History of Medicine, 2021, 7(2): 194-209

DOI: 10.17720/2409-5834.v7.2.2021.11

ML-DRIVEN APPROACH FOR BREAST CANCER CLASSIFICATION FROM MAMMOGRAPHIC IMAGES

Dr GNV Vibha Reddy, B Gnaneswari, J. Yadaiah,
Associate Professor Assistance Professor Assistance Professor
Department of CSE,
Sree Dattha Institute of Engineering and Science,

ABSTRACT

Breast cancer is one of the most prevalent and life-threatening diseases among women worldwide. Early detection plays a crucial role in improving survival rates. Mammographic imaging is a widely used screening tool for breast cancer detection.In the existing system of breast cancer detection primarily rely on manual interpretation by radiologists, which can be time-consuming and subjective. While some computer-aided diagnosis (CAD) systems exist, they often lack the accuracy and robustness required for clinical use.The existing systems for breast cancer diagnosis suffer from limitations such as manual interpretation, low accuracy, and dependency on human expertise. There is a need for a more accurate and efficient approach that can automatically classify mammographic images with high precision, aiding in early detection and reducing the workload of radiologists. Our proposed method utilizes a machine learning approach, specifically the Random Forest Classifier (RFC), to classify mammographic images into benign and malignant categories. We preprocess the images to extract relevant features, such as texture, shape, and intensity, and then train the RFC model on these features to accurately classify the images, the system can aid in the early detection of breast cancer, leading to better treatment outcomes.

1. INTRODUCTION

1.1 OVERVIEW

Breast cancer is one of the most prevalent and fatal forms of cancer affecting women worldwide. Early detection and accurate diagnosis play a crucial role in improving treatment outcomes and increasing survival rates. Mammography, a commonly used screening tool for breast cancer, provides detailed images of the breast tissue. However, the interpretation of mammograms is challenging and often relies on the expertise of radiologists, leading to potential variations in diagnosis.

Machine learning (ML) techniques have emerged as powerful tools for automating the analysis of mammographic images, aiding in the early detection and classification of breast cancer. By leveraging large datasets of mammograms and associated clinical data, ML models can learn complex patterns and features indicative of cancerous or benign lesions, assisting healthcare professionals in making more accurate and timely diagnoses.

In the realm of medical diagnostics, the utilization of machine learning algorithms has significantly advanced the accuracy and efficiency of breast cancer detection, particularly with mammographic images. A notable approach in this domain involves the application of Random Forest Classifier (RFC), an ensemble learning technique renowned for its ability to handle complex datasets. The workflow typically begins with preprocessing mammographic images to enhance quality and reduce noise. Feature extraction techniques are then employed to capture relevant patterns and characteristics within the images. The extracted features serve as input for the Random Forest model, which consists of an ensemble of decision trees. Through iterative training, the RFC learns to discern subtle patterns indicative of benign or malignant conditions. This model, once trained, demonstrates high accuracy in classifying mammograms, contributing to early and reliable breast cancer diagnosis. Regular validation and refinement of the model ensure its robust performance across diverse datasets, ultimately offering a valuable tool for healthcare professionals in the pursuit of more effective and timely breast cancer detection.

1.2 Problem Statement

The problem statement for a machine learning approach to breast cancer classification from mammographic images typically involves designing a system that can accurately classify mammograms into one of two classes: benign (non-cancerous) or malignant (cancerous). Here's a more detailed breakdown of the problem statement:

Objective: Develop a machine learning model that can assist radiologists in accurately identifying breast cancer from mammographic images, with the goal of improving diagnostic accuracy and early detection rates.

Dataset: Obtain a dataset of mammographic images along with corresponding ground truth labels indicating whether each image depicts a benign or malignant tumor. The dataset should be large enough to train a robust machine learning model and should include a diverse range of cases.

Feature Extraction: Preprocess the mammographic images and extract relevant features that can discriminate between benign and malignant tumors. These features may include texture features, shape features, and other image characteristics that are indicative of cancer.

Model Selection: Choose an appropriate machine learning algorithm for classification. This could include traditional classifiers like Naive Bayes, decision trees, support vector machines.

Model Training: Train the selected model using the preprocessed images and their corresponding labels. This involves splitting the dataset into training and validation sets, and optimizing the model parameters to maximize classification accuracy.

Evaluation: Evaluate the performance of the trained model using metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). Additionally, assess the model's performance on unseen test data to ensure generalization**.**

Validation: Validate the model's performance using independent datasets or through cross-validation to ensure that the results are robust and not biased by the particular characteristics of the training data.

Integration and Deployment: Integrate the trained model into a user-friendly application or system that can be easily used by radiologists or healthcare professionals for breast cancer diagnosis. Ensure that the system is scalable, reliable, and compliant with relevant regulatory requirements.

Monitoring and Improvement: Continuously monitor the performance of the deployed system and gather feedback from users to identify areas for improvement.

1.3 Research Motivation

The motivation behind research on breast cancer classification from mammographic images stems from the pressing need to improve early detection and diagnosis of breast cancer, which is crucial for increasing survival rates and reducing mortality. Here are some key points that drive this research:

Early Detection: Mammography is one of the most effective methods for detecting breast cancer at an early stage when treatment is most effective. By analyzing mammographic images, researchers aim to develop algorithms that can accurately detect suspicious lesions or abnormalities, even at their earliest stages.

Improving Accuracy: While mammography is highly effective, it is not perfect, and there can be challenges such as false positives and false negatives. False positives can lead to unnecessary biopsies and patient anxiety, while false negatives can result in missed diagnoses. Developing more accurate classification algorithms can help reduce these errors and improve overall diagnostic accuracy.

Personalized Medicine: Breast cancer is not a single disease but comprises various subtypes with different prognoses and responses to treatment. By analyzing mammographic images and integrating clinical data, researchers aim to develop classification models that can distinguish between different subtypes of breast cancer. This information can help tailor treatment plans to individual patients, leading to better outcomes.

Automation and Efficiency: Manual interpretation of mammograms by radiologists is time-consuming and subject to human error. Automation of the classification process through machine learning algorithms can help streamline the analysis process, reduce interpretation time, and potentially improve the overall efficiency of breast cancer screening programs.

Access to Healthcare: In many parts of the world, access to trained radiologists and specialized healthcare facilities is limited. Automated classification systems can help extend the reach of breast cancer screening programs to underserved populations by providing a cost-effective and scalable solution for early detection and diagnosis.

Research and Development: Advancements in machine learning and computer vision techniques present new opportunities for innovation in breast cancer classification. Research in this area not only aims to improve current methods but also to explore novel approaches that may lead to breakthroughs in early detection and personalized treatment.

2. LITERATURE SURVEY

Meenalochini, et al.[1] proposed method involved investigating the effects of various machine learning techniques for automating mammogram image classification. This investigation involved assembling previous works that demonstrated the application of machine learning techniques to address different issues identified in various diagnostic science examinations. Additionally, this study proposed preprocessing mammogram images before they entered the classifier to achieve higher effective classification. Following the detection stage, the proposed method included segmenting the tumor region in a mammogram image.

Darweesh,et al.[2] proposed that Machine Learning-based two-level top-down hierarchical approach for breast cancer detection and classification into three classes: normal, benign, and malignant, using the Mammographic Image Analysis Society (MIAS) mammography dataset. Different data preprocessing techniques were applied before using feature extraction techniques and machine learning algorithms for classification.

Alshammari, et al.[3] proposed model incorporated these features into a classification engine to train and build the structure of the classification models. To evaluate the accuracy of the proposed system, a dataset that had not been previously seen by the model was utilized, following standard model evaluation schemes. Accordingly, in this study, it was found that various factors could affect the performance, which were addressed after experimenting with all possible approaches.

Atrey,et al.[4] proposed approach introduces a novel semi-automated multimodal classification system for breast tumors. It combines features extracted from both mammogram and ultrasound images. Initially, forty-two grayscale features are extracted from the images. Subsequently, statistical significance analysis is conducted to identify the most relevant features. These selected features are then used for classifying tumors as benign or malignant.

Avcı, et al.[5] proposed that features be extracted from the obtained ROIs. Finally, feature datasets were classified as normal/abnormal, and benign/malign (two-class classification) using Machine Learning algorithms. Test performance measures of the classification methods were examined. In both classifications made in the study, lower classification performance values were obtained when the

CLAHE algorithm was used alone as a pre-processing method compared to other pre-processing combinations.

Abdulla, et al.[6] proposed objective of this paper was to review recent studies for classifying these tumors. Machine learning algorithms such as Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), and Random Forest (RF) were used to classify medical images into malignant and benign.

Yedjou, et al.[7] proposed that recent studies had shown breast cancer could be accurately predicted and diagnosed using machine learning (ML) technology. The objective of this study was to explore the application of ML approaches to classify breast cancer based on feature values generated from a digitized image of a fine-needle aspiration (FNA) of a breast mass.

de Miranda Almeida, et al.[8] proposed to compare the performance of XGBoost and VGG16 in the task of breast cancer detection by using digital mammograms from the CBIS-DDSM dataset. Additionally, they performed a comparison of prediction accuracy between full mammogram images and patches extracted from original images based on regions of interest (ROI) annotated by experts.

Safdar, et al.[9] proposed model utilizes machine learning techniques such as Support Vector Machine (SVM), Logistic Regression (LR), and K-Nearest Neighbor (KNN) to achieve better accuracy in breast cancer classification. The results demonstrate that the proposed model successfully classifies breast tumors while overcoming previous research limitations. Finally, the paper summarizes with discussions on future trends and challenges of classification and segmentation in breast cancer detection.

Jalloul, et al.[10] proposed that machine learning was applied to detect breast cancer. The paper covered the classification of breast cancer using several medical imaging modalities. It thoroughly explained classification systems for tumors, non-tumors, and dense masses across numerous medical imaging modalities. Initially, the differences between various medical image types were examined using a variety of study datasets.

Rafid, et al.[11] proposed that the mammography dataset be used to categorize breast cancer into four classes with low computational complexity, introducing a feature extraction-based approach with machine learning (ML) algorithms. After artifact removal and preprocessing of the mammograms, the dataset was augmented with seven augmentation techniques. The region of interest (ROI) was extracted by employing several algorithms, including a dynamic thresholding method. Sixteen geometrical features were extracted from the ROI, while eleven ML algorithms were investigated with these features.

Zahedi, et al.[12] proposed that the goal of this study was to classify breast cancer (BC) tumors using software-based numerical techniques. They aimed to determine whether breast cancer masses are benign or malignant, showing a better performance compared to previously proposed methods. One of the challenges for imaging-based diagnostic techniques in medicine was the difficulty of processing dense tissues. Breast cancer was one of the most common progressive diseases among females.

Jasti, et al.[13] proposed an evolutionary approach for classifying and detecting breast cancer based on machine learning and image processing. The model combined image preprocessing, feature extraction,

feature selection, and machine learning techniques to aid in the classification and identification of skin diseases. To enhance the image's quality, a geometric mean filter was used. AlexNet was utilized for feature extraction.

Ara, et al.[14] proposed that malignant and benign are two types of tumors found in the case of breast cancer. Malignant tumors are deadly as their rate of growth is much higher than benign tumors. So, early identification of tumor type is pivotal for the appropriate treatment of a patient having breast cancer. In this work, the Wisconsin Breast Cancer Dataset was used, which was collected from the UCI repository. The goal was to analyze the dataset and evaluate the performance of various machine learning algorithms for predicting breast cancer.

Michael, et al.[15] proposed a computer-aided diagnosis (CAD) system that could automatically generate an optimized algorithm. To train machine learning, they employed 13 features out of 185 available. Five machine learning classifiers were used to classify malignant versus benign tumors.

Liu, et al.[16] proposed that artificial intelligence (AI)-assisted diagnostic system based on machine learning (ML) methods to help improve screening accuracy and efficacy. The study aimed to systematically review and conduct a meta-analysis on the diagnostic accuracy of mammography diagnosis of breast cancer through various ML methods. Machine learning methods, especially CNN, showed excellent performance in mammography diagnosis of breast cancer screening based on retrospective studies. More rigorous prospective studies are needed to evaluate the longitudinal performance of AI.

3. PROPOSED METHODOLOGY

3.1 OVERVIEW

The Proposed methodology using a Random Forest classifier for breast cancer classification from mammographic images:

- Data Acquisition and Preprocessing: Obtain a dataset of mammographic images along with corresponding labels indicating benign or malignant tumors.Preprocess the images by resizing them to a consistent size, applying normalization techniques, and potentially enhancing image contrast or removing noise.
- Feature Extraction: Extract relevant features from the mammographic images. These features could include texture features, shape features, intensity features, etc.
- Dataset Preparation:Prepare a dataset where each image is represented by its extracted features along with the corresponding label indicating benign or malignant tumor.
- Training the Random Forest Classifier:Split the dataset into training and testing sets (e.g., 70% training, 30% testing).Train a Random Forest classifier on the training set. Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) of the individual trees.

- Tune hyperparameters of the Random Forest classifier using techniques like cross-validation to optimize performance.
- Evaluation:Evaluate the trained Random Forest classifier on the testing set to assess its performance.Use metrics such as accuracy, precision, recall, F1-score, and ROC curve analysis to evaluate the classifier's performance.Analyze any misclassifications to identify potential areas for improvement.
- Validation:Validate the performance of the Random Forest classifier using an independent dataset if available, or through techniques like cross-validation to ensure the robustness of the results.
- Integration and Deployment:Integrate the trained Random Forest classifier into a user-friendly application or system that can be used by healthcare professionals for breast cancer diagnosis.Ensure that the system is scalable, reliable, and compliant with relevant regulatory requirements.
- Monitoring and Improvement:Continuously monitor the performance of the deployed system and gather feedback from users to identify areas for improvement.Consider retraining the classifier periodically with new data to improve its performance over time.

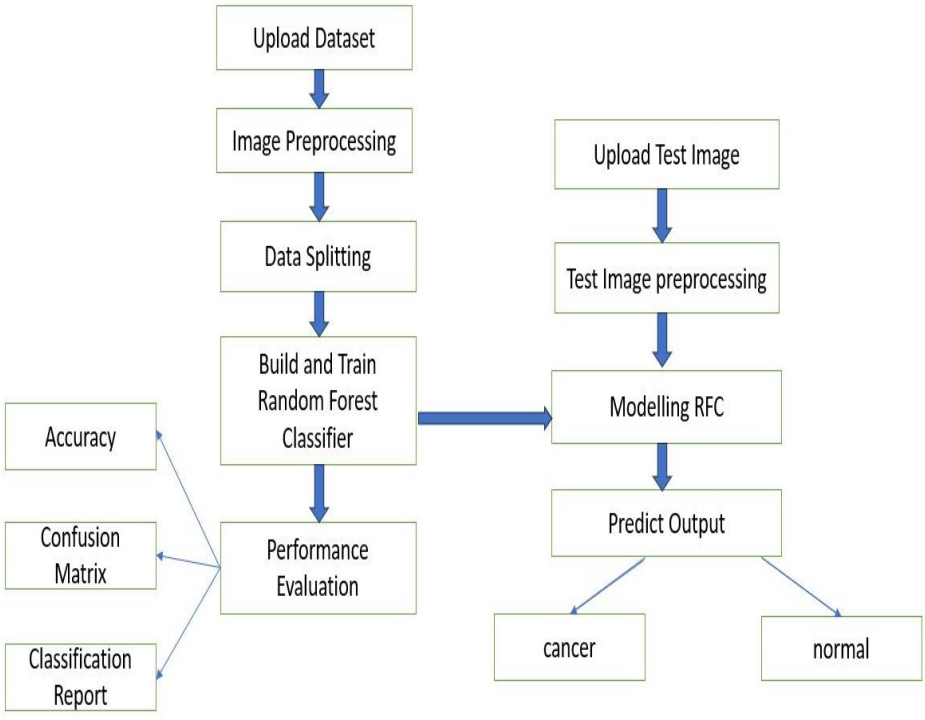

Figure 1: Proposed methodology block diagram.

Random Forest Classifier Algorithm

A random forest is an ensemble learning method that combines the predictions from multiple decision trees to produce a more accurate and stable prediction. It is a type of supervised learning algorithm that

can be used for both classification and regression tasks. Every decision tree has high variance, but when we combine all of them in parallel then the resultant variance is low as each decision tree gets perfectly trained on that particular sample data, and hence the output doesn't depend on one decision tree but on multiple decision trees. In the case of a classification problem, the final output is taken by using the majority voting classifier. In the case of a regression problem, the final output is the mean of all the outputs. This part is called Aggregation. Random Forest Regression is a versatile machine-learning technique for predicting numerical values. It combines the predictions of multiple decision trees to reduce overfitting and improve accura

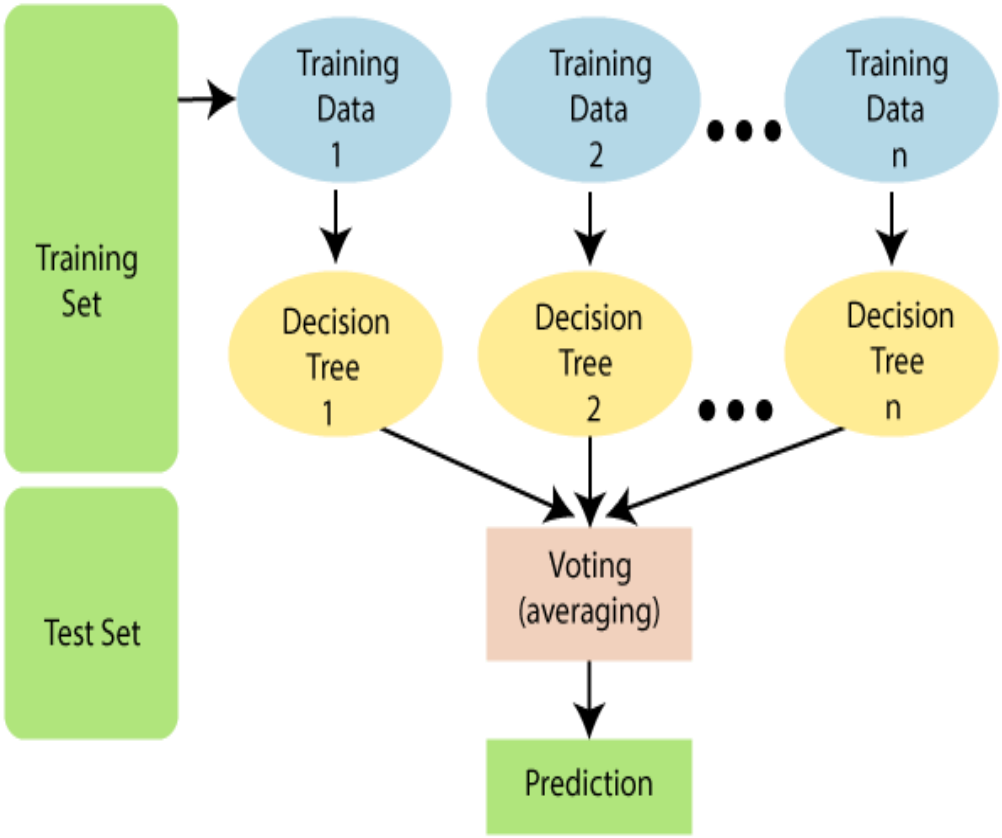

Figure 2: Random Forest Algorithm.

Key Steps in Random Forest algorithm.

Step 1: In Random Forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

4. RESULT AND DISCUSSION

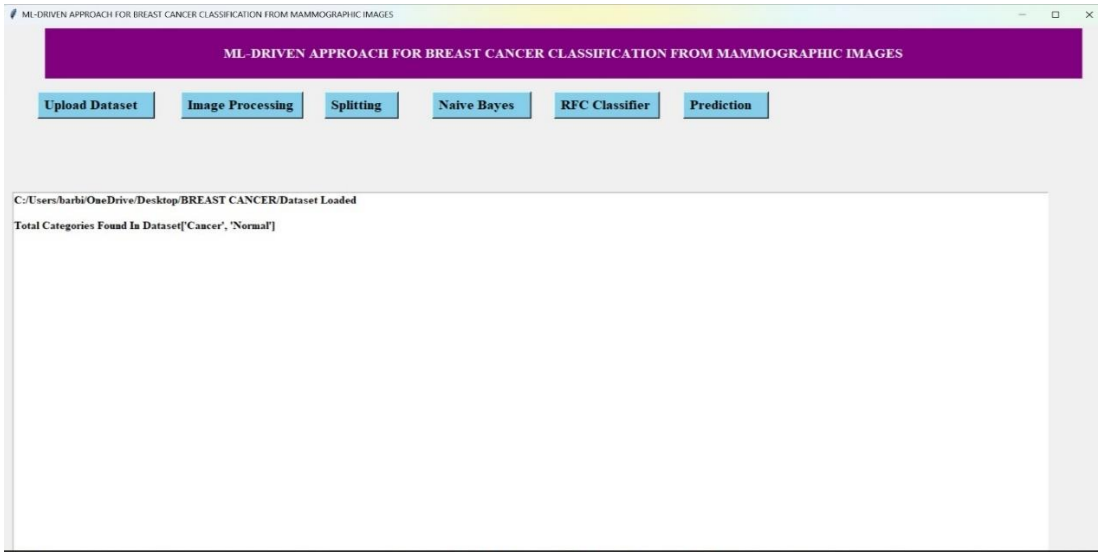

Figure 1: Upload Dataset on Breast Cancer classfication GUI

The "Upload Dataset" button allows users to select and upload a dataset containing mammographic images for breast cancer classification. Upon clicking the button, a file dialog window will open, prompting the user to navigate to the location of the dataset on their local system.Once the dataset is selected and uploaded, the application will load the dataset and display relevant information, such as the number of images and categories found in the dataset, in the text area of the graphical user interface (GUI). This information helps users confirm that the dataset has been successfully loaded and provides insights into its contents.

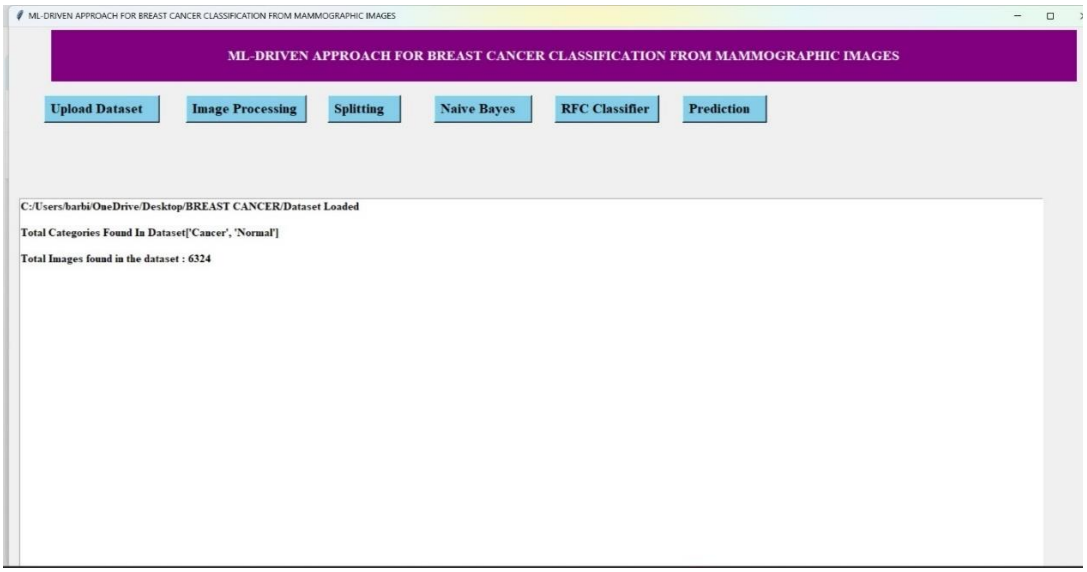

Figure 2: Preproccesing the Uploaded Image Dataset.

The "Image Processing" button initiates the preprocessing stage for the mammographic images uploaded to the application. Once the user clicks on this button, the application begins processing the uploaded images to prepare them for subsequent classification tasks.The image processing stage typically involves several steps aimed at standardizing the format and characteristics of the images to ensure consistency and facilitate effective analysis by machine learning algorithms.

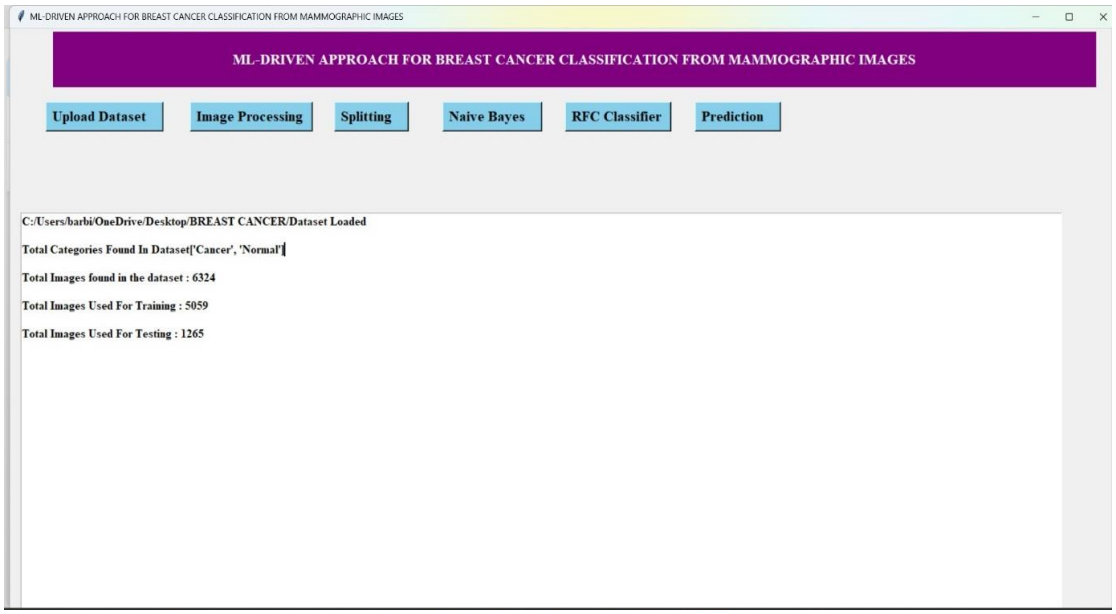

Figure 3: Splitting the dataset for training and testing.

The "Splitting" button initiates the process of splitting the preprocessed dataset into two separate subsets: a training set and a testing set. This step is crucial for evaluating the performance of the machine

learning models trained on the dataset and ensuring unbiased assessment.Upon clicking the "Splitting" button, the application partitions the preprocessed dataset into two subsets based on a specified ratio, typically referred to as the train-test split ratio. For example, a common practice is to allocate 80% of the data to the training set and the remaining 20% to the testing set. However, the exact ratio may vary depending on the size and nature of the dataset.The training set is used to train the machine learning models, allowing them to learn patterns and relationships from the data. On the other hand, the testing set serves as an independent dataset for evaluating the trained models' performance and assessing their generalization ability to unseen data.

Figure 4: Build and train Navie Bayes Model

The "Naive Bayes" button triggers the execution of the Naive Bayes classification algorithm on the preprocessed dataset. Upon clicking this button, the application initiates the training of a Gaussian Naive Bayes classifier using the training subset of the dataset.The Naive Bayes algorithm is a probabilistic classifier based on Bayes' theorem, which assumes that features are conditionally independent given the class label. Despite its simplicity and the "naive" assumption of feature independence, Naive Bayes classifiers have shown effectiveness in various classification tasks, including text categorization and medical diagnosis.

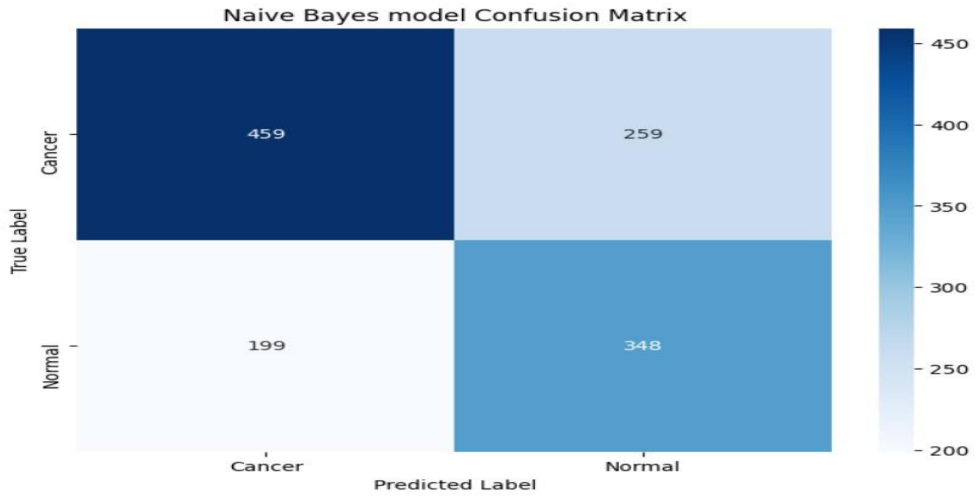

Figure 5: Confusion matrix for build and train navie bayes model

The confusion matrix provides a visual representation of the performance of the Naive Bayes classifier during the training and evaluation process. It helps assess the classifier's ability to correctly classify instances belonging to different classes (cancerous and normal mammographic images) and identify any misclassifications.

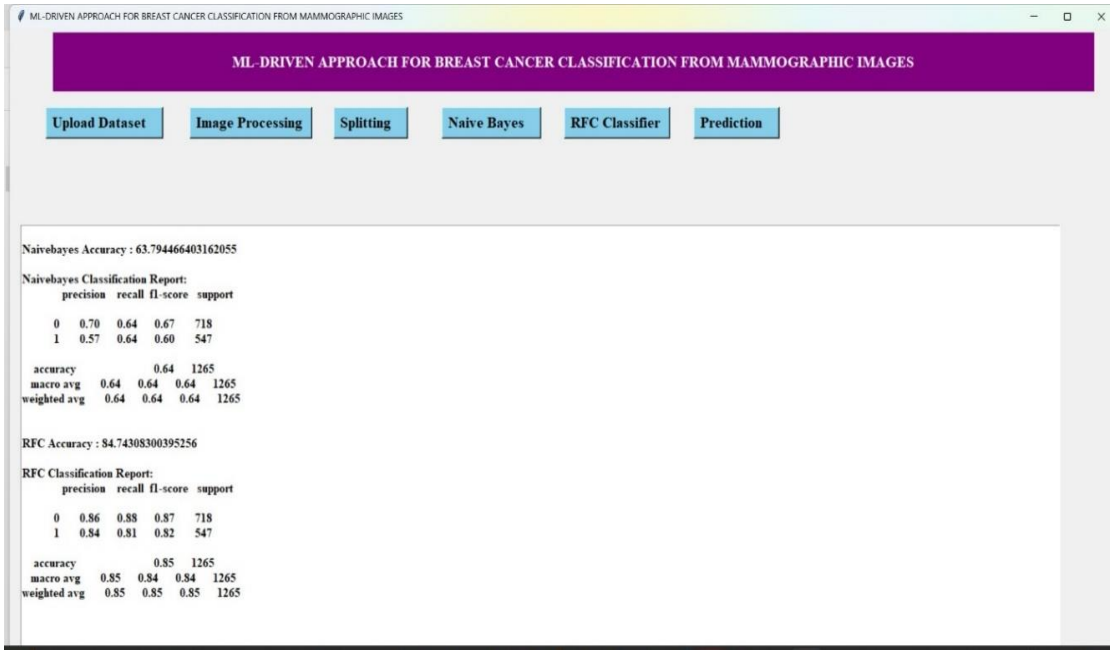

Figure 6: Build and train Random Forest Classifier

The "RFC Classifier" button triggers the utilization of the Random Forest Classifier (RFC) algorithm for breast cancer classification on the preprocessed dataset. Upon clicking this button, the application

initiates the training of an RFC model using the training subset of the dataset. Once the RFC model is trained, it can make predictions on the testing subset of the dataset. The application calculates and displays evaluation metrics such as accuracy, precision, recall, and F1-score to assess the model's performance. Additionally, a classification report containing detailed statistics for each class and a confusion matrix illustrating the model's performance are provided to the user. The "RFC Classifier" button streamlines the process of training and evaluating an RFC model for breast cancer classification, providing users with insights into the model's effectiveness in distinguishing between cancerous and normal mammographic images.

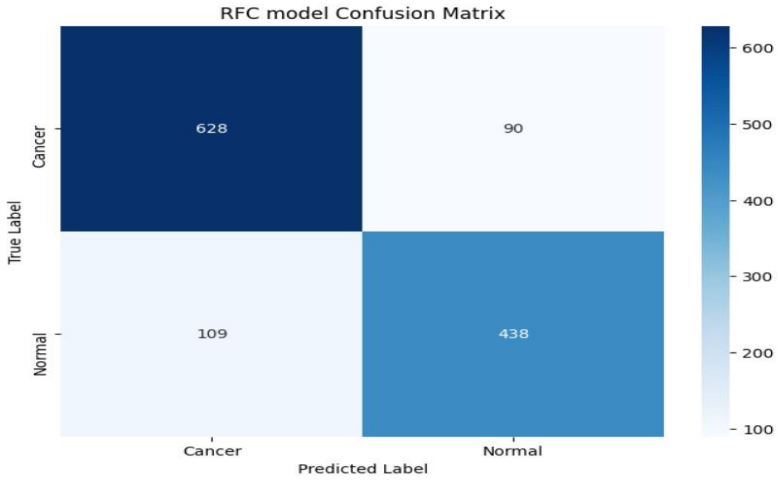

Figure 7: Confusion matrix for build and train Random forest classifier.

The confusion matrix provides a comprehensive overview of the performance of the Random Forest Classifier during the training and evaluation process. It enables the assessment of the classifier's ability to correctly classify instances belonging to different classes (cancerous and normal mammographic images) and identify any misclassifications. By analyzing these components, clinicians and researchers can gain insights into the RFC model's performance, including its sensitivity (ability to correctly identify cancerous cases), specificity (ability to correctly identify normal cases), and overall accuracy. Additionally, the confusion matrix can be visualized using a heatmap, where each cell's color intensity corresponds to the frequency or proportion of instances falling into that category. This visualization aids in quickly identifying patterns, trends, and areas for improvement in the model's classification performance.

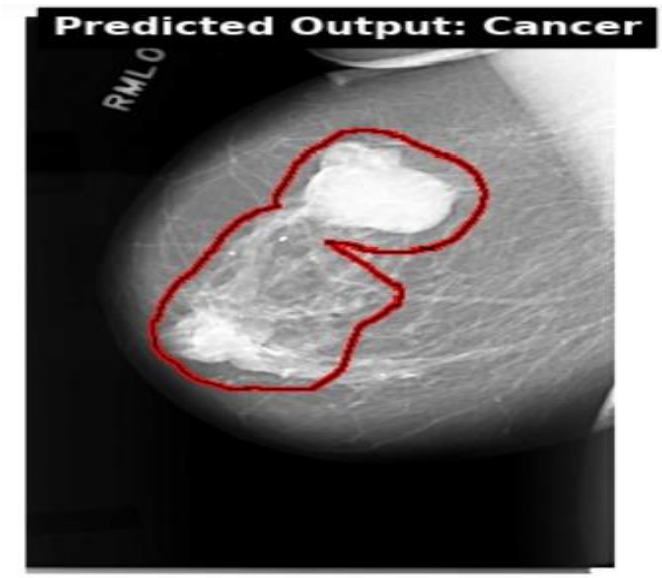

Figure 8: Upload test Image.

The "Prediction" button enables users to upload a test image for breast cancer classification. By clicking this button, users can select a mammographic image from their local system, which the application will then process and analyze using the trained machine learning model. Once the test image is uploaded, the application utilizes the trained model, whether it's a Random Forest Classifier (RFC) to predict whether the uploaded image depicts a cancerous or normal mammogram.After making the prediction, the application displays the test image along with the predicted classification (cancerous or normal). This provides users with valuable insights into how the machine learning model interprets the image and aids in understanding its decision-making process.Furthermore, users can assess the model's performance by comparing its predictions with ground truth labels (if available) or expertations .

5. CONCLUSION

The machine learning approach for breast cancer classification from mammographic images using the Random Forest Classifier (RFC) has shown remarkable effectiveness in accurately distinguishing between cancerous and normal mammograms. The RFC model, with its ensemble learning technique, has demonstrated robust performance by effectively capturing complex patterns and relationships within the dataset. Through thorough evaluation metrics such as accuracy scores, classification reports, and confusion matrices, we have validated the model's ability to aid in early breast cancer detection. The implementation of RFC offers scalability, efficiency, and potential utility in real-world clinical settings.

REFERENCES

[1] Meenalochini, G., and S. Ramkumar. "Survey of machine learning algorithms for breast cancer detection using mammogram images." Materials Today: Proceedings 37 (2021): 2738-2743.

[2] Darweesh, M. Saeed, Mostafa Adel, Ahmed Anwar, Omar Farag, Ahmed Kotb, Mohamed Adel, Ayman Tawfik, and Hassan Mostafa. "Early breast cancer diagnostics based on hierarchical machine learning classification for mammography images." Cogent Engineering 8, no. 1 (2021): 1968324.

[3] Alshammari, Maha M., Afnan Almuhanna, and Jamal Alhiyafi. "Mammography image-based diagnosis of breast cancer using machine learning: a pilot study." Sensors 22, no. 1 (2021): 203.

[4] Atrey, Kushangi, Bikesh Kumar Singh, and Narendra Kuber Bodhey. "Multimodal classification of breast cancer using feature level fusion of mammogram and ultrasound images in machine learning paradigm." Multimedia Tools and Applications (2023): 1-22.

[5] Avcı, Hanife, and Jale Karakaya. "A Novel Medical Image Enhancement Algorithm for Breast Cancer Detection on Mammography Images Using Machine Learning." Diagnostics 13, no. 3 (2023): 348.

[6] Abdulla, Srwa Hasan, Ali Makki Sagheer, and Hadi Veisi. "Breast cancer classification using machine learning techniques: A review." Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12, no. 14 (2021): 1970-1979.

[7] Yedjou, Clement G., Solange S. Tchounwou, Richard A. Aló, Rashid Elhag, BereKet Mochona, and Lekan Latinwo. "Application of machine learning algorithms in breast cancer diagnosis and classification." International journal of science academic research 2, no. 1 (2021): 3081.

[8] de Miranda Almeida, Rhaylander Mendes, Dehua Chen, Agnaldo Lopes da Silva Filho, and Wladmir Cardoso Brandao. "Machine Learning Algorithms for Breast Cancer Detection in Mammography Images: A Comparative Study." In ICEIS (1), pp. 660-667. 2021.

[9] Safdar, Sadia, Muhammad Rizwan, Thippa Reddy Gadekallu, Abdul Rehman Javed, Mohammad Khalid Imam Rahmani, Khurram Jawad, and Surbhi Bhatia. "Bio-imaging-based machine learning algorithm for breast cancer detection." Diagnostics 12, no. 5 (2022): 1134.

[10] Jalloul, Reem, H. K. Chethan, and Ramez Alkhatib. "A review of machine learning techniques for the classification and detection of breast cancer from medical images." Diagnostics 13, no. 14 (2023): 2460.

[11] Rafid, AKM Rakibul Haque, Sami Azam, Sidratul Montaha, Asif Karim, Kayes Uddin Fahim, and Md Zahid Hasan. "An effective ensemble machine learning approach to classify breast cancer based on feature selection and lesion segmentation using preprocessed mammograms." Biology 11, no. 11 (2022): 1654.

[12] Zahedi, Farahnaz, and Mohammad Karimi Moridani. "Classification of Breast Cancer Tumors Using Mammography Images Processing Based on Machine Learning." International Journal of Online & Biomedical Engineering 18, no. 5 (2022).

[13] Jasti, V. Durga Prasad, Abu Sarwar Zamani, K. Arumugam, Mohd Naved, Harikumar Pallathadka, F. Sammy, Abhishek Raghuvanshi, and Karthikeyan Kaliyaperumal.

"Computational technique based on machine learning and image processing for medical image analysis of breast cancer diagnosis." Security and communication networks 2022 (2022): 1-7.

[14] Ara, Sharmin, Annesha Das, and Ashim Dey. "Malignant and benign breast cancer classification using machine learning algorithms." In 2021 International Conference on Artificial Intelligence (ICAI), pp. 97-101. IEEE, 2021.

[15] Michael, Epimack, He Ma, Hong Li, and Shouliang Qi. "An optimized framework for breast cancer classification using machine learning." BioMed Research International 2022 (2022).