Advances in Computational Design, *Vol. 11, No. 1 (2026) 79-96*
https://doi.org/10.12989/acd.2026.11.1.079

Techno-Press

A data-driven approach to predicting breast cancer recurrence with hybrid machine learning models

Deepa B.G.*[1], Velmurugan R.[2a], Narender M.[3b], Suhaas K.P.[3c]

[1]*Department of Computer Science, Christ University, Bangalore, India*
[2]*Department of Computer Science, Kristu Jayanti (Deemed to be University), Bangalore, India*
[3]*Department of Computer Science and Engineering, The National Institute of Engineering, Mysore, India*

(Received September 30, 2025, Revised October 19, 2025, Accepted October 20, 2025)

Abstract. Breast cancer recurrence is one of the most significant medical concern, and accurate recurrence models can assist in early intervention and treatment planning. Breast cancer recurrent remains as one of the most critical concern for patients prognosis and treatment planning. Accuracy Predicting individual recurrence risk is crucial for the development of precise therapy, specialy for those patients with high-risk profiles. In the study proposes a hybrid machine learning approach that uses the computational modeling and the medical information to predict the recurrence of breast cancer in a patient. The dataset contains the medical and patient information like the age, tumor size, lymph node involvement, malignancy degree, location, irradiation status and recurrence class. This proposed approach begins with the process of data processing, handling the missing data values, features normalization and encoding of categorical variable into numerical format. The dataset is divided into two parts the training set and the testing set and the two selected models' random forest and logistic regression models are trained independently. The predictions form both the model is stacked and a logistic regression meta-model is trained on these combined predictions. The evaluation of the model was conducted using the metrics such as accuracy, precision, recall, and F1 score. The designed hybrid model was able to achieve the accuracy of 97.66% with the precision, recall and F1 score all reaching around 98.15%. This study highlights the potential of hybrid machine learning techniques, improving the accuracy and reliability of machine learning models for breast cancer recurrence prediction. This development model can serve as a valuable tool for the medical industry to support decision making and assist in personalized treatment decisions, offering early detection of recurrence. This can enhance the treatment of a patient by supporting early detection and patients' outcomes through targeted therapy.

Keywords: breast cancer; logistic regression; machine learning; recurrence; random forest

1. Introduction

Breast cancer is one of the most common forms of cancer affecting women worldwide, with millions of new cases diagnosed every year [1]. While advancements in medical treatments have significantly improved survival rates, the risk of recurrence remains a major concern for many

*Corresponding author, Associate Professor, Ph.D., E-mail: deepabg03@gmail.com
[a] Associate Professor, Ph.D., E-mail: velmurugan@kristujayanti.com
[b] Associate Professor, Ph.D., E-mail: narender@nie.ac.in
[c] Associate Professor, Ph.D., E-mail: kpsuhaas@gmail.com

ISSN: 2092-7614 (Print), 2092-7622 (Online)

patients [2]. Recurrence refers to the return of cancer after treatment, and it can occur months or even years later, making it difficult to predict [3]. Early detection of recurrence is essential as it allows for timely intervention and more personalized treatment plans, which can greatly improve patient outcomes.

In recent years, the integration of machine learning into healthcare has opened new possibilities for predictive modeling, especially in diseases like cancer. Machine learning models can analyze complex data, identifying patterns that may not be evident through traditional methods. Predicting breast cancer recurrence using machine learning provides an opportunity to offer more precise risk assessments, allowing clinicians to make better-informed decisions regarding patient care [4-7].

This study explores a hybrid machine learning approach to predict breast cancer recurrence. By combining multiple machine learning algorithms and leveraging both clinical and patient data, the proposed model aims to improve the accuracy of recurrence predictions [8-10]. This hybrid approach not only enhances prediction capabilities but also offers a more robust tool for early intervention, ultimately supporting the development of more targeted therapies for those at high risk of recurrence.

2. Related work

The authors [11] employ machine learning to differentiate between recurrence and non-recurrence events in breast cancer patients. The approach uses functional classifiers that evaluate clinical features such as patient age, tumor size, and other medical indicators to provide more accurate predictions. The results of the study indicated a significant improvement in prediction accuracy, achieving over 90% accuracy in classifying recurrence events and around 88% accuracy for non-recurrence events. These results suggest that functional classifiers can offer a robust method for identifying patients at risk of recurrence, allowing for better-informed treatment decisions and early interventions. The study adds valuable insights into the role of machine learning in enhancing breast cancer prognosis.

The authors [12] explore the application of neural networks alongside traditional machine learning algorithms to enhance the prediction of breast cancer recurrence. The researchers developed a model combining artificial neural networks (ANN) with machine learning techniques, leveraging various patient and tumor characteristics to classify cases effectively. Their approach achieved high classification accuracy, with the ANN model yielding results of over 92% accuracy in predicting recurrence. This integration of neural networks with machine learning was particularly effective in identifying subtle patterns in the data, offering an improvement over single-method approaches. The study demonstrates the potential of combining ANN and machine learning techniques to better assess the risk of recurrence, enabling more personalized treatment and early intervention strategies for breast cancer patients.

The Classification of Non-Recurrence-Events and Recurrence-Events Using Function Classifiers," the authors investigate the use of functional classifiers to enhance the accuracy of predicting breast cancer recurrence. By analyzing a comprehensive dataset of patient information and clinical characteristics, the study focuses on effectively distinguishing between non-recurrence and recurrence events. The findings reveal that the functional classifier achieved an impressive accuracy rate of approximately 91% for classifying recurrence events and around 87% for non-recurrence events. This demonstrates the model's capability to capture complex patterns within the data, facilitating better risk assessment for patients. The research underscores the importance of

advanced classification techniques in oncology, suggesting that such models can significantly contribute to personalized treatment strategies and improved patient outcomes by enabling earlier interventions for those at higher risk of recurrence [13].

Authors [14] investigate how explainable artificial intelligence (AI) can clarify the factors influencing breast cancer recurrence predictions. By analyzing clinical features such as tumor size and lymph node involvement, the study identifies these factors as crucial indicators of recurrence risk. The findings show that the model achieved an accuracy of approximately 90% while providing insights into feature importance, with tumor characteristics being the most significant contributors to the predictions. This research underscores the value of interpretable models in clinical practice, enhancing trust in automated predictions and facilitating personalized treatment strategies for patients at risk of recurrence.

The authors [15] explore the use of thermal imaging alongside machine learning and deep learning techniques for breast cancer detection. The findings reveal that the model achieved an accuracy of about 95% in identifying cancerous lesions, underscoring the effectiveness of thermal imaging as a diagnostic tool. This research emphasizes the potential of advanced algorithms combined with non-invasive imaging methods to improve early detection of breast cancer, ultimately supporting timely interventions and better patient outcomes.

3. The proposed algorithm

Step 1: Dataset Input

This is the process of loading the dataset, it includes features such as age, menopause status, tumour size, lymph nodes, node caps, degree of malignancy, breast location, breast quadrant, irradiation and the recurrence class.

The dataset used for this study contains the clinical and the pathological features of the patients which includs:

1.1 Start Age: Age at which the patient was diagnosed.
1.2 End Age: Age at which the treatment was completed.
1.3 Menopause: Menopausal status (premeno, ge40, lt40).
1.4 Start Tumour Size: Size of the tumour at the start of the treatment.
1.5 End Tumour Size: Size of the tumour at the end of the treatment.
1.6 Start_env_nodes: Number of involved lymph nodes at the start.
1.7 End_env_nodes: Number of involved lymph nodes at the end.
1.8 Node-caps: Presence of node capsulation (yes, no).
1.9 Deg-malig: Degree of malignancy (1-3).
1.10 Breast: Affected breast (left, right).
1.11 Breast-quad: Breast quadrant of the tumor (left_up, left_low, right_up, right_low, central).
1.12 Irradiat: Whether the patient received irradiation (yes, no).
1.13 Class: Recurrence status (recurrence-events, no-recurrence-events).

Step 2: Data Preprocessor

2.1 Data Separation: here the dataset is separated into classes representing different outcomes, such as recurrence-events and no-recurrence events. This process of separation is crucial for training the model to distinguish between different recurrence statuses accordingly and accurately.

2.2 Data Cleaning: this is the process of handling missing values or inconsistencies in the dataset and converting categorical variables into numerical format using encoding techniques.

2.3 Feature Engineering: Create new features as it is necessary to improve the model's predictive power.

2.4 Data Normalization: Normalizing numerical features to a range [0,1]. Normalization helps in standardizing the input data and improving the convergence speed of the training process.

2.5 Data Encoding: Encode categorical features using one-hat encoding to convert them into a numerical format suitable for machine learning algorithms.

Step 3: Data Splitting

The dataset is split into training and testing subsets which is used for the evaluation of the models performance, helping to access its generalization ability and prevent overfitting.

Step 4: Data Modelling:

4.1 Random Forest: It is an ensemble learning method that created multiple decision trees using different subsets of the training data and a random subset of features. Each tree is used to predicts the target variable independently.

4.2 Logistics Regression: It is a linear model which is used for binary classification, which learns the weights for each feature by optimizing the log-loss function.

Step 5: Stacking

5.1 Generate Base Predictions: process of generating each instance in the test dataset, both models generate their predictions. These predictions are then treated as new features for the meta-model.

5.2 Combine Predictions: they are later combined as the predictions from the base models to form a new feature vector.

5.3 Train Meta-Model: the meta model (Logistic Regression) is trained on the combined predictions using the actual class labels from the test set as the target.

Step 6: Model Evaluation

Calculate Metrics: Evaluating the performance of the models using metrics such as accuracy, precision, recall, and F1 score.

Plot Accuracy and Loss Curves: Using matplotlib to plot the accuracy and loss curves which helps to visualize the model's learning progress.

Step 7: Model Saving

Save the trained models for future use using joblib.

Step 8: Model Loading

Load the saved models when needed to make new predictions on classifications.

Step 9: New Predictions

9.1 Load New Data: Load new patient data with features such as age, tumour size, etc which are used for the process of classification.

9.2 Preprocess Data: Preprocess the new data using the same steps as the training data (encoding categorical variables, normalizing numerical features).

9.3 Make Predictions:

a. Generate predictions using the loaded base models.
b. Combine the predictions to form a new feature vector.
c. Use the meta-model to make the final prediction based on these inputs.

This proposed algorithm leverages the strengths of both Random Forest and Logistic Regression models to enhance prediction accuracy and robustness, offering a reliable tool for predicting breast cancer recurrence.

Fig. 1 represents the flow chart of the model training process, it represents the steps involved for the proposed algorithm for predicting breast cancer recurrence.

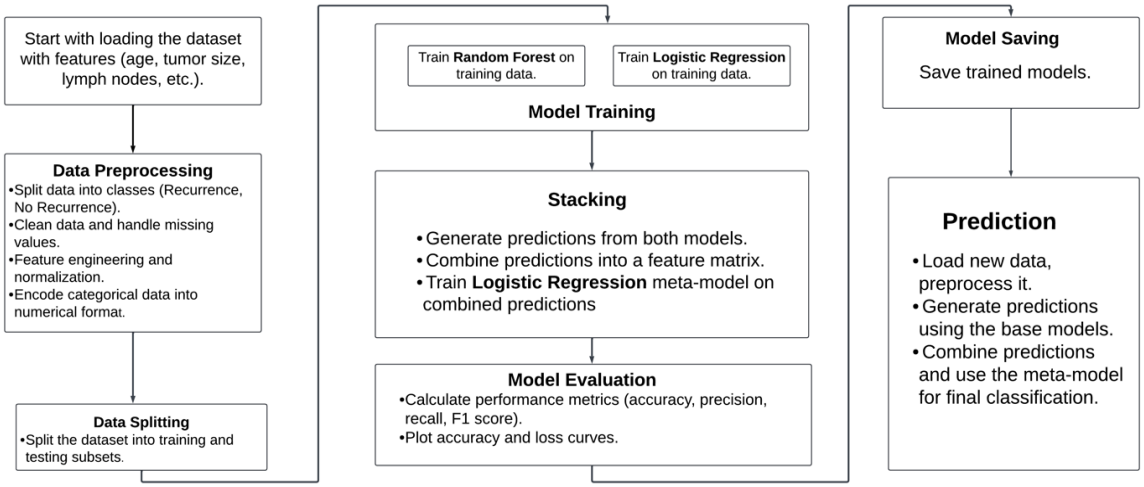

Figure 1. Flow chart

4. Breaking down the mathematics behind logistic registration and random forest models

4.1 Logistics regression

4.1.1 Logistics Regression (sigmoid function):

n Logistics Regression the sigmoid function is used to convert a linear combination of the input features into a probability score ranging from 0 to 1. This function is defined as:

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{1}$$

Where z is the linear combination of input features and weighs:

Z = aT x + b

a = Vector of weights corresponding to input featues.

x = Input feature vector, including variables such as tumour size age.

b = Bias term.

Where the output σ(z) represents the probability that a tumour is malignant.

4.1.2 Decision boundary

The decision boundary in logistic regression is where the models predicted probability is 0.5 this is used to classify the tumour as malignant or benign:

$$\sigma(z) = 0.5 \tag{2}$$

$$aT\ x + b = 0 \tag{3}$$

4.1.3 Cost Function (Log Loss)

The models performance is measured using the cost function, which quantifies the difference between predicted probabilities and actual outcomes

$$J(a,b) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log\left(h_{w,b}\left(x^{(i)}\right)\right) + \left(1-y^{(i)}\right)\log\left(1-h_{w,b}\left(x^{(i)}\right)\right)\right] \quad (4)$$

m : Number of training samples.
$y^{(i)}$: Actual label for the i-th sample.
$h_{w,b}\left(x^{(i)}\right)$: Predicted probability for the i-th sample

4.1.4 Gradient descent

Optimize the weights and bias by the processing of minimizing the cost function by using the Gradient Descent:

$$w := w - \alpha\frac{\partial J(w,b)}{\partial w} \quad (5)$$

$$b := b - \alpha\frac{\partial J(w,b)}{\partial b} \quad (6)$$

α: learning rate.

4.2 Random forest

Random Forest model is used in this project to improve the classification accuracy of the model by aggregating multiple decision trees, each making a vote for the class label. Here's the mathematical breakdown of the Random Forest in this model:

4.2.1 Decision trees in random forest

Here each decision tree is built recursively by splitting the dataset based on the criteria such that Gini impurity or information gain:

Gini Impurity: Is used to measures the impurity of a node split and it is defined as:

$$\text{Gini} = 1 - \sum_{i=1}^{c} p_i^2 \quad (7)$$

Where pi is the proportion of samples belonging to class i.

4.2.2 Ensemble voting

In Random Forest, each tree contributes a vote for the class label, and the final classification is determined by majority vote. majority vote is the class with the highest number of votes from all trees is selected.

4.2.3 Out-of-Bag error:

Out of bag error or (OOB) error is internal validation metric for Random Forest. It evaluates samples that were not used in training specific trees, OOB error is the average error rate of prediction made using the tree that did not include the sample in their training subset.

In Summary this breast cancer classification model uses, Logistic Regression which provides a probabilistic for determining whether a tumour is malignant or benign for the model. It uses a sigmoid function to map input features to the probabilities and determines a decision boundary to classify tumours and Random Forest combines multiple decision trees to improve classification accuracy. It utilizes ensemble voting and manages complex feature interactions effectively. In this project both models are evaluated using metrics such as accuracy, precision, recall, and F1 score to

ensure reliable and accurate classification of breast cancer tumours.

5. Logistic regression and random forest in overall

The developed framework for breast cancer classification has been able to demonstrate significant accuracy and efficiency by leveraging the combined power of Logistic Regression and Random Forest models. The Logistic Regression and Random Forest model is very well known for handling structured data and have been effectively applied to the breast cancer dataset to predict tumour recurrence.

Logistic Regression is a supervised machine learning algorithm which uses mathematical statistical approach primarily used for binary classification tasks, for predicting whether a breast cancer tumour is malignant or benign also treated as Class 1 or Class 0. Logistic Regression is more closely related to classification problems than traditional regression.

This model's strength lies in its ability to provide interpretable results, where the coefficients represent the relationship between each feature and the likelihood of tumour recurrence. For example, a positive coefficient indicates that an increase in that feature increases the probability of the tumour being malignant. The Logistic Regression algorithm predicts the probability of a binary outcome value, where the value of the dependent variable can take one of two discrete values. This algorithm uses a sigmoid function to map the input features to the probability value between the range of 0 to 1. This function produces an "S" shaped curve, which represents the probability of the dependent variable being in a particular class. The decision boundary is determined by the threshold value. If the predicted probability value is above this threshold value, the instance is classified as malignant else it is classified as benign.

The Logistic Regression is a type of regression analysis used for binary classification problems, where the output is categorical rather than continuous. This process transforms the output of a linear regression model into a probability value that lies between the values 0 and 1 by using a sigmoid function. This transformation is used to enable the model to handle classification tasks. Considering the set of independent features X which can be represented as

$$X\begin{Bmatrix} X_{11} & X_{12} & \dots & X_{1M} \\ X_{21} & X_{22} & \dots & X_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ X_{N1} & X_{N2} & \dots & X_{NM} \end{Bmatrix} \tag{8}$$

Now the dependent variable Y is a binary value which can take only any one of the two possible classes (0 or 1).

$$Y=\begin{cases} 1 \,/\, \text{TRUE} & \text{If CLASS A} \\ 0/\text{FALSE} & \text{If CLASS B} \end{cases} \tag{9}$$

The Logistic regression model computes a linear combination of all the needed input features and then applies the sigmoid function to the values to determine the output probability. This probability value indicates the likelihood that the given instance belongs to Class 0 or 1 depending on the values. To the context of breast cancer classification, The Logistic Regression used here has been applied to predict whether a tumour is likely to be malignant based on various input features. The model than provides a probabilistic value that can be used to determine the likelihood of the

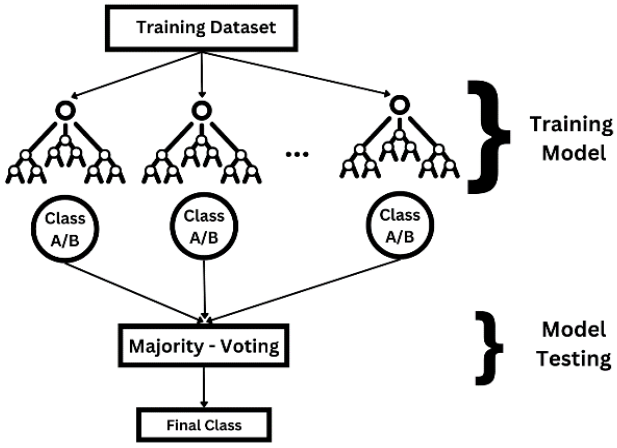

Figure 2. Random forest model

tumour being malignant. By adjusting the decision threshold value, the sensitivity (recall) and the specificity (precision) of the model can be tuned to achieve optimal performance. The approach used allows for an interpretable model where the significance of each feature in predicting the malignancy of the tumour can be understood, making it a valuable tool in medical diagnostics.

Random Forest as shown in the above Fig. 2, introduced by Leo Breiman in 2001, is a robust machine learning algorithm that uses the process of combining the power of multiple decision trees to improve the prediction accuracy and to reduce overfitting. It is widely applied in both classification and regression tasks. Random Forest algorithm uses the method of ensemble learning that by constructing multiple decision trees during the phase of training, each using a random subset of the dataset and features. This algorithm can aggregate the results from all trees, either by the process of majority voting (for classification) or by the process of averaging (for regression), to make a final prediction. This process ensures that the model is stable and precise for effectively handling complex data. This helps in focusing on the most relevant aspects of the data. By aggregating the decisions from the multiple trees, Random Forest algorithm Random Forest can be used to effectively manage these high-dimensional data by the process of selecting random subsets of features form each tree, preventing the model from becoming overwhelmed. While Random Forest is more complex when compared with a single decision tree, it still retains a level of interpretability. Which allows us to access the importance of the different features, providing insights into which the aspects of the data which are most critical for making accurate predictions. In summary, the Random Forest algorithm is considered to be a robust and effective method for handling complex classification tasks, offering both accuracy and interpretability.

6. Results and discussion

6.1 Dataset description

The dataset which is utilized in this research pertains to breast cancer diagnosis, this is specifically focusing on the prediction of cancer recurrence post-treatment. The dataset comprises 286 instances, in which each instance represents a patient’s clinical and demographic information.

The features present in the dataset were carefully selected to provide a comprehensive view of the factors that may influence the likelihood of breast cancer recurrence. The used dataset includes a total number of 13 attributes, which are categorized into independent variables (features) or a dependent variable (target).

The list of independent variables present in the dataset are as:

1. Start Age: Represents the lower bound value of the patient's age range.
2. End Age: Represents the upper bound value of the patient's age range.
3. Menopause: Represents the menopausal status of the patient, which is categorized as
   a. ge40: Greater than 40 years.
   b. lt40: Less than 40 years.
   c. premeno: Pre-menopausal.
4. Start Tumor Size: Represents the value of initial size of the tumor in millimeters.
5. End Tumor Size: Represents the value of the size of the tumor after a period of observation or treatment.
6. Start_env_nodes: Represents the value of initial number of axillary lymph nodes involved.
7. End_env_nodes: Represents the value of number of axillary lymph nodes involved at a later stage.
8. Node-caps: Represents whether the lymph nodes are encapsulated, which is again categorized as:
   a. yes: Encapsulation present.
   b. no: Encapsulation absent.
9. Deg-malig: Represents the degree of malignancy, where higher values indicate more aggressive cancer.
10. Breast: Represents the breast affected by cancer, which is categorized as:
    a. left: Left breast.
    b. right: Right breast.
11. Breast-quad: Represents the quadrant of the breast where the tumor is located, which is categorized as:
    a. left_up: Upper left.
    b. left_low: Lower left.
    c. right_up: Upper right.
    d. right_low: Lower right
    e. central: Central region.
12. Irradiate: Indicates whether the patient received radiation therapy, categorized as:
    a. yes: Radiation therapy was administered.
    b. no: No radiation therapy.

The list of dependent variables present in the dataset are as:

a. Class: Represents the outcome of the cancer treatment, indicating whether the cancer recurred, categorized as:
b. recurrence-events: experienced a recurrence of breast cancer.
c. no-recurrence-events: did not experience a recurrence.

6.2 Data preprocessing

The used dataset contains a sample size of 286 instances. The instance contains a feature count of 12 independent variables and 1 dependent variable. In these each instance contains 6 categorical

```
[121] df = pd.read_csv("breast cancer diagnosis .csv")

[122] print(df.head(3))

...    Start Age  End Age menopause  Start tumor size  End tumor size  \
    0         40       49   premeno                15              19
    1         50       59      ge40                15              19
    2         50       59      ge40                35              39

       Start_env_nodes  end_env_nodes node-caps  deg-malig breast breast-quad  \
    0                0              2       yes          3  right     left_up
    1                0              2        no          1  right     central
    2                0              2        no          2   left    left_low

      irradiat                 Class
    0       no     recurrence-events
    1       no  no-recurrence-events
    2       no     recurrence-events
```

```
   Start Age  End Age  menopause  Start tumor size  End tumor size  \
0         40       49          2                15              19
1         50       59          0                15              19
2         50       59          0                35              39
3         40       49          2                35              39
4         40       49          2                30              34

   Start_env_nodes  end_env_nodes  node-caps  deg-malig  breast  breast-quad  \
0                0              2          2          3       1            2
1                0              2          1          1       1            0
2                0              2          1          2       0            1
3                0              2          2          3       1            1
4                3              5          2          2       0            5

   irradiat  Class
0         0      1
1         0      0
2         0      1
3         1      0
4         0      1
```

Figure 3. Dataset used for training process

variables which where post-encoding and 7 numerical variables including both age and tumour size-related features. The descriptive analysis of this dataset indicated a balanced distribution of the categorical variables which was post-encoding, which is a essential process for ensuring that the machine learning models are not biased toward any particular class.

The dataset used contained some missing values. During the training process these missing entries were replaced with ‘NaN’ values, and a missing value analysis was conducted to make sure that the dataset is clear.

The dataset contains categorical attributes on which label encoding was applied to convert them into a numerical value which is used for the machine learning algorithms. The categorical

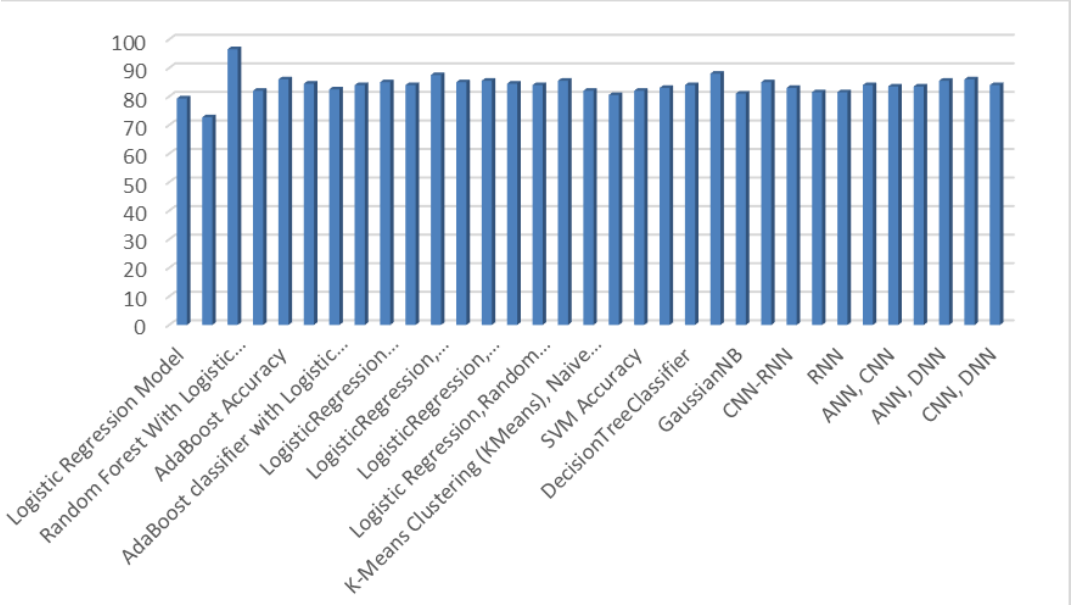

Figure 4. Accuracy Value

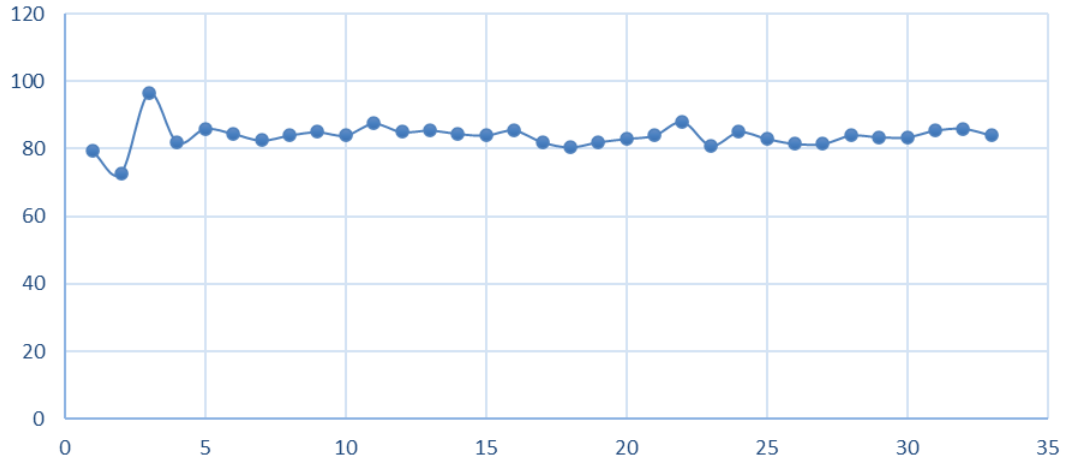

Figure 5. Accuracy Value

attributes that were labelled or encoded are ‘menopause’, ‘node-caps’, ‘breast’, ‘breast-quad’, ‘irradiate’, and ‘Class’. The encoded values were documented.

During the process of model training the dataset was partitioned into two major subsets training and testing subsets with the ratio of 80-20 split. The training subset was utilized for the process of model training only, while the testing subset was reserved only for the process of evaluating model performance. This partitioning was done to ensure that the models could generalize well to unseen data also, which is important for robust predictive analytics.

The dataset provides a rich source of information which was crucial for modelling the recurrence of breast cancer. By the process of data preprocessing, handling missing values and

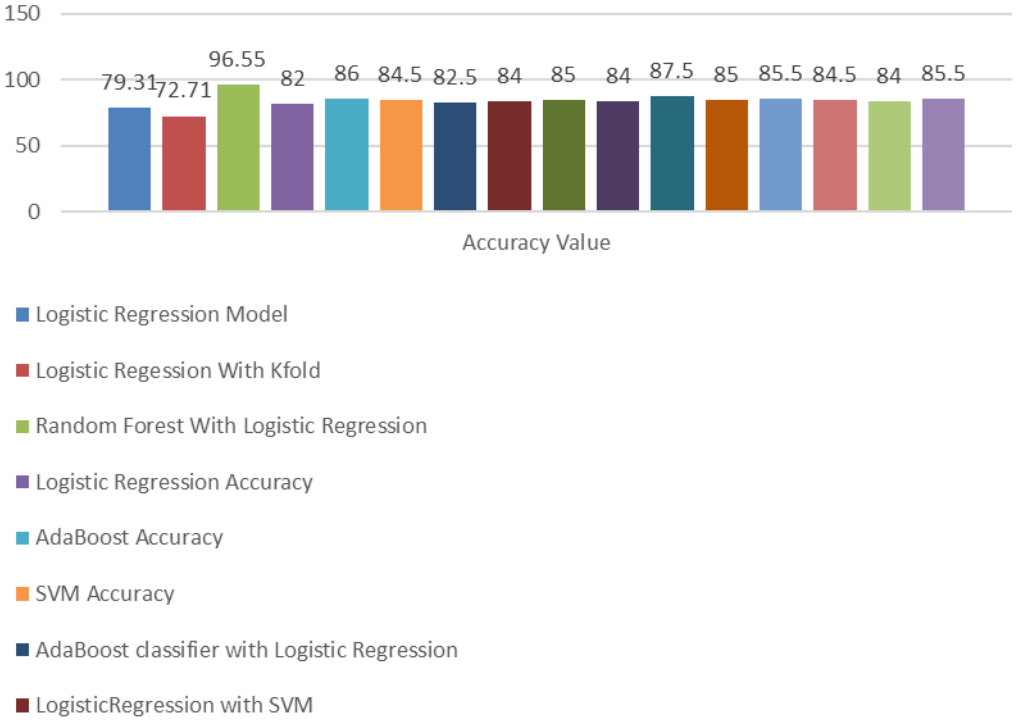

Figure 6. Accuracy Value

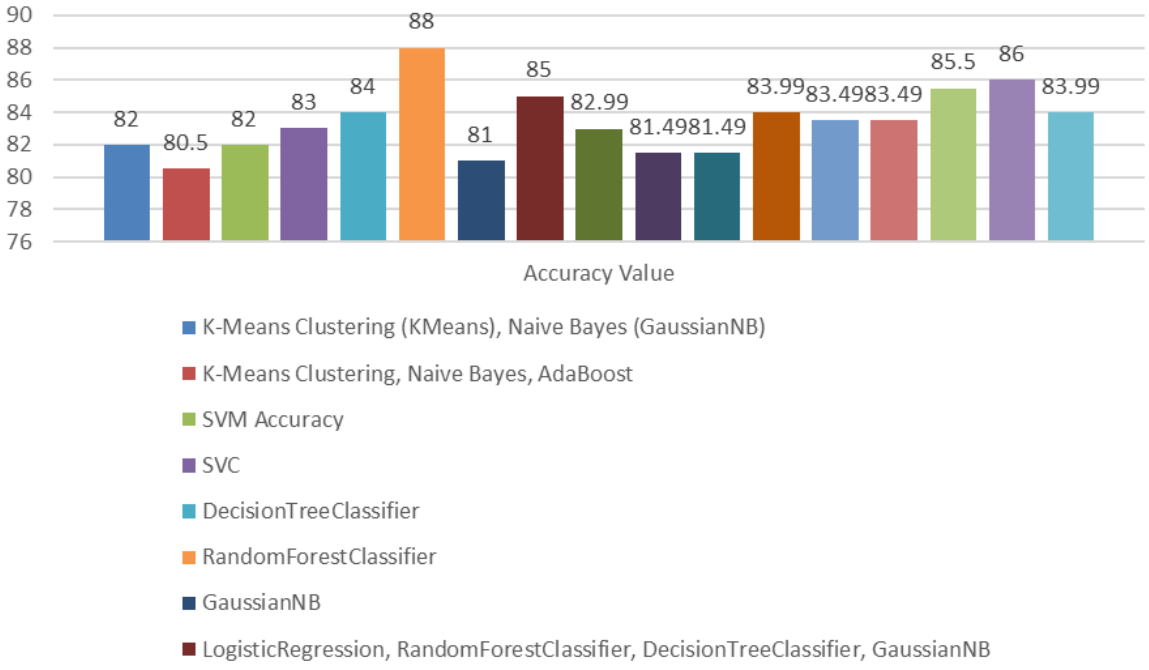

Figure 7. Accuracy Value

encoding categorical variables, the dataset was well-prepared to be used as input into various machine learning models.

7. Analysis of results

While evaluating the performance of various machine learning models, the results of the used models demonstrate a diverse range of accuracy levels as shown the figures 5 and 6, also indicating the relative effectiveness of the model in handling the dataset. Among the tested individual models, the Random Forest classifier model shows the highest accuracy value of 88%, displaying its high performance in classification tasks. The AdaBoost model also seen to exhibit notable effectiveness by achieving the accuracy level of 86%. With the comparison to the Support Vector Machines (SVM) and Decision Tree classifiers is seen to provide moderate accuracy value of 82% and 84%, respectively. These models did not reach the peak performance as of Random Forest. Gaussian Naive Bayes and K-Means Clustering have seen to be with the accuracies of 81% and 82%, these models performed well but fall short behind the higher-performing models.

The models when combined with each other, the accuracy metrics shows significant improvements in the accuracy values. The combination of Logistic Regression and Random Forest had been able to achieves the highest overall accuracy of 96.55%, illustrating a robust performance that was achieved through effective model integration. Other combinations of the model include Logistic Regression with Random Forest and Decision Tree classifiers, which yields an accuracy value of 87.5%, and the various other multi-model combinations involving Logistic Regression, Random Forest, Decision Trees, Gaussian Naive Bayes, and K-Nearest Neighbours, which have consistently deliver accuracies around 85.5%. The deep learning models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) exhibit high performance. CNN model was able to achieve an accuracy value of 81.49%, while RNN models were able to match the same performance level. The fully connected architecture of CNN model was able to improves its result to 83.99%, and combined architectures such as Artificial Neural Networks (ANN) along with CNN was able to demonstrate the accuracy value of 83.49%.

The Deep Neural Network (DNN) was also tested with the dataset and has also performed comparably, with accuracies around 83.49%. When tested with the combination of ANN and DNN models they were able to achieves a accuracy value of 85.5%, and the integration of these models RNN with DNN was able to provide a high accurate value of 86%. Despite these improvements it was observed that the best-performing combinations of traditional models was able surpass the deep learning approaches. Overall, at the end of this analysis it was observed that individual models and deep learning techniques have their merits, combining the traditional models such as Logistic Regression and Random Forest was able to provide superior results.

These information’s are important for developing a robust classification system, it is also observed that integrating multiple models often provides the best performance. The future work may explore further improvements and additional model combinations which may lead to enhance accuracy and improve generalization capabilities.

8. Training procedure

In the model the training process had the following workflow which involves initializing and training two distinct models to tackle the problem for predicting the recurrence, the models where selected because of their excellent performance and accuracy.

Figure 8. Accuracy of logistic regression with random forest model

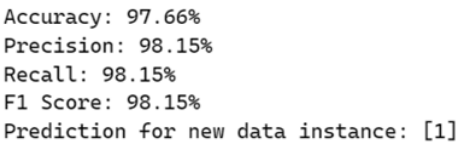

Figure 9. Model evaluation metrics

The Random Forest Classifier is the first model which was used this model creates a multitude of decisions trees, each trained on a subset of the data, and then combined with their individual prediction to make a final decision. We configure it with 100 trees, which is mainly helps in enhances its ability to capture complex patterns and relationships present within the data. This model was trained on our dataset, where it learns from the features present in the dataset.

Alongside, we train a Logistic Regression model. This model operates differently when compared to that of the functionality of Random Forest. It assumes a linear relationship between the features and the target variable. Logistic Regression mainly works on binary classification tasks, where it estimates the probabilities and make the decision based on the logistic functions. The model is trained again on the same dataset for its learning to predict the probability of the recurrence of cancer.

After the successful completion of the training process of both the models, we then proceed to generate predictions. The prediction of both the Random Forest and Logistic Regression model are essentially important for the next stage of the process.

To enhance the predictive performance of these frameworks we combine the presentations from both the models into a new feature matrix. This matrix contains columns corresponding to the prediction made by each of the two models. By stacking these predictions values, we create a richer feature set for the next stage.

After which we train the meta-model, which is another instance of the Logistic Regression, on this combined feature matrix with the actual labels. This stage improves the strengths of both models to improve the overall accuracy of this model.

Finally, the accuracy of the meta-models predicted is assessed by the process of comparing to the true labels from the test set. To ensure that the trained models are preserved and can be used in future the complete model pipeline is saved to a pkl file. This saved model can be loaded later to make predictions on the new dataset, ensuring consistent.

This comprehensive approach, involving training individual models, combining their predictions leading to improved performance in predicting cancer recurrence.

9. Error validation

The above confusion matrix is a tool used to measure the performance for the machine learning classification models. It provides a clear visualization of how well the model performed in the test

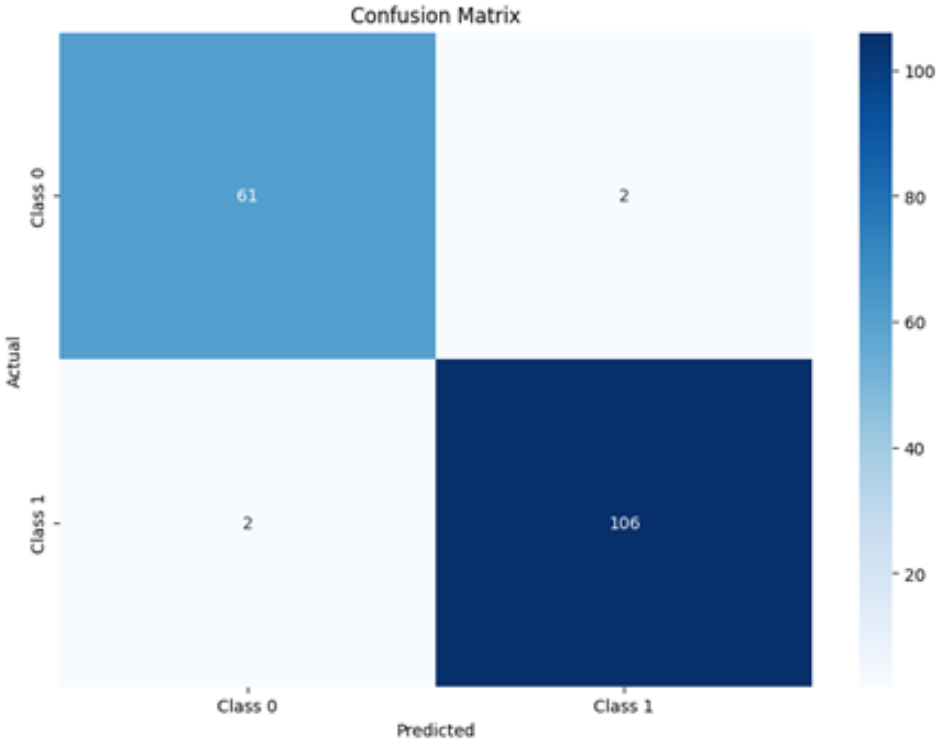

Figure 10. Confusion matrix

data, by understanding the above confusion matrix it can be stated that the model is predicting whether the breast cancer will be a recurrence event or non-recurrence event we know that the two classes class 0 (No Recurrence) and class 1 (Recurrence). The matrix represents the count of the predictions belonging to the categories like ‘true positive’, ‘true negative’, ‘false negative’ and ‘false positive’.

where the category ‘true positive’ which represents that the model correctly predicted that cancer recurred. The next category ‘true negative’ which represents that the model correctly predicted that cancer did not recur. The other category ‘true negative’ represents the model’s incorrect predictions that cancer did not recur when it actually did. The last category ‘false positive’ represents the model’s incorrect predictions that cancer recurred when it didn’t.

The information gained by the confusion matrix is as follows, the 61 instances of the ‘class 0’ was correctly predicted as non-recurrence event which belongs to the category of True Negative. The 2 instances of the same were incorrectly predicted as recurrence event which belongs to the category of ‘False Positive’. It is also observed that for the prediction for the ‘class 1’ event it was seen that 106 instances were correctly predicted as recurrence which falls under the category of True Positive, the next 2 instances were incorrectly predicted as no recurrence which falls to the category of False Negative.

From the above confusion matrix, it can be concluded as the models shows high accuracy in prediction of the data for both the classes. With only ‘2 False Positive’ and ‘2 False Negative’, it indicates that the model is performing well and able to correctly distinguish between the two cases.

When a given total of 171 predictions on this dataset was made only 4 were incorrect, the model can be considered reliable for making predictions on this dataset, hence this can be highly beneficial for predicting the recurrence of the breast cancer. False negative outcomes are especially considered to be critical in medical applications because a missed recurrence could delay treatment which may further lead to worst outcomes. Here, the model only made 2 False Negative errors which represents it is quite very effective at identifying these events. False Positive models can also lead to unnecessary stress for patients and may lead to further unnecessary treatments here the model only made 2 False Positive predictions keeps these errors cases low, which represents that this model is predicting recurrence event accurately.

This confusion matrix highlights that the model is highly effective in predicting the breast cancer recurrence event. This balance between the True Positive and True Negative, combined with the low count of False Positive and False Negative, shows that the model is robust and reliable for clinical applications. This is particularly considered to be very important in the field of healthcare, where the false prediction may lead to unnecessary treatment or worst cases.

10. Applications

The framework developed is for the prediction of the recurrence of the breast cancer using the combination of Random Forest and Logistic Regression models discussed above. The framework has a major importance in the healthcare and clinical domains.

One of the most significant applications of this framework is its ability to predict the recurrence of breast cancer in patients who had undergone initial treatments. By Early and accurate distinguishing between patients at high and low risk of breast cancer recurrence can help the patients to undergo any additional treatment options if needed. With the model's strong predictive performance, the treatment plan can be made based on the patient's risk profile. The patients with high risk for recurrence may benefit from more aggressive treatment or close monitoring and the patients with low risk can avoid unnecessary treatments and their effects.

The model's higher accuracy minimizes the chances of false predictions which leads to a more efficient allocations of healthcare resource and help reduce the financial burden on patients by avoiding all the unnecessary treatments.

The models help in improving the life quality by predicting cancer recurrence, the model helps patients avoid the unnecessary treatments their side effects and reduce the financial burdens, also as the model can identifies the patients with high chances of recurrence, they can continue the treatment with specialized treatment and extra care.

11. Conclusions

In this study, we have developed and implemented a breast cancer recurrence predictive framework utilizing the ability of both Random Forest and Logistic Regression models. By integrating a model which combines the predictions of both these models (Random Forest and Logistic Regression), we were able to significantly improve the overall accuracy and robustness of the predictive system. The results of this model were able to demonstrate a highly effective value, with a low incidence of false positive and false negative prediction value making it a valuable tool for medical applications.

The confusion matrix evaluates the model and showed the models ability to accurately distinguish between the recurring and non-recurring events with a minimal error. The false prediction of the model was very low, where the false prediction of the model can lead to either unnecessary treatments or the missed opportunities for proper care. The model can be lifesaving for the patients at the risk of recurrence.

The training procedure and procedure used for the evaluation process of the model indicates that this approach used is highly reliable, scalable, and well-suited for deployment in the medical field for the clinical decision-support systems. By minimizing diagnostic errors one can improve the patients' health monitoring, the developed framework has the potential to enhance the quality of patient's health care, facilitating personalized treatment plans for the patients with high chance of breast cancer recurrence and also help significantly reduce the unnecessary treatment provided for patients with low chances of breast cancer recurrence reducing the financial burden along with the side effects of the treatments.

In overall, this study demonstrates the potential of combining machine learning models which can result in improved performance and accuracy in predicting or distinguishing between the occurrence and non-occurrence events. Future work can expand this existing framework to the ability to detect other types of cancers or disease, as well the model's performances can be enhanced on more diverse datasets. This successful application of artificial intelligence in this study serves as a step forward in interacting data-driven technologies into the modern healthcare systems, supporting more accurate diagnoses and personalized treatment if needed for the better patient's outcome.

References

1. Pati, A., Panigrahi, A., Parhi, M., Giri, J., Qin, H., Mallik, S., ... Agrawal, U.K. (2024). Performance assessment of hybrid machine learning approaches for breast cancer and recurrence prediction. PloS One, 19(8), e0304768. https://doi.org/10.1371/journal.pone.0304768
2. Othman, N.A., Abdel-Fattah, M.A., Ali, A.T. (2023). A hybrid deep learning framework with decision-level fusion for breast cancer survival prediction. Big Data and Cognitive Computing, 7(1), 50. https://doi.org/10.3390/bdcc7010050
3. Sahu, B., Mohanty, S.N., Rout, S.K. (2019). A hybrid approach for breast cancer classification and diagnosis. EAI Endorsed Transactions on Scalable Information Systems, 6(20).
4. Pritom, Ahmed Iqbal, Md Ahadur Rahman Munshi, Shahed Anzarus Sabab, and Shihabuzzaman Shihab. "Predicting breast cancer recurrence using effective classification and feature selection technique." In 2016 19th International Conference on Computer and Information Technology (ICCIT), 310-314. Bangladesh, February.
5. Mohebian, M.R., Marateb, H.R., Mansourian, M., Mañanas, M.A., Mokarian, F. (2017). A hybrid computer-aided-diagnosis system for prediction of breast cancer recurrence (HPBCR) using optimized ensemble learning. Computational and Structural Biotechnology Journal, 15, 75-85. https://doi.org/10.1016/j.csbj.2016.11.004
6. Dawngliani, M.S., Chandrasekaran, N., Lalmuanawma, S., Thangkhanhau, H. (2019). Prediction of breast cancer recurrence using ensemble machine learning classifiers. In International Conference on Security with Intelligent Computing and Big-data Services, 232-244. Cham: Springer International Publishing, December.
7. Alzu'bi, A., Najadat, H., Doulat, W., Al-Shari, O., Zhou, L. (2021). Predicting the recurrence of breast cancer using machine learning algorithms. Multimedia Tools and Applications, 80(9), 13787-13800. https://doi.org/10.1007/s11042-020-10448-w

8. Maigari, A., Zainol, Z., Xinying, C. (2025). Multi-modal stacked ensemble model for breast cancer prognosis prediction. Statistics, Optimization & Information Computing, 13(3), 1013-1034. https://doi.org/10.19139/soic-2310-5070-2100
9. Jakhar, A.K., Gupta, A., Singh, M. (2024). SELF: a stacked-based ensemble learning framework for breast cancer classification. Evolutionary Intelligence, 17(3), 1341-1356. https://doi.org/10.1007/s12065-023-00824-4
10. Kumar, M., Singhal, S., Shekhar, S., Sharma, B., Srivastava, G. (2022). Optimized stacking ensemble learning model for breast cancer detection and classification using machine learning. Sustainability, 14(21), 13998. https://doi.org/10.3390/su142113998
11. Mohamed, T.S., Khalifah, S.M. (2022). Breast Cancer Prediction: The Classification of Non-Recurrence-Events and Recurrence-Events Using Functions Classifiers. In 2022 3rd Information Technology to Enhance e-learning and Other Application (IT-ELA), 55-60. IEEE. Iraq, December.
12. Gupta, S.R. (2022). Prediction time of breast cancer tumor recurrence using machine learning. Cancer Treatment and Research Communications, 32, 100602. https://doi.org/10.1016/j.ctarc.2022.10060
13. Kaushik, A., Madhuranath, B., Rao, D., Dey, S.R., Sampatrao, G.S. (2024). Interpreting Breast Cancer Recurrence Prediction Models: Exploring Feature Importance with Explainable AI. In 2024 3rd International Conference on Artificial Intelligence For Internet of Things (AIIoT), 1-6. IEEE. India, May.
14. Mohammed, S.A., Darrab, S., Noaman, S.A., Saake, G. (2020). Analysis of breast cancer detection using different machine learning techniques. In International conference on data mining and big data, 108-117. Singapore, July.
15. Allugunti, V.R. (2022). Breast cancer detection based on thermographic images using machine learning and deep learning algorithms. International Journal of Engineering in Computer Science, 4(1), 49-56.