IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 1, January 2025

DOI: 10.17148/IJARCCE.2025.14122

Symptom-Based Breast Cancer Detection and Carcinoma Type Identification Using GLCM Feature Extraction and RF Classification

Ramya P M[1], Sanvitha S Acharya[2], Isha S Shetty[3], Nisha[4], Lekha[5]

Assistant Professor, Information Science and Engineering, AJIET, Mangalore, India[1]

Student, Computer Science and Engineering, AJIET, Mangalore, India [2-5]

Abstract: Early and accurate detection of breast cancer is essential for improving patient outcomes and tailoring treatment strategies. This study introduces a two-step machine learning framework for symptom-based breast cancer detection and carcinoma type identification. The initial step utilizes Random Forest (RF) to detect the presence of breast cancer based on extracted symptoms. If cancer is detected, the second step confirms the carcinoma type, specifically identifying ductal carcinoma, using Gray-Level Co-Occurrence Matrix (GLCM) for feature extraction and RF classification. The proposed system demonstrates enhanced accuracy and reliability, leveraging the strength of feature-based methods and ensemble learning techniques. This paper provides an in-depth analysis of methodologies, results, and related datasets, emphasizing the practicality and effectiveness of the system in clinical applications.

Keywords: Breast cancer detection, GLCM, Random Forest, Ductal carcinoma, Feature extraction, Machine learning, Symptom-based analysis, Classification.

I.INTRODUCTION

Breast cancer remains one of the leading causes of cancer-related mortality worldwide, with early detection being crucial to improving survival rates and treatment outcomes. Despite advancements in medical imaging and diagnostic techniques, limitations such as imbalanced datasets, high false-positive rates, and lack of holistic profiling hinder accurate and timely detection. Identifying carcinoma types, especially ductal carcinoma, is equally vital for guiding appropriate treatment plans.

Traditional diagnostic approaches often rely on radiological methods and manual interpretation, which can be error-prone and time-consuming. Additionally, these methods frequently fail to incorporate patient-specific symptoms and imaging features holistically, leading to suboptimal results. Addressing these challenges requires innovative solutions that combine machine learning (ML) techniques with feature extraction methods to enhance diagnostic precision.

Machine learning, particularly ensemble models like Random Forest (RF), offers robust capabilities in handling complex datasets and deriving meaningful insights. Furthermore, integrating feature extraction techniques such as Gray Level Co-Occurrence Matrix (GLCM) allows for capturing intricate imaging details critical to carcinoma type identification. This study aims to develop a two-stage detection framework leveraging RF and GLCM to address these gaps and advance breast cancer diagnostic practices.

II. PROBLEM STATEMENT

Breast cancer diagnosis, particularly in its early stages, poses significant challenges due to limitations in existing diagnostic methodologies. Conventional approaches often overlook patient-specific symptoms and fail to effectively manage imbalanced datasets, resulting in lower accuracy and reliability. Additionally, identifying carcinoma types such as ductal carcinoma from imaging data requires advanced techniques that can extract and classify subtle features. These gaps in detection and classification hinder timely treatment and adversely affect patient outcomes. To address these challenges, we propose a system leveraging machine learning techniques to detect breast cancer and classify carcinoma types. By integrating symptom-based analysis with feature extraction methods, the system aims to enhance diagnostic accuracy and enable personalized, data-driven treatment strategies.

III.OBJECTIVES

The primary objectives of the proposed system include:

- Developing a two-step framework for breast cancer detection using Random Forest (RF) and carcinoma type identification using GLCM feature extraction and RF classification.
- Enhancing diagnostic accuracy by incorporating patient symptoms and imaging data for holistic profiling.
- Addressing the challenge of imbalanced datasets to improve performance metrics, particularly in early detection scenarios.
- Providing reliable and efficient tools to aid clinicians in decision-making and personalized treatment planning.

IV. REQUIREMENT SPECIFICATION

Hardware Interfaces

- Laptop with Operating System: Windows 11.

Software Interfaces

- Programming Language: Python.
- Framework: Django Web Server.
- Database: MySQL.
- Development Environment: Visual Studio Code.
- Front-End Technologies: HTML, CSS.

V. SYMPTOM AND IMAGING ANALYSIS

Symptom and imaging analysis form the foundation of the proposed breast cancer detection system. This stage involves extracting critical features from patient-reported symptoms and medical imaging data for accurate diagnosis and carcinoma type classification.

Datasets are often sourced from public repositories such as Kaggle, GitHub, or medical imaging platforms. These datasets undergo pre-processing to ensure they are suitable for training and evaluating machine learning models. For this project, multiple datasets were utilized to cover a range of patient profiles, symptoms, and imaging modalities.

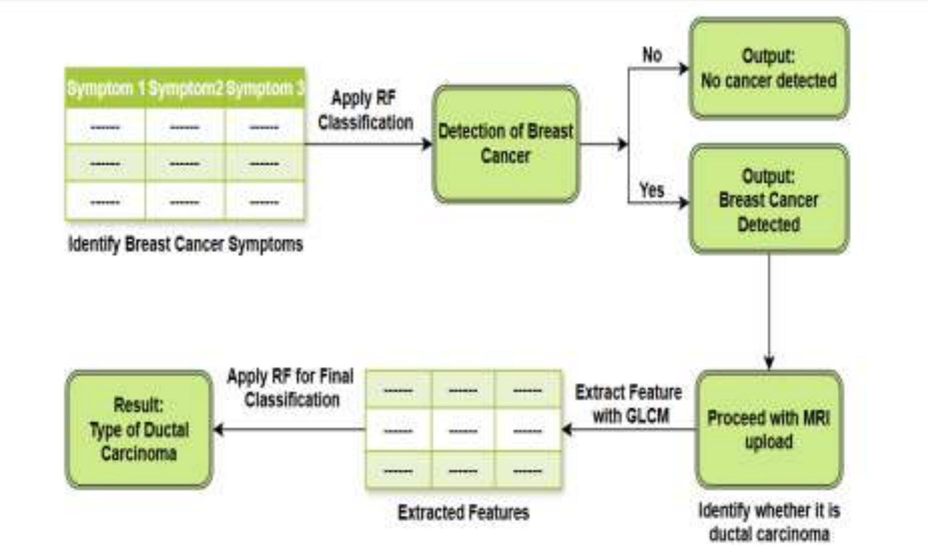
Symptom 1 Symptom2 Symptom 3
Identify Breast Cancer Symptoms
Apply RF Classification
Detection of Breast Cancer
No
Output: No cancer detected
Yes
Output: Breast Cancer Detected
Proceed with MRI upload
Identify whether it is ductal carcinoma
Extract Feature with GLCM
Extracted Features
Apply RF for Final Classification
Result: Type of Ductal Carcinoma

Fig 1. Symptom and Imaging Analysis

Dataset 1: Breast Cancer Diagnostic Data

- **Source**: UCI Machine Learning Repository.
- **Purpose**: Symptom-based breast cancer detection.
- **Attributes**:
  - ➢ Radius, Texture, Perimeter, and Area: Statistical features of cell nuclei.
  - ➢ Concavity and Symmetry: Measures of tumor irregularities.
  - ➢ Diagnosis: Binary classification (Cancerous/Non-Cancerous).

Dataset 2: Mammographic Mass Data

- **Source**: Open Access Mammography Research Platform.
- **Purpose**: Carcinoma type identification.
- **Attributes**:
  - ➢ Image texture features extracted using GLCM (contrast, correlation, energy, homogeneity).
  - ➢ Age and Family History: Patient demographics.
  - ➢ Diagnosis: Carcinoma classification labels.

Dataset 3: Breast Cancer Histopathological Images

- **Source**: Kaggle - Breast Histopathology Images.
- **Purpose**: Advanced carcinoma detection using image-based classification.
- **Attributes**:
  - ➢ High-resolution histopathological images.
  - ➢ Annotated regions highlighting malignant tissues.
  - ➢ Labels: Binary classification for malignant and benign conditions.

These datasets were chosen for their relevance, comprehensive feature coverage, and availability of diagnostic labels necessary for training and validating machine learning models. Integrating symptom-based data with imaging features ensures a holistic diagnostic approach, addressing challenges such as data imbalance and enhancing classification accuracy.

VI. MACHINE LEARNING ALGORITHMS

The proposed system employs the Random Forest (RF) algorithm for both symptom-based breast cancer detection and image-based carcinoma type identification. The workflow involves the following steps:

1. Utilize input datasets containing patient symptoms and imaging features for training the model.
2. Split the dataset into training and testing sets for model validation.
3. Build the RF model, incorporating both symptom-based features and extracted image features.
4. Train the model using the training set and validate it using the test set to measure performance.
5. Evaluate the model using accuracy, confusion matrices, precision, recall, and F1-scores.

A. Random Forest (RF)

Random Forest for Breast Cancer Detection: The RF algorithm is employed for the initial detection of breast cancer based on patient symptoms. It analyzes various clinical features, such as tumor size, shape, and patient demographics, to determine the likelihood of cancer. The robustness of RF ensures that it performs well even in the presence of imbalanced datasets and varied feature distributions.

Random Forest for Image-based Carcinoma Identification: In the second stage, RF is used to classify carcinoma types, such as ductal carcinoma, from imaging data. Features extracted from mammography or histopathological images using techniques like Gray Level Co-Occurrence Matrix (GLCM) are fed into the RF model. RF's ensemble nature enables it to accurately classify carcinoma types by leveraging the extracted texture-based features from the images.

DOI: 10.17148/IJARCCE.2025.14122

B. Grey Level Co-occurrence Matrix

A co-occurrence matrix or co-occurrence distribution (also referred to as: gray-level co-occurrence matrices (GLCMs) is a matrix that is defined over an image, the distribution of co-occurring pixel values (grayscale values, or colors) at a given offset. It is used as an approach for texture analysis with various applications including medical image analysis. Whether considering the intensity or grayscale values of the image or various dimensions of color, the co-occurrence matrix can measure the texture of the image. Because co-occurrence matrices are typically large and sparse, various metrics of the matrix are often taken to get a more useful set of features. Key steps for calculating a GLCM include:

Quantization of Gray Levels: The first step often involves reducing the number of gray levels in the image to make the GLCM size manageable, as well as to reduce the computational complexity.

Defining the Distance and Angle: The distance (d) and angle (θ) specify the spatial relationship between the pixel pairs considered for the matrix. Common angles used are 0°, 45°, 90°, and 135°.

Matrix Calculation: For each pair of pixels separated by distance d and angle θ, the occurrence of pixel intensity ‘i’ next to pixel intensity ‘j’ is counted.

VII. SYSTEM DESIGN

The provided flowchart represents the system design for the breast cancer detection system, which is centred around machine learning models for both symptom-based detection and carcinoma type identification. The design incorporates two main stages: Symptom-based Prediction and Imaging-based Classification. The webpage design mirrors these stages, offering options for users to interact with the system for symptom-based diagnosis and image-based classification.

1. Model Development

- **Central Hub**: This phase involves the development and training of machine learning models using patient symptom data and imaging features. The outputs of these models feed into both prediction (for cancer detection) and classification (for carcinoma type identification).

- **Feature Extraction**: Key features, such as symptom severity and GLCM-based imaging features, are extracted to train the models and enhance performance.

2. Symptom-Based Prediction

- **Cancer Detection**: The symptom-based model uses datasets containing patient profiles, including demographics and clinical symptoms. It predicts whether a patient is likely to have breast cancer based on these factors.

- **Prediction Flow**: The input symptoms are processed through the RF model to determine the probability of cancer presence. The model analyses tumor characteristics (mean area, mean radius, and mean smoothness) to make a diagnosis.

3. Imaging-Based Classification

- **Carcinoma Type Identification**: After detecting breast cancer in the first step, the second model classifies the carcinoma type (e.g., ductal carcinoma in-situ or invasive ductal carcinoma). This model uses features extracted from mammogram or histopathological images via GLCM.

- **Image Processing**: High-resolution imaging data is analysed to identify texture patterns indicative of carcinoma types. The RF model utilizes these patterns to classify the images and predict the carcinoma type accurately.

4. Integration

- **Comprehensive Diagnostic Support**: The outputs from both prediction and classification models are integrated to provide a holistic breast cancer diagnosis.

- **System Flow**: If cancer is detected in the first stage, the second stage confirms the carcinoma type. This integration ensures that the entire diagnostic process, from initial detection to classification, is seamless and efficient.

DOI: 10.17148/IJARCCE.2025.14122

- **User Interface**: The system offers a user-friendly interface that displays the results of both the cancer detection and carcinoma type classification processes, helping doctors and patients make informed decisions.

This design ensures that the system is capable of delivering a comprehensive, accurate diagnosis by integrating symptom-based and imaging-based models.

VIII. PERFORMANCE EVALUATION

The performance of the models developed in this project was assessed using various evaluation metrics and visual tools to ensure their accuracy and reliability in breast cancer detection and carcinoma type classification. These metrics include confusion matrices, classification accuracy, and regression metrics, providing a comprehensive view of the models' effectiveness.

1. Symptom-Based Breast Cancer Detection Model

- **Evaluation Metrics**: The model’s performance was evaluated using confusion matrices, which compare actual and predicted outcomes for cancer detection. Accuracy, precision, recall, and F1-score were also calculated to assess the classification performance.
- **Training vs. Testing Accuracy**:
  - **Training Accuracy**: Consistently high, nearing 100%, indicating that the model is effectively learning from the training data.
  - **Testing Accuracy**: Fluctuations were observed, especially when predicting unseen data, which suggests potential overfitting or insufficient generalization.
  - **Improvement Strategies**: Techniques like cross-validation or regularization could be employed to improve the model’s generalization ability and stability for real-world usage.
- **Confusion Matrix**: The confusion matrix showed strong classification accuracy, with most values along the diagonal, indicating correct predictions. The minimal off-diagonal elements reflected the model's reliability in distinguishing malignant from benign cases.

2. Carcinoma Type Classification Model (Image-Based)

- **Evaluation Metrics**: For carcinoma type classification, confusion matrices were used to assess how well the model predicted different carcinoma types. Accuracy, precision, recall, and F1-score were also computed.
- **Results**:
  - **High Precision and Recall**: The model demonstrated high precision and recall, especially for common carcinoma types like ductal carcinoma.
  - **F1-Score**: Balanced F1-scores for each carcinoma type suggested that the model was not biased toward any specific class.

3. Model Evaluation Overview

- **Confusion Matrices**: These matrices provided insight into the performance of the detection and classification models. The diagonal elements indicated the correct classification of malignant vs. benign cases and carcinoma types, while the off-diagonal elements were minimal, demonstrating good model accuracy.
- **Training vs. Testing Performance**:
  - **Consistency**: The model showed consistent results on training data, but fluctuations in testing accuracy suggested areas for improvement in generalization to new, unseen data.
  - **Overfitting**: Observed fluctuations in accuracy highlighted the possibility of overfitting, suggesting the need for regularization techniques to improve real-world performance.

DOI: 10.17148/IJARCCE.2025.14122

4. Performance Evaluation Tools

- **Accuracy Metrics**: Accuracy, precision, recall, and F1-score were calculated for both the symptom-based detection and carcinoma classification tasks to assess overall classification effectiveness.
- **Confusion Matrices**: These matrices were instrumental in visualizing the true positives, false positives, true negatives, and false negatives, providing a deeper understanding of model performance.

5. Recommendations for Future Improvements

- **Data Augmentation**: Implementing image augmentation techniques to improve the training dataset for image-based classification.
- **Model Optimization**: Further tuning of the Random Forest parameters (e.g., tree depth, number of trees) could improve model accuracy, particularly for the testing dataset.
- **Cross-Validation**: Using k-fold cross-validation would provide more robust evaluation by training the model on multiple subsets of the data.

IX. RESULT

The results of the breast cancer detection system are demonstrated through a user-friendly web interface, which showcases the predictions and classifications generated by the machine learning models. These interfaces were designed to ensure an intuitive and seamless experience, enabling users—such as doctors, clinicians, and researchers—to access valuable insights effortlessly.

Web Interface Functionalities:

1. **Symptom-Based Breast Cancer Detection**:
   - This page displays whether a breast cancer case based on the input symptoms and associated features. After entering the necessary symptoms and other relevant information, the model processes the data and provides a classification result with an associated confidence score.
2. **Carcinoma Type Identification (Image-Based)**:
   - For this functionality, users can upload mammogram images, and the system processes them to classify the carcinoma type. This result is displayed alongside a confidence score for the prediction, providing insights into the type of cancer present based on image features.

Pages and Key Features:

- **Homepage**: The homepage provides an overview of the tool's features and allows easy access to the symptom-based detection and image-based carcinoma identification systems. It offers clear sign-up and login options for both medical professionals and administrators, ensuring secure access to personalized features and management tools.
  - **Symptom-Based Detection**: Enter symptoms and characteristics for breast cancer prediction.
  - **Image Upload and Carcinoma Classification**: Upload mammogram images to classify the carcinoma type.
- **Login Page:** A simple, secure login page allows users to access their accounts by entering a registered email Id and password. If users forget their credentials, the “Forgot Password” link facilitates password recovery.
- **Symptom Based Breast Cancer Detection Result:** After entering the symptoms, the system processes the data and displays the result along with the model's confidence level. This page helps users understand the likelihood of breast cancer based on symptom input.
- **Carcinoma Type Classification Result:** For users uploading mammogram images, the page displays the predicted carcinoma type (e.g., ductal carcinoma in-situ or invasive ductal carcinoma) and the model’s confidence score. This result is accompanied by additional insights on the characteristics of the detected carcinoma type.

X. FUTURE WORK

1. **Enhanced Feature Selection and Model Optimization**
   - Incorporating advanced feature selection techniques to identify the most significant parameters for carcinoma detection and classification.
   - Implementing hyperparameter tuning strategies for the Random Forest (RF) classifier to enhance its accuracy and robustness.
2. **Integration of Deep Learning Models**
   - Utilizing convolutional neural networks (CNNs) for automated feature extraction and improved classification accuracy, especially for complex imaging data.
   - Comparing RF-based results with deep learning models to assess performance and scalability.
3. **Multi-class Classification for Carcinoma Subtypes**
   - Expanding the system to classify multiple carcinoma subtypes with finer granularity, aiding in more precise diagnostic workflows.
   - Incorporating advanced algorithms to handle multi-class imbalances effectively.
4. **Real-time Diagnostic Tool Development**
   - Creating a cloud-based platform that integrates the detection system for real-time image analysis and decision-making.
   - Optimizing the system for edge computing to facilitate low-latency analysis in resource-limited healthcare setups.
5. **Integration of Patient Data for Personalized Diagnosis**
   - Combining imaging data with patient history to provide holistic and personalized diagnostic insights.
   - Developing predictive models to evaluate the risk of recurrence or metastasis based on patient-specific data.
6. **Mobile and Cross-platform Application**
   - Developing a mobile application to provide medical professionals and patients with easy access to diagnostic insights.
   - Including an offline mode for remote or rural areas with limited internet connectivity.
7. **Collaboration with Healthcare Systems**
   - Collaborating with hospitals and research institutions to validate the system using real-world datasets.
   - Integrating the tool with existing hospital management systems for streamlined workflows.
8. **Explainable AI for Transparency**
   - Developing explainable AI (XAI) techniques to help radiologists and clinicians understand the system's decision-making process.
   - Building trust and facilitating system adoption by healthcare professionals.

XI. CONCLUSION

The proposed breast cancer detection system using the Random Forest (RF) classifier and carcinoma type identification leveraging Gray-Level Co-occurrence Matrix (GLCM) feature extraction offers a robust and accurate solution for early detection and classification. By integrating machine learning and image processing techniques, the system effectively identifies breast cancer types based on imaging data, improving diagnostic accuracy and supporting clinical decision-making.

Key features such as the extraction of textural features through GLCM and the reliable performance of the RF classifier make the system both efficient and scalable for diverse medical datasets. The system addresses critical challenges faced in healthcare, including delayed diagnosis and limited resources for specialized cancer detection.

This comprehensive approach enhances early detection and allows for tailored treatment strategies, ultimately improving patient outcomes. Future work will focus on integrating advanced machine learning and deep learning techniques, expanding multi-class classification capabilities, and ensuring the system’s accessibility and usability in real-world medical environments. By addressing these aspects, the system aims to bridge the gap between technology and clinical needs, driving advancements in cancer diagnosis and contributing to global efforts in combating breast cancer effectively.

REFERENCES

[1] U. Uniyal, “Ultrasound RF time series for classification of breast lesions,” IEEE Trans. Med. Imaging, vol. 34, no. 5, pp. 786-793, May 2015.

[2] S. Zhang, Y. Liu, W. Liao, R. Ron Zee Tan, R. Bi, and M. Olivo, “Ex vivo tissue classification using broadband hyperspectral imaging endoscopy and artificial intelligence: A pilot study,” IEEE J. Biomed. Health Inform., vol. 29, no. 3, pp. 445-455, Mar. 2022.

[3] F. Azour and A. Boukerche, “Design guidelines for mammogram-based computer-aided systems using deep learning techniques,” IEEE Access, vol. 7, pp. 179870-179880, 2020.

[4] U. Haq, J. P. Li, I. Khan, B. L. Y. Agbley, S. Ahmad, M. I. Uddin, W. Zhou, S. Khan, and I. Alam, “DEBCM: Deep learning-based enhanced breast invasive ductal carcinoma classification model in IoMT healthcare systems,” IEEE Access, vol. 9, pp. 152206-152225, 2021.

[5] Anaya-Isaza, L. Mera-Jiménez, J. M. Cabrera-Chavarro, L. Guachi-Guachi, D. Peluffo-Ordóñez, and J. I. Rios-Patiño, “Comparison of current deep convolutional neural networks for the segmentation of breast masses in mammograms,” IEEE Access, vol. 9, pp. 152206-152225, 2021.

[6] R. Martínez-Licort, C. de la Cruz León, D. Agarwal, B. Sahelices, I. de la Torre, J. P. Miramontes-González, and M. Amoon, “Breast carcinoma prediction through integration of machine learning models,” IEEE Access, vol. 8, pp. 156870-156880, 2021.

[7] U. Naseem, J. Rashid, L. Ali, J. Kim, Q. E. U. Haq, M. J. Awan, and M. Imran, “An automatic detection of breast cancer diagnosis and prognosis based on machine learning using ensemble of classifiers,” Int. J. Comput. Sci. & Appl., vol. 30, pp. 45-53, 2022.

[8] S. Sharmin, T. Ahammad, M. A. Talukder and P. Ghose, "A Hybrid Dependable Deep Feature Extraction and Ensemble-Based Machine Learning Approach for Breast Cancer Detection," in *IEEE Access*, vol. 11, pp. 87694-87708, 2023.

[9] R. Hou, M. A. Mazurowski, L. J. Grimm, J. R. Marks, L. M. King, C. C. Maley, E. S. S. Hwang, and J. Y. Lo, “Prediction of upstaged ductal carcinoma in situ using forced labeling and domain adaptation,” IEEE Trans. Med. Imaging, vol. 40, no. 8, pp. 2101-2113, Aug. 2021.

[10] M. Abd-Elnaby, M. Alfonse, and M. Roushdy, “Classification of breast cancer using microarray gene expression data: A survey,” J. Comput. Sci. & Technol., vol. 38, pp. 13-24, 2021.

[11] H. D. Chenga, J. Shana, W. Jua, Y. H. Guoa, and L. Zhang, “Automated breast cancer detection and classification using ultrasound images,” IEEE Trans. Med. Imaging, vol. 28, no. 4, pp. 754-764, Apr. 2009.

[12] M. K. Muttair and M. Z. Lighvan, “Breast cancer classification utilizing deep learning techniques on medical images: A comprehensive review,” J. Med. Imaging, vol. 8, no. 1, pp. 125-138, Jan. 2022.