

Multimodal Artificial Intelligence for Breast Cancer Detection Using Medical Imaging and Clinical Data

Thota Pallavi¹, Allam Thriveni², perumallapalli Sri Sai Snigdha³, Vavilala Divya Raj⁴

¹Thota Pallavi, Department of Computer Science and Engineering,
G. Narayanamma Institute of Technology and Science, Hyderabad, India
pallavithota338@gmail.com

²Allam Thriveni, Department of Computer Science and Engineering,
G. Narayanamma Institute of Technology and Science, Hyderabad, India
allamthriveni3@gmail.com

³Perumallapalli Sri Sai Snigdha, Department of Computer Science and Engineering,
G. Narayanamma Institute of Technology and Science, Hyderabad, India
psnigdha13@gmail.com

⁴Vavilala Divya Raj, Department of Computer Science and Engineering,
G. Narayanamma Institute of Technology and Science, Hyderabad, India
divyalaraj.vavilala@gnits.ac.in

Abstract - Breast cancer diagnosis typically involves both medical images and clinical text reports, but many existing AI systems analyze only one type of data at a time. This makes it difficult to connect information from scans and clinical notes, which can limit diagnostic accuracy. In this paper, we propose an AI-based multimodal framework for breast cancer detection that integrates both imaging data and textual information. The system supports text-only input, image-only input, or both, which are processed using separate encoders and combined through a fusion mechanism. By learning a shared representation of these data types, the proposed approach improves breast cancer classification, tumor localization, and clinical decision support, making the results more accurate and reliable.

Key Words: Multimodal Learning, Medical Image Analysis, Tumor Classification, Multimodal Fusion, Clinical Decision Support.

I. INTRODUCTION

Breast cancer is one of the most frequent malignancies among women worldwide, and the early diagnosis is necessary to keep the survival rate high. In general, it requires interpreting mammography images and clinical records findings, which is a time-consuming and challenging process. Artificial Intelligence (AI) algorithms have demonstrated extraordinary potential in many fields, especially in the medical sphere. The reason for that is the possibility to utilize different types of data

for finding patterns that can help doctors make accurate diagnoses.

However, most of the already existing solutions are only focused on one type of information, such as images or text. Although that approach has its advantages, it cannot provide an accurate overview of the situation, since there is no connection between different data modes. Some of the solutions are trying to establish that association and fusion features from text and images, but they lack the required effectiveness in generating plausible reasoning.

That is why, in this paper, the author presents a novel multimodal AI algorithm for detecting breast cancer. It uses free text from the accompanying clinical records and mammography images for better prediction of tumor locations. The models can accept both text and images as input or utilize them individually by concatenating their respective text encoders. Then, the features are combined and enhanced to make the breast cancer classification and localization more effective, which contributes to an improved outcome of the clinical examination.

II. LITERATURE REVIEW

Recent advances in biomedical AI are focused on developing generalizable models that can perform multiple medical tasks. Such models as Med-PaLM M leverage multimodal data such as medical images and clinical text. However, they still encounter challenges such as limited dataset diversity and lack of real-world clinical validation.

Explainable AI is becoming essential in medical diagnosis to ensure transparency in decision-making.

Techniques like Grad CAM help visualize important regions in medical images for cancer detection, but multimodal explainable approaches are still limited.

Multimodal deep learning models that combine medical images and patient data have been applied for cancer diagnosis and prognosis prediction. While they improve prediction performance, challenges persist due to heterogeneous data sources and limited model interpretability.

The integration of natural language processing with medical imaging through multimodal fusion can improve clinical decision-making. However, these approaches are limited by insufficient high-quality datasets and misalignment between textual and image features.

In breast cancer diagnosis, multimodal fusion combined with explainable AI has demonstrated improved diagnostic performance and clinical support. However, standardized evaluation methods and further clinical validation are still required.

III. METHODOLOGY

Figure 1 presents the working principle of the proposed deep learning-based breast cancer detection system using mammogram images. The figure illustrates the stepwise approach performed to process the input mammogram image, extract features using a convolutional neural network, classify the tumor label, and visualize the region of interest using visualization algorithms.

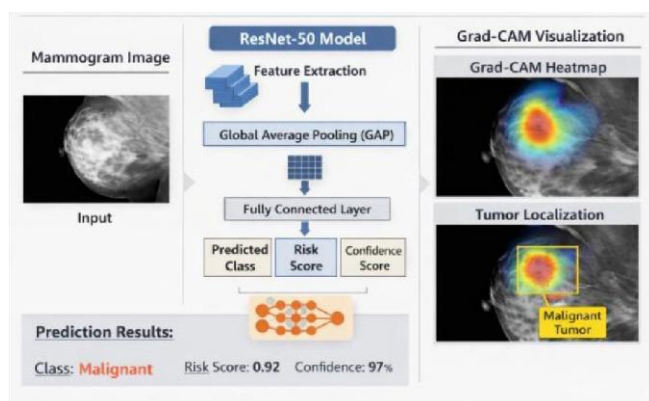


Fig – 1: ResNet-50 for Breast Cancer Detection

A. Input Layer – Mammogram Image

The process starts with a mammogram image of the breast, which is commonly used for an early detection of breast cancer. Mammography assists in the identification of abnormal masses or tissue structures that could lead to tumor formations. This image will then be supplied to the deep learning system as an input for processing.

B. Image Preprocessing

Before the image is passed to the deep learning model, a preprocessing step is carried out to improve the quality and consistency of the input data. Preprocessing helps to remove noise and ensures that all images are in a suitable format for the model.

The preprocessing operations usually comprise such step as

- image resizing to fit the input size required by the model
- pixel value normalization to improve learning efficiency
- noise removal and contrast enhancement steps to highlight relevant patterns of breast tissue.

This phase is critical to ensure that all mammogram images are standardized to extract features accurately.

C. Feature Extraction – ResNet-50 Model

After the preprocessing step, the mammogram image is fed into the ResNet-50 model, which is a deep convolutional neural network that automatically extracts relevant features during processing. This particular model comprises multiple convolutional layers and residual blocks that transform the input mammogram image.

In the process, the network will recognize distinct features like:

- Presence of tissue patterns that resemble normal/abnormal tissues
- Unique tumor shapes that could signify cancerous formation
- Unusual tissue appearances indicating abnormalities

The convolutional layers enable the network to effectively learn and extract features to differentiate healthy tissues from cancerous regions in the mammogram.

D. Fully Connected Layer

The feature vector resulting from the GAP layer is then fed into a Fully Connected (Dense) Layer responsible for final classification.

This layer analyzes the extracted features to classify the mammogram image as either containing a **Benign Tumor** (non-cancerous tissue growth) or a **Malignant Tumor** (cancerous tumor requiring medical attention). This layer represents the decision-making stage of the model.

E. Multimodal Feature Fusion

The suggested architecture takes into account information from two different modalities, namely mammography images and clinical text. In particular,

the image part of the architecture uses a deep CNN-based model to extract visual features from the processed mammogram image, and the clinical text branch extracts features from the input clinical text to get its representation.

After extracting features from the different modalities, these features are combined to produce one multimodal representation. Within the proposed approach, the feature vectors are concatenated to keep information from both modalities. Thus, the combined feature vector includes complementary visual and textual information and it is then used in the classification stage.

The multimodal fusion procedure can be depicted as follows:

Image $\rightarrow$ Preprocessing $\rightarrow$ Image Encoder $\rightarrow$ Image Features

Clinical Text $\rightarrow$ Text Preprocessing $\rightarrow$ Text Encoder $\rightarrow$ Text Features

Image Features + Text Features $\rightarrow$ Feature Concatenation $\rightarrow$ Fused Representation $\rightarrow$ Classifier $\rightarrow$ Prediction

This kind of the multimodal fusion allows taking into account the visual and contextual information and helps to make the classification process more effective.

F. Prediction Output Layer

After classification, the system produces the final prediction results. The outputs include:

Predicted Class: indicates whether the tumor is benign or malignant

Risk Score: represents the likelihood of a cancer

Confidence Score: reflects how confidently the model is predicting the result. For example, in the diagram:

Predicted Class: Malignant

Risk Score: 0.92

Confidence Score: 97%

These outputs can help clinicians assess the likelihood of breast cancer.

G. Explainable AI Visualization – Grad-CAM

In order to increase the model's interpretability, we apply a Grad-CAM visualization, which highlights the regions of the mammogram image that were most influential on the model's decision.

Two types of visualizations are generated:

- **Grad-CAM Heatmap:** A colored map showing important areas influencing the model's decision

- **Tumor Localization:** A bounding box highlighting the detected tumor region.

This visualization helps doctors understand which region of the mammogram led to the malignant prediction, improving transparency and trust in AI-based breast cancer diagnosis.

IV. RESULTS

A. Input Mammogram

Fig. 2 shows an example of a mammogram image that serves as the input to the breast cancer detection system. The image is characterized by the presence of breast tissue structures which are further analyzed by the deep learning model to detect potential abnormalities.

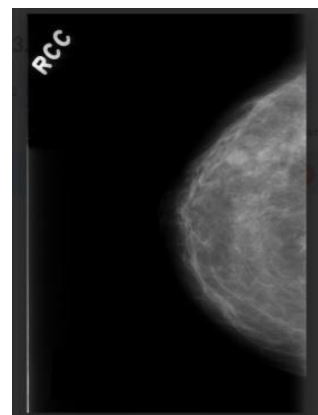


Fig – 2: Sample Diagnostic Image

B. Preprocessed Image

Fig. 3 below is the preprocessed mammogram image after the rescaling, normalization, and noise reduction stages. This stage involves preprocessing the input image to enhance the quality of the input data to the model.

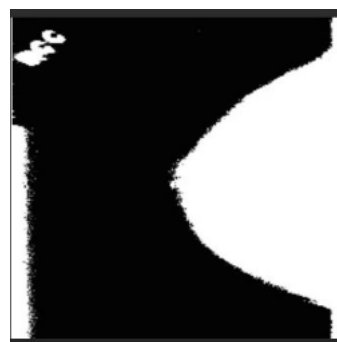


Fig – 3: Preprocessed image

C. Tumor Localization

Fig. 4 below is the tumor localization image with the heat map highlighting the region spotted by the Grad-CAM algorithm. The boxed area is the region that prompted suspicion of a malignant tumor in the model's output.

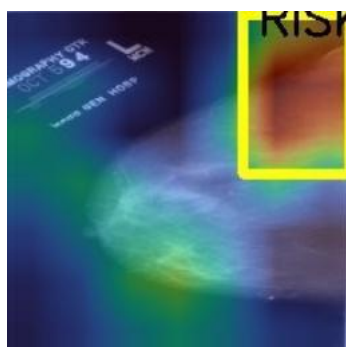


Fig- 4 :Localized image for Breast cancer

D. Confusion Matrix

The confusion matrix represents the performance of the model on the basis of the mammogram image for classifying breast cancer. This matrix provides the comparison of actual labels with predicted labels indicating the number of correctly and incorrectly classified benign and malignant tumors in fig 5.

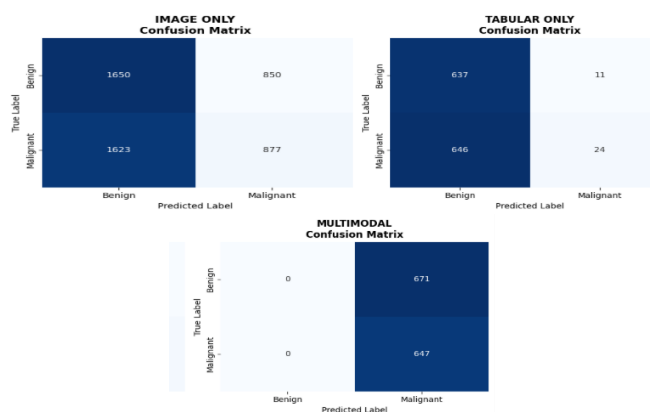


Fig – 5: Confusion Matrix for Image-Only Breast Cancer Detection Model

E. ROC Curve

Fig 6 represents the performance of different models used for detecting breast cancer like image only, text only, and multimodal. The comparison was done on the basis of true positive rates and false positive rates which conclude that multimodal performs slightly better in the classification than individual models.

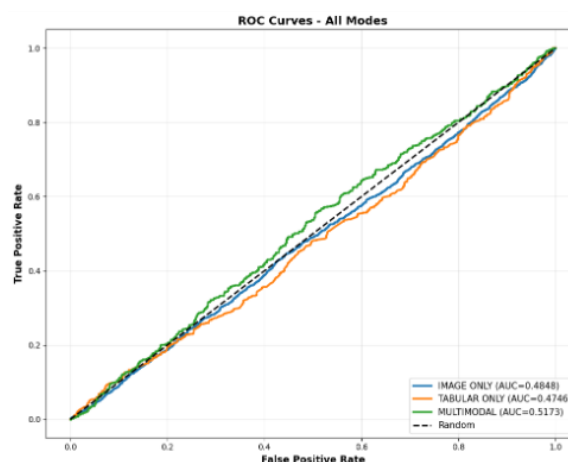


Fig- 6 :ROC Curve for Breast Cancer Detection Models

F. Performance Metrics

Table 1 compares the performances of the text-only, image-only, and fusion models in terms of accuracy, recall, precision. It can be observed that the fusion model outperforms individual models in most of the evaluation metrics, including a higher recall and F1-score.

TABLE-1: ANALYSIS OF BREAST CANCER DETECTION MODELS

Model	Accuracy	Precision	Recall
Text only	0.5015	0.6857	0.3508
Image only	0.5015	0.5078	0.3508
(Fusion)Text+Image	0.4909	0.4909	1.0000

G. Algorithm Comparison

Table 2 presents a comparison of the results of different deep learning algorithms, including ResNet18, ResNet50, and EfficientNetB3, used for breast cancer detection. Based on the accuracy and F1-score metrics, it can be concluded that

ResNet50 outperforms other algorithms in terms of performance.

TABLE - 2 : BREAST CANCER ALGORITHM COMPARISON

Algorithm	Accuracy	Precision	Recall	F1-score
ResNet18	0.71	0.775	1.0	0.873
ResNet50	0.725	0.784	0.987	0.987
EfficientNetB3	0.595	0.773	0.993	0.87

H. User Interface

Fig. 7 presents the interface of the text-only model used for breast cancer information and query analysis. Users can input text-based questions related to breast cancer, and the system analyzes the input to produce informative responses that enhance understanding of the disease.



Fig- 7: Text-Only Model Interface

Fig. 8 presents the interface of the image-only model, which was utilized for the diagnosis of breast cancer. In this model, users can upload a mammogram image, which will be analyzed by the model to detect abnormalities and predict breast cancer.

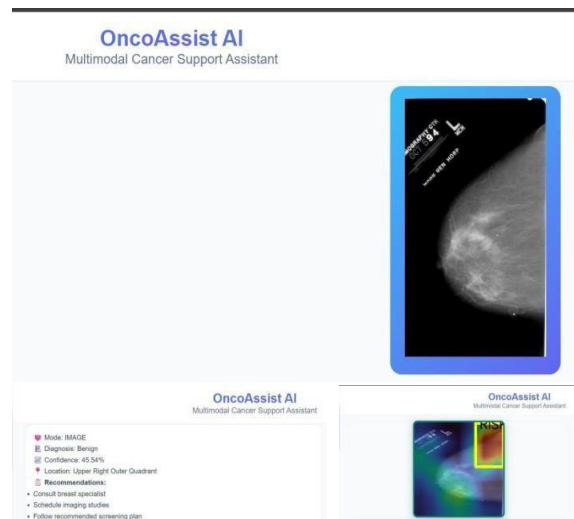


Fig- 8: Image-Only Model Interface

Fig. 9 presents the interface of the multimodal model, where a user can enter text and upload a mammogram image for improved breast cancer prediction.

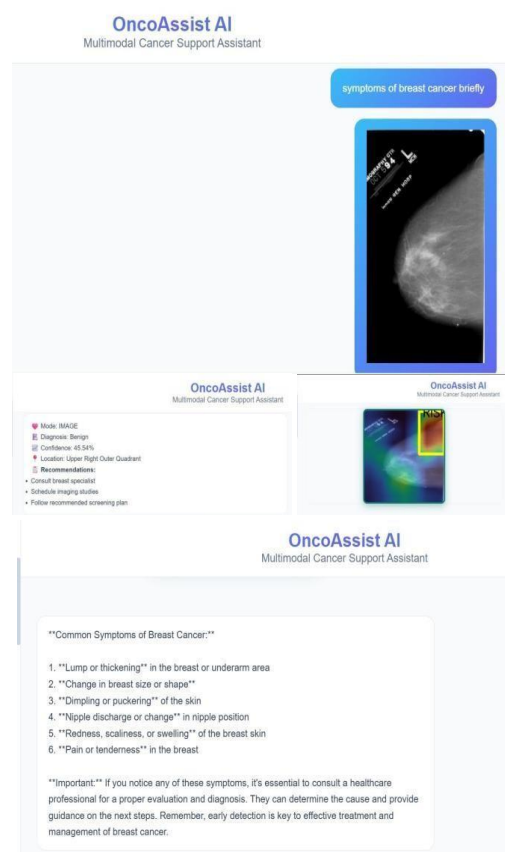


Fig- 9. Multimodal Model Interface

V. DISCUSSIONS

The results obtained through the application of the multimodal machine learning model show promising results in the context of the use of the multimodal framework for breast cancer detection on the basis of mammogram images and clinical text information. The comparative analysis of the performance of the text-based, image-based, and multimodal fusion models suggests that the two modalities can bring complementary information for the problem of breast cancer detection.

The fusion model was able to achieve recall of 1.0000 and F1-score of 0.6585 on the dataset under evaluation. The high value of recall suggests that the model was able to detect all positive cases. Nevertheless, the accuracy and precision of the multimodal model reached only 0.4909, which indicates that there were quite a number of mispredictions among those. Therefore, while demonstrating the potential of the multimodal classification of images and text, the current results prove that further optimization of the process is needed to enhance classification accuracy.

The comparison of the results obtained through the use of the three image-based deep learning models shows that the highest accuracy among them (0.725) was achieved by ResNet50 architecture, followed by ResNet18 (0.71) and EfficientNetB3 (0.595). The comparison of the results proves that the use of ResNet50 allowed for the most efficient feature extraction from the medical images. Thus, the choice of deep learning architecture plays an important role in medical image classification problems.

The Grad-CAM visualization allows improving the explanation of the results obtained through the use of image-based models. Grad-CAM generates heatmaps showing the regions of the image responsible for the prediction made by the model, while the localization model highlights suspicious regions associated with the prediction.

Thus, the integration of medical imaging, clinical text, deep learning, and explainable AI technologies offers a promising framework for breast cancer diagnosis assistance. However, the preliminary results reported above demonstrate the limitations of the fusion model in terms of precision and overall accuracy. Therefore, further experiments involving larger and more diverse dataset, better multimodal feature alignment, hyperparameters optimization, and validation are needed. The suggested system should be used for the purposes of supporting the clinical decision making but cannot be regarded as the means of replacement of professional medical diagnosis.

VI. CONCLUSIONS

A breast cancer detection and analysis system that is based on the intelligence framework of mammogram images and clinical text information. This system enables the user to upload an image or enter text and get the diagnostic results as output. In this system, the results are displayed as the risk score, confidence level, severity level, and localisation of tumors using heatmaps and bounding boxes to indicate the suspicious regions in the image. Finally, the system also provides an AI chatbot to answer the questions and provide relevant information about breast cancer. The chatbot and results are stored in a database, and the user can view their history of analysis and queries in a history page.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to K. Ramesh Reddy, Principal of G. Narayanamma Institute of Technology and Science for providing the facilities to carry out this work. The authors also thank M. Seetha, N. Kalyani, and A. Sharada for their valuable support and guidance. Special thanks are extended to K. L. S. Soujanya, the project coordinators, and the internal guide for their encouragement and assistance throughout the project. The authors also thank the faculty, staff, parents, and friends for their support in completing this work.

REFERENCES

- [1] Acosta J.N., Falcone G.J., Rajpurkar P., et al. (2022) Integrated Multimodal Artificial Intelligence Framework for Healthcare Applications. PubMed.
- [2] Baltrušaitis T., Ahuja C., Morency L.P. (2022) Multimodal Learning for Medical Decision Support Systems. Pattern Recognition Letters (Elsevier).
- [3] Chen Z., Wang Y., et al. (2023) Vision–Language Models for Medical Imaging: Survey. ACM Computing Surveys.
- [4] Huang S., Yang J., et al. (2021) Multimodal Deep Learning for Precision Oncology. IEEE Access.
- [5] Khan S., Hussain M., et al. (2022) A Systematic Review of Multimodal Fusion and Explainable AI in Breast Cancer Diagnosis. Computer Modeling in Engineering & Sciences.
- [6] Li X., et al. (2024) Vision–Language Foundation Models for Medical Imaging: A Review. Elsevier.
- [7] Singhal K., Azizi S., Tu T., et al. (2024) Towards Generalist Biomedical AI. NEJM AI.
- [8] Tjoa E., Guan C. (2020) Explainable AI for Cancer Diagnosis Using Deep Learning, Artificial Intelligence in Medicine (Springer).
- [9] Zhang Y., Jiang M., et al. (2023) Integrating Language into Medical Visual Recognition and Reasoning: A Survey. PubMed.

