breast cancer prediction and classification", 2018 International Conference on Electronics Control Optimization and Computer Science (ICECOCS), pp. 1-5, 2018.
14. N. K. Sinha, M. Khulal, M. Gurung and A. Lal, "Developing a web based system for breast cancer prediction using xgboost classifier", International Journal of Engineering Research Technology (IJERT), vol. 9, 2020.
15. X. Jia, W. Meng, S. Li, Z. Tong and Y. Jia, "A rare case of intracystic Her-2 positive young breast cancer," 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX, USA, 2021, pp. 2598-2602.
16. Suman Mann, Amit Kumar Bindal, Archana Balyan, Vijay Shukla, Zatin Gupta, Vivek Tomar, Shahajan Miah, "Multiresolution-Based Singular Value Decomposition Approach for Breast Cancer Image Classification", BioMed Research International, vol.2022, pp.1, 2022.
17. H. Al-Shamlan and A. El-Zaart, "Feature extraction values for breast cancer mammography images," 2010 International Conference on Bioinformatics and Biomedical Technology, Chengdu, China, 2010, pp. 335-340.
18. A. Bharat, N. Pooja and R. A. Reddy, "Using Machine Learning algorithms for breast cancer risk prediction and diagnosis," 2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C), Bangalore, India, 2018, pp. 1-4.

19. N. Khuriwal and N. Mishra, "Breast Cancer Detection From Histopathological Images Using Deep Learning," 2018 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE), Jaipur, India, 2018, pp. 1-4.
20. A. Qasem et al., "Breast cancer mass localization based on machine learning," 2014 IEEE 10th International Colloquium on Signal Processing and its Applications, Kuala Lumpur, 2014, pp. 31-36.
21. M. R. Al-Hadidi, A. Alarabeyyat and M. Alhanahnah, "Breast Cancer Detection Using K-NerestNeighbor Machine Learning Algorithm," 2016 9th International Conference on Developments in eSystems Engineering (DeSE), Liverpool, UK, 2016, pp. 35-39.
22. H. Sami, M. Sagheer, K. Riaz, M. Q. Mehmood and M. Zubair, "Machine Learning-Based Approaches For Breast Cancer Detection in Microwave Imaging," 2021 IEEE USNC-URSI Radio Science Meeting (Joint with AP-S Symposium), Singapore, Singapore, 2021, pp. 72-73.
23. L. Wang, "Early diagnosis of breast cancer", Sensors, vol. 17, no. 7, pp. 1572, 2017.
24. T. Reimer, J. Sacristan and S. Pistorius, "Improving the diagnostic capability of microwave radar imaging systems using machine learning", 2019 13th European Conference on Antennas and Propagation (EuCAP), pp. 1-5, 2019.

25. R. C. Conceição, H. Medeiros, M. O'Halloran, D. Rodriguez-Herrera, D. Flores-Tapia and S. Pistorius, "SVM-based classification of breast tumor phantoms using a UWB radar prototype system", 2014 XXXIth URSI General Assembly and Scientific Symposium (URSI GASS), pp. 1-4, 2014.
26. V. Oleksyuk, F. Saleheen, D. F. Caroline, S. A. Pascarella and C. Won, "Classification of breast masses using Tactile Imaging System and machine learning algorithms", 2016 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), pp. 1-4, 2016.
27. B. Gerazov and R. C. Conceicao, "Deep learning for tumor classification in homogeneous breast tissue in medical microwave imaging", IEEE EUROCON 2017 -17th International Conference on Smart Technologies, pp. 564-569, 2017.
28. S. P. Rana, M. Dey, G. Tiberi et al., "Machine learning approaches for automated lesion detection in microwave breast imaging clinical data", Sci Rep, vol. 9, pp. 10510, 2019.
29. F. A. Spanhol, L. S. Oliveira, C. Petitjean and L. Heutte, "A Dataset for Breast Cancer Histopathological Image Classification," in IEEE Transactions on Biomedical Engineering, vol. 63, no. 7, pp. 1455-1462, July 2016.
30. R. Dhanya, I. R. Paul, S. S. Akula, M. Sivakumar and J. J. Nair, "A comparative study for breast cancer prediction using machine learning and feature selection", 2019 International Conference on Intelligent Computing and Control Systems (ICCS), pp. 1049-1055, 2019.
31. M. M. Islam, H. Iqbal, M. R. Haque and M. K. Hasan, "Prediction of breast cancer using support vector machine and k-nearest neighbors", 2017 IEEE Region 10 Humanitarian Technology Conference (R10- HTC), pp. 226-229, 2017.
32. M. S. Yarabarla, L. K. Ravi and A. Sivasangari, "Breast cancer prediction via machine learning", 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), pp. 121-124, 2019.
33. V. Chaurasia, S. Pal and B. Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques", Journal of Algorithms & Computational Technology, vol. 12, no. 2, pp. 119-126, 2018.
34. N. Fatima, L. Liu, S. Hong and H. Ahmed, "Prediction of breast cancer comparative review of machine learning techniques and their analysis", IEEE Access, vol. 8, pp. 150 360-150 376, 2020.
35. A. Toprak, "Extreme learning machine (elm)-based classification of benign and malignant cells in breast cancer", Medical science monitor: international medical journal of experimental and clinical research, vol. 24, pp. 6537, 2018.
36. D. S. Jacob, R. Viswan, V. Manju, L. PadmaSuresh and S. Raj, "A survey on breast cancer prediction using data miningtechniques", 2018 Conference on Emerging Devices and Smart Systems (ICEDSS), pp. 256-258, 2018.
37. K. L. Kashyap, M. K. Bajpai and P. Khanna, "Breast cancer detection in digital mammograms", 2015 IEEE international conference on imaging systems and techniques (IST), pp. 1-6, 2015.
38. F.-T. Johra and M. M. H. Shuvo, "Detection of breast cancer from histopathology image and classifying benign and malignant state using fuzzy logic", 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), pp. 1-5, 2016.

39. O. V. Singh and P. Choudhary, "A study on convolution neural network for breast cancer detection", 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), pp. 1-7, 2019.
40. A. Gür, "Deep Feature Synthesis for Accurate Breast Cancer Prediction," 2022 Medical Technologies Congress (TIPTEKNO), Antalya, Turkey, 2022, pp. 1-4.
41. Senthil, P., Suganya, M., Baidari, I., & Sajjan, S. P. (2022). Enhancement Sushisen algorithms in images analysis technologies to increase computerized tomography images. International Journal of Information Technology, 14(1), 375-387.
42. Senthil, P., Stanly, M., & Inakshi, S. S. (2020). Improve Multidimensional 5G OFDM Based MIMO Sushisen Algorithms Merge Multi-Cell Transmission. International Journal of Recent Engineering Science, 7(2), 17-21.
43.Vijayaletchumy Subramaniam, Kavenia Kunasegran, P. Senthil THE MULTISENSORY METHOD IN LITERACY MASTERY OF DYSLEXIC STUDENTS . 2023 Mar. 13;94(1): 1250-1272.
44.Senthil, P., & Suganya, M. (2018). Exchanged Nonlinear Third Order Differential Equation Ordinary Differential Equation. Journal for Research| Volume, 4(05).
45.Senthil, P., Stanly, M., & Inakshi, S. S. (2020). Improve Multidimensional 5G OFDM Based MIMO Sushisen Algorithms Merge Multi-Cell Transmission. International Journal of Recent Engineering Science, 7(2), 17-21.
46.Senthil, P. (2016). Image Mining In ranking Approach under Interval-Valued Hesitant Fuzzy Set Gr Selection. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 1(2), 105-114.
47.Senthil, P. (2016). Image Mining Brain Tumor Detection using Tad Plane Volume Rendering from MRI (IBITA). Journal of computer science, 1(1), 1-13.