Deep Multi-modal Breast Cancer Detection Network

Noor Ul Huda Shah, Tanveer Hussain, Amr Ahmed, Yonghuai Liu, Usman Ali*, Ardhendu Behera

Department of Computer Science, Edge Hill University, Ormskirk, England

Emails: {26263114, hussaint, , ahmeda, liuyo, beheraa}@edgehill.ac.uk

*Computer Science and Engineering, Sejong University, Seoul, South Korea

*Email: usman.ali@sejong.ac.kr

Abstract—Automated breast cancer detection via computer vision techniques is challenging due to the complex nature of breast tissue, the subtle appearance of cancerous lesions, and variations in breast density. Mainstream techniques primarily focus on visual cues, overlooking complementary patient-specific textual features that are equally important and can enhance diagnostic accuracy. To address this gap, we introduce Multi-modal Cancer Detection Network (MMDCNet) that integrates visual cues with clinical data to improve breast cancer detection. Our approach processes medical images using computer vision techniques while structured patient metadata patterns are learned through a custom fully connected network. The extracted features are fused to form a comprehensive representation, allowing the model to leverage both visual and clinical information. The final classifier is trained based on the joint features embedding space of visual and clinical cues and experiments prove enhanced performance, improving accuracy from 79.38% to 90.87% on a Mini-DDSM dataset. Additionally, our approach achieves 97.05% accuracy on an image-only dataset, highlighting the robustness and effectiveness of visual feature extraction. These findings emphasise the potential of multi-modal learning in medical diagnostics, paving the way for future research on optimising data integration strategies and refining AI-driven clinical decision support systems.

Index Terms—Multi-Modal, Cancer Detection, Deep Learning, Breast Cancer

I. INTRODUCTION

According to WHO (World Health Organisation) and ACS (American Cancer Society), BC affects around 1.7 million people a year. Breast cancer (BC) is one of the most common forms of malignancies diagnosed in women worldwide. Although it can affect both men and women, but it is more widely diagnosed in women [1]. BC is caused by clotting or abnormal growth of cells. Early cancer identification is a key step in improving the results of a patient, as early diagnosis and quantifiable measures can increase the success rate of treatment, have considerable influence, and save lives [2]. Although effective, traditional diagnostic methods often have limitations, such as various interpretations of symptoms and data and the need for specific knowledge. Artificial intelligence approaches have become transformative in cancer detection and classification. These methods can analyse enormous amounts of medical images and genomic data and identify subtle patterns that may mislead human experts. For example, mainstream AI models, including convolutional neural network (CNN),

have demonstrated high accuracy in mapping breast cancer on mammograms, comparing expert radiologists [3]

Text-based approaches to breast cancer diagnosis have also attracted increasing attention due to their potential to improve diagnostic accuracy by incorporating additional patient information. Several studies have investigated the integration of clinical and demographic data such as age, genetic and family history data, and medical imaging. For example, a research method [4] used machine learning model to fuse demographic and clinical data to predict breast cancer risk. The promising results show that these models can outperform conventional methods in some cases. Similarly, a study by [5] showed that the integration of textual data, including patient histories and pathology reports, can increase the predictive power of AI models, thus leading to better risk stratification. However, text-based methods have several drawbacks, such as their reliance on raw data, which can lead to biased or inaccurate results. In this study [6], the performance of the k-means algorithm is evaluated by evaluating the classification accuracy. They used K-Means clustering on BCW dataset (Breast Cancer Wisconsin), which estimates various parameters such as initialization (foggy and random), distance measures (Manhattan, Euclidean and Pearson) and data normalization methods (simple and variance). The accuracy of this algorithm is about 92%. This study demonstrates that k-means can effectively distinguish benign from malignant tumours and also highlights the importance of parameter selection for optimal classification. In this paper, only the numerical transformation was evaluated, and the integration of additional data (e.g., imaging, or clinical data) can increase the accuracy and robustness of the model. Furthermore, although NLP techniques are powerful, raw textual data often requires extensive preprocessing to avoid misunderstandings or loss of valuable information in patient records. These questions highlight the importance of integrating disparate data sources to improve prediction models.

Although text data provides valuable insights, artificial intelligence-based image models using large amounts of visual data have proven to mitigate many limitations by leveraging large-scale, unstructured visual data. For example, using CNN, mammography studies have achieved highly accurate results and have provided valuable information about tumour characteristics that may be overlooked in textual data. How-
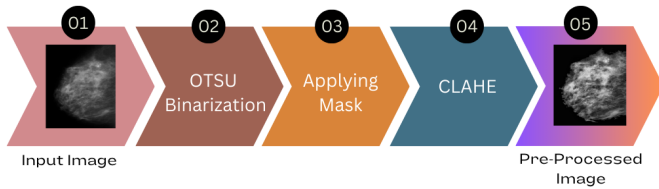
Fig. 1. Pre-processing of Mammography Dataset for Multi-modal deep neural network.

ever, image-based models also have some limitations, such as the need for high-definition images and the difficulty of generalising to different groups of patients. To overcome this problem, some studies combine data from the multi-modalities i.e., image and text to obtain more representative inputs for the patients. For example, DeepClinMed-PGM presents a deep learning-based multi-modal model designed for breast cancer prognosis by integrating pathological data with molecular insights for disease-free survival (DFS) prediction [7]. This study highlights the effectiveness of DeepClinMed-PGM in integrating multi-modal clinical and molecular data, significantly improving breast cancer prognosis and supporting personalised treatment strategies.DeepClinMed-PGM enhances DFS prediction by combining pathology imaging, molecular profiles, and clinical data. The model achieved high AUC values, with the training cohort scoring 0.979, 0.957, and 0.871 for 1-, 3-, and 5-year DFS predictions. In external testing, AUC values were 0.851, 0.878, and 0.938 for 1-, 2-, and 3-year predictions. Strong predictive performance was further validated by hazard ratios (HR: 0.027, 0.117, 0.061) and C-index values (0.925, 0.823, 0.864) across cohorts.

Cancer predictions based on radiology reports are based on structured data such as patient demographics, clinical history, or genetic markers. Models such as logistic regression and decision trees analyse these features to make predictions and perform well on well-structured datasets. However, its generalisation ability is limited compared to models that process complex or large-scale data such as medical images. For example, traditional machine learning methods have had some success in using textual metadata for risk prediction [4]. Although these techniques are effective for certain use cases, they often perform poorly when capturing diverse patterns, highlighting the need for more sophisticated techniques (multi-modal deep learning techniques) to improve results.

Multi-modal artificial intelligence models show great potential to detect breast cancer by combining different types of data, such as medical images and clinical data. For example, DeepMind's RETAIN combine clinical data with structured metadata to predict health outcomes [8]. MedFuse explores a multi-modal approach to predictive modelling in healthcare. It integrates clinical time-series data with chest X-ray images for tasks such as in-hospital mortality prediction and phenotype classification. MedFuse was designed to accommodate both uni-modal and multi-modal inputs, addressing the challenge of asynchronous data collection, where not all modalities

are available for every patient. The authors demonstrate that MedFuse outperforms more complex fusion strategies on fully paired test sets and remains robust even with missing data in partially paired test sets. This underscores its practical applicability, as it has been validated with real clinical data. Additionally, the release of the code ensures reproducibility, enabling further research and development in multi-modal healthcare applications. [9].

A novel breast tumour classification architecture, termed SW-ForkNet, was proposed in [10]. This architecture integrates a DenseNet121 backbone with the Swin Transformer framework. SW-ForkNet incorporates stress and spatial excitation (sSE) blocks to enhance spatial detail extraction, while the Swin Transformer component captures long-range dependencies within the global context. Strategic interconnections within the DenseNet121 architecture further optimise the performance of sSE and Swin Transformer. The network processes three distinct feature sets, ultimately generating a final feature map through vectorization, concatenation, and subsequent processing. A softmax classifier then predicts the tumour classification. Evaluations across three benchmark datasets (BUSI, GDPH, and SYSUCC) shows SW-ForkNet's superior performance, achieving higher accuracy and F1-scores compared to existing methods. Notably, SW-ForkNet achieved 93.12% accuracy and 92.27% F1-score on BUSI, 96.15% accuracy and 96.04% F1-score on GDPH, and 94.88% accuracy and 94.03% F1-score on SYSUCC. These results suggest that SW-ForkNet presents a promising and efficient architecture for breast cancer classification

Another recent method [11] introduced a novel framework, KRC-APM (Key Region Cropping with Artificial Prior Model and Principal Region Planting), for breast cancer detection in ultrasound images. To enhance the accuracy of tumour detection, KRC-APM addresses the limitations of relying on standard regions of interest (ROIs). A critical region cropping (CRC) method is employed to identify and isolate high-confidence tumour areas, effectively increasing the ratio of tumour area to background. This is achieved by analysing the reliability level of potential tumour regions. Moreover, with the guidance of experienced oncologists, the framework integrates three types of diagnostic artificial priors (APs) to capture essential tumour characteristics: shape modelling, margin analysis, and echogenicity pattern identification. The performance of KRC-APM was evaluated on two distinct datasets. Results demonstrated that the proposed method significantly improved breast cancer recognition accuracy. On the UDIT dataset, UNet++ achieved the highest performance among other methods, with an Intersection over Union (IoU) of 78.58%, a Dice Coefficient (DC) score of 76.49%, and an average score of 77.54%. Similarly, on the BUSI dataset, UNet++ exhibited optimal performance with an IoU of 61.74%, a DC score of 60.57%, and an average score of 61.16%. The research [12] proposesed a deep learning model to classify tumours by integrating gene expression data and collected histopathological images. This model uses normalization and dimensionality reduction to process gene expression data from the TCGA-

BRCA dataset and analyses histopathological images to extract features. Both models (gene expression and image processing) were combined in the final layer to predict the outcome. With 10 cross-validations and evaluated for accuracy, sensitivity, specificity, and AUC, the model achieved an overall accuracy of 88.07%, with an AUC of 0.9427 between subtypes and an AUC of 0.9427 between subtypes and multiple methods. Better research than traditional methods.

A. Motivation

Breast cancer classification faces significant challenges, particularly in handling incomplete, noisy data and integrating multi-modal data. Traditional methods primarily rely on visual cues from medical images, often overlooking complementary patients data, such as clinical history and demographics, which can greatly enhance classification accuracy. Moreover, variations in breast tissue density and subtle cancerous lesions make it difficult to achieve reliable results in less-than-ideal conditions. To address these challenges, our approach incorporates deep learning with attention mechanisms to reconstruct noisy images and combines both image and clinical data. This multi-modal strategy enhances classification accuracy and robustness, overcoming the limitations of single-modal methods. The main objectives of this study are:

- Many current approaches primarily focus on image-only data and do not incorporate other important patient information. We explore how deep learning models can improve performance on the Mini-DDSM dataset by using both image and patient data

- A significant challenge in current methods is their inability to handle noisy or incomplete data. To address this, we apply image preprocessing techniques, including Otsu binarization for noise reduction, CLAHE for contrast enhancement, and resizing to standardize input dimensions. These preprocessing steps improve image quality, enhancing feature extraction and ultimately enhancing the performance of existing backbone models and providing more accurate classification results.

- Evaluate the overall impact of combining multi-modal deep learning techniques on the classification accuracy and reliability of various types of medical image datasets, through a series of ablation studies to assess the contribution of individual components like attention mechanisms, unfreezing layers, and the integration of patient metadata.

II. MULTI-MODAL CANCER DETECTION NETWORK

The proposed network is designed to address challenges in existing approaches, particularly the reliance on single-modality models, which miss valuable information. By integrating both image data and textual data (such as age and breast density), this multi-modal network enhances feature representation, providing a more comprehensive understanding of breast abnormalities. This combination of modalities aims to improve classification accuracy and overcome the limitations of traditional image-only models.

A. Data Acquisition and Preprocessing

For this study, we considered the MINI-DDSM [13] and BUSI [14] datasets. The BUSI dataset is often used exclusively for breast cancer classification, while MINI-DDSM is utilized for the implementation of multi-modal learning to enhance classification performance. Each mammographic image in MINI-DDSM is accompanied by patient metadata, including age and breast density. A structured CSV file provides the necessary mapping between image files and their corresponding metadata.

For the multi-modal approach, we preprocess the mammographic images using a custom preprocessing pipeline as shown in Figure 1. Preprocessing includes removing unnecessary details and image enhancement. Otsu binarisation is applied to mammograms to generate a binary mask that can effectively distinguish between the background region and feature region [15] [16]. Following this, the largest white region, which corresponds to the breast tissue, is retained by selecting the largest contour, while eliminating noise and extraneous elements such as labels, skin-adjacent tissue, and black borders [17].

Once the breast region is isolated, Contrast Limited Adaptive Histogram Equalization (CLAHE), as discussed by [18], is employed to enhance local contrast, improving the visibility of micro-calcifications and other critical structures. To ensure consistency in feature extraction, the mammogram is converted to gray-scale before being resized to 224×224 pixels, standardizing the input dimensions for deep learning models. Finally, a batch processing approach is implemented, organizing images within a structured folder-based system to automate preprocessing for large datasets [19]. The same preprocessing method has been adopted by [20] on the same dataset, leading to improved results in their experiments.

The datasets are further processed using the Custom data generator module. Image transformations include RGB conversion, tensor normalization, and ensuring compatibility with deep learning frameworks. The age and density values are extracted and converted into numerical tensors. The datasets are then divided into training and validation subsets to evaluate model performance.

B. MMDCNet Training

In this work, we first evaluate multiple pre-trained deep learning models, including ResNet50, EfficientNet, DenseNet, and Vision Transformer (ViT) [21] [22], to detect breast cancer using image data. These models are initialized with ImageNet weights and adapted for the task by freezing the initial layers. The number of layers frozen depends on the model architecture, with earlier layers typically frozen to retain general feature representations, while the final layers are fine-tuned for our specific task.
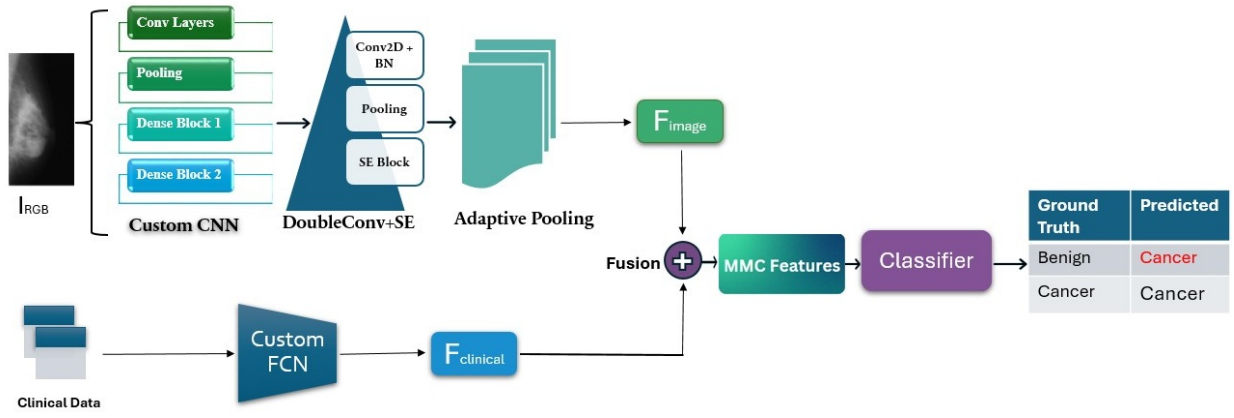
Fig. 2. The proposed MMDCNet Framework. The imaging pathway processes mammogram input (IRGB) through a Custom CNN with convolution layers, pooling, and two dense blocks, followed by a specialized DoubleConv+SE module (incorporating Conv2D with batch normalization, pooling, and Squeeze-Excitation block) and adaptive pooling to generate image features (Fimage). The clinical data pathway processes patient information through a Custom FCN to produce clinical features (Fclinical). These complementary features are merged through a fusion mechanism to create MMC (Multi-Modal Combined) features, which are then classified to predict cancer. The architecture demonstrates its performance through a classification matrix comparing ground truth against predicted outcomes, leveraging both visual and clinical indicators for enhanced diagnostic accuracy.

After evaluating these models, we trained them on the BUSI dataset for image classification and the MINI-DDSM dataset for both image and textual data-based detection. For image-based classification, the best-performing deep learning model is selected based on accuracy, precision, recall, and F1 score. The results for both datasets are shown in Table I and Table II. For Multi-modal classification, the extracted mammogram features are combined with patient metadata (age and density), enabling a more comprehensive approach to classification. The Multi-modal architecture leverages different deep learning models as the image backbone while fusing textual data using self attention attention mechanisms.

1) Multi-modal Learning: Our proposed MultiScaleClassifier, implemented in PyTorch, integrates both image and textual features for improved breast cancer diagnosis. The proposed model is explained in subsequent sections.

a) Feature Extraction and Fusion Strategy: Our approach integrates deep learning-based image feature extraction with structured patient metadata to enhance breast cancer classification accuracy, as shown in the architecture diagram Figure 2. We utilize DenseNet [22] as the backbone model, leveraging its dense connectivity to enable efficient feature propagation and gradient flow. This architecture is particularly suited for medical imaging, as it captures fine-grained textures essential for identifying subtle abnormalities in breast tissue. To further refine feature representation, a Squeeze-and-Excitation (SE) attention block is incorporated, adaptively recalibrating channel-wise feature responses to emphasize diagnostically significant patterns while suppressing less relevant information.

Following feature extraction, additional convolutional layers with batch normalization are employed to refine spatial patterns and ensure training stability. Batch normalization mitigates internal covariate shifts, improving model generalization across diverse breast tissue characteristics. This ensures

robustness against variations in imaging conditions and patient demographics, thereby enhancing the model's capacity to distinguish between benign, cancerous, and normal tissues. Concurrently, structured patient metadata, including age and breast density, undergoes normalization and processing through a fully connected neural network. This module extracts meaningful patterns from the metadata while filtering out noise, ensuring that only relevant contextual information contributes to the final prediction.

The extracted textual features are projected into a 128-dimensional space, optimizing the balance between computational efficiency and discriminative power. This representation enables effective fusion with image-derived features, facilitating a unified and comprehensive diagnostic model. To further enhance feature interaction, a self attention attention mechanism with eight heads is applied, allowing the model to capture intricate relationships between visual and textual data. This attention mechanism improves cross-modal learning by enabling the model to focus on the most informative aspects of both modalities, ensuring a richer representation of patient-specific diagnostic factors.

The fused representation is processed through fully connected layers, culminating in classification into three categories: Benign, Cancer, or Normal. By integrating structured metadata with deep visual features, this approach overcomes the limitations of single-modality analysis, leading to a more comprehensive diagnostic framework. The inclusion of attention mechanisms and normalization techniques ensures that the model remains transparent, robust, and adaptable across diverse datasets, ultimately contributing to improved breast cancer Classification accuracy.

b) Self-Attention Mechanism: To capture spatial dependencies within an image, a self-attention mechanism is used, which computes attention over the feature map by projecting it into query, key, and value spaces [23]. The self-attention

mechanism computes attention for each spatial position as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

where: - $Q = \text{Conv1x1}(x)$ is the query matrix, - $K = \text{Conv1x1}(x)$ is the key matrix, - $V = \text{Conv1x1}(x)$ is the value matrix.

The query, key, and value matrices are derived by applying 1x1 convolutions to the input feature map. The softmax function is applied to the dot product of the query and key, which gives the attention scores. These attention scores are then used to weight the values and obtain the output feature map. The final output is obtained by applying a residual connection:

$$\text{Output} = \gamma \times \text{Attention}(Q, K, V) + x$$

where γ is a learnable scaling parameter, and x is the input feature map. This mechanism allows the model to focus on different regions of the input feature map based on their relevance, enabling the capture of long-range dependencies [24].

III. EXPERIMENTAL RESULTS

1) BUSI Dataset: This study makes use of a dataset of breast ultrasound images for the analysis of breast cancer. It is very useful for the tumour classidication and detection in ultrasound images affected by noise and artifacts. The dataset includes three categories: normal, benign, and malignant. When combined with machine learning techniques, breast ultrasound images have shown considerable potential for accurate classification, detection, and segmentation of breast cancer. The initial data collection involved breast ultrasound images from 600 women aged between 25 and 75 years, gathered in 2018. The dataset consists of 780 images, each with an average resolution of 500x500 pixels, and stored in PNG format. Ground truth annotations are provided alongside the original images. The images are classified into three categories: normal, benign, and malignant [14].

Our evaluation on the BUSI dataset highlights that BCDNet α achieves the highest accuracy of 97.05%, among the models tested. Specifically, our model with BCDNet α outperforms previous work such as DenseNet-121, which achieved 88.46% accuracy, EfficientNetB0 and ResNet101, which reported accuracies of 86.14% and 91.22%, respectively. The addition of an attention mechanism generally leads to improvements in model performance.The SelfAttention mechanism is applied to the features extracted by the DenseNet-121 backbone. This attention mechanism allows the model to focus on important regions within the image, enhancing the model's ability to prioritize critical features that are relevant for identifying cancerous or benign tissues. Specifically, the attention mechanism recalibrates feature maps and helps in learning more robust representations, thus improving the overall detection accuracy. This approach is particularly beneficial when dealing with

Method	Accuracy	Precision	Recall	F1 Score
DenseNet-121 [25]	88.46	77.50	83.78	80.52
DenseNet-40 [25]	90.77	80.00	88.89	84.21
VGG-Like [25]	85.38	75.00	76.92	75.95
DenseNet-161 [25]	94.62	90.00	92.31	91.14
SW-ForkNet [10]	93.12	92.27	92.27	92.27
Swin Transformer [10]	52.45	50.13	51.46	50.79
ConvMixer [10]	83.42	81.72	80.65	81.18
DenseNet121 [10]	88.02	86.84	86.80	86.82
ResNet101 [10]	91.22	90.25	90.62	90.43
EfficientNetB0 [10]	86.14	83.50	84.51	84.00
Xception [10]	90.19	89.31	89.29	89.30
VGG16 [10]	87.23	86.65	85.05	85.84
BCDNetα	97.05	96.91	96.49	96.66
BCDNetβ	96.67	95.87	97.26	96.53
BCDNetγ	92.56	94.44	89.93	91.79

TABLE I
PERFORMANCE METRICS FOR BUSI DATASET DATASET. RED ARE THE BEST RESULTS.
α: DENSENET-121 BACKBONE, β: EFFICIENTNETB0 BACKBONE, γ: VGG16 BACKBONE

complex and noisy medical images, as it boosts the model's ability to capture subtle patterns that are crucial for accurate classification.

2) Mini-DDSM Dataset: This study investigates a dataset of breast ultrasound images for the analysis of breast cancer. The Mini-DDSM dataset provides features recorded in mammograms so that generative models can be evaluated in the context of traditional cancer detection methods. It provides a comprehensive basis for performing generative modeling in a variety of imaging modalities. The dataset is classified into benign, malignant, and normal. In addition to the image data, the dataset includes patient-specific attributes such as age and breast density. The dataset provides a folder for each patient containing the original ultrasound image, a filename identifier, and a binary mask delineating the suspicious or tumour contour. This comprehensive dataset provides valuable information for developing and evaluating breast cancer detection and diagnosis algorithms. The binary mask of the lesion is constructed based on the original Freeman chain coding, so this data set protects you from that inconvenience [13].

Our results on the Mini-DDSM dataset show that MMCDNet$_\alpha$, with the DenseNet201 backbone, outperforms other traditional CNNs in accuracy. The MMCDNet$_\alpha$ achieves a high accuracy of 90.87%, while models such as DenseNet201 from previous works report an accuracy of 86.51%. Additionally, models like Xception, ResNet50, and MobileNet demonstrate lower accuracies, ranging from 78.30% to 82.99%. Although the increase in performance between the our suggest MMCDNet and [20] is not dramatically higher, the critical point lies in the difference in approach. The BRINT-based model [20] combines handcrafted texture features from BRINT with pre-trained deep learning architectures like VGG16, ResNet50, and MobileNetV2. This hybrid approach leverages the strengths of both texture descriptors

Method	Accuracy	Precision	Recall	F1 Score
DenseNet201 [26]	86.51	86.53	86.54	86.54
Xception [26]	82.99	83.08	83.04	83.06
ResNet50 [26]	82.70	82.70	82.71	82.70
MobileNet [26]	81.23	81.23	82.24	81.23
Inception [26]	78.30	78.53	78.39	78.46
BRINT + VGG16 [20]	90.26	90.53	89.98	90.25
BRINT + MobilNetV2 [20]	78.03	78.78	77.34	78.05
BRINT + Custom CNN [20]	71.22	72.71	69.11	70.87
BRINT + ResNet50 [20]	58.85	93.93	26.26	41.05
BRINT + VGG19 [20]	39.00	98.50	5.38	10.20
DenseNet201	79.38	79.35	79.98	79.25
EfficientNetB0	70.00	70.32	70.31	70.31
VGG16	55.00	55.99	55.98	55.98
VGG16 + Atten	62.00	63.03	63.02	63.02
MMCDNet$_\alpha$	90.87	91.55	91.61	91.58
MMCDNet$_\beta$	84.73	85.33	85.50	85.37
MMCDNet$_\gamma$	78.46	79.68	79.28	79.44

TABLE II

PERFORMANCE METRICS FOR MINI-DDSM DATASET. RED COLOUR INDICATES THE BEST RESULTS.
α: DENSENET201 BACKBONE, β: EFFICIENTNETB0 BACKBONE, γ: VGG16 BACKBONE

and robust feature extraction from pre-trained networks, enhancing the model's ability to identify subtle differences in tissue patterns crucial for cancer classification. The use of well-established pre-trained models ensures scalability and efficiency, while combining low-level texture features with high-level semantic features improves diagnostic accuracy. Overall, it offers a versatile and powerful solution for breast cancer detection. On other hand, by leveraging a Multi-modal approach, MMCDNet$_\alpha$ demonstrates how integrating both image and metadata with a robust backbone like DenseNet201 leads to better feature extraction and classification, outperforming previous methods. The improvements in MMCDNet$_\alpha$ are likely due to the ability of the DenseNet201 backbone to capture complex spatial relationships in the image data, while the multi-modal approach introduces complementary information from metadata, leading to a more comprehensive representation of the input data. Additionally, incorporating attention mechanisms like SelfAttention and SEBlock within the DenseNet201 backbone allows the model to focus on critical regions of the mammogram, which helps distinguish between benign, normal, and malignant cases. This integration of attention mechanisms enhances feature selection and model interpretability, making the network more robust to variations and improving the classification performance.

A. Evaluation Metrics

Model performance is assessed using accuracy, precision, recall, and F1-score. Accuracy measures overall correctness as the ratio of correct predictions to total samples. Precision evaluates the proportion of correctly predicted positive cases, minimizing false alarms, while recall quantifies the model's ability to detect actual positives.

ABLATION STUDY

We carried out several ablation studies to analyze the effect of various parameters in the proposed MMDCNet framework. M1 uses a custom CNN with convolutional layers, pooling, and Dense Blocks 1 and 2 but does not incorporate attention mechanisms or unfreezing layers. M2, M3, and M4 extend M1 by adding attention mechanisms and unfreezing deeper layers, using VGG16, DenseNet201, and EfficientNetB0, respectively. M5 evaluates single-modality performance with only image features from the Mini-DDSM dataset. Finally, MMDCNet integrates both image features and patient metadata (age and density) via a self-attention fusion mechanism, using Pretrained backbone with unfreezing layers and a Squeeze-and-Excitation (SE) attention block. This multi-modal approach shows the best performance across all experiments. Next, we provide our analysis of these results.
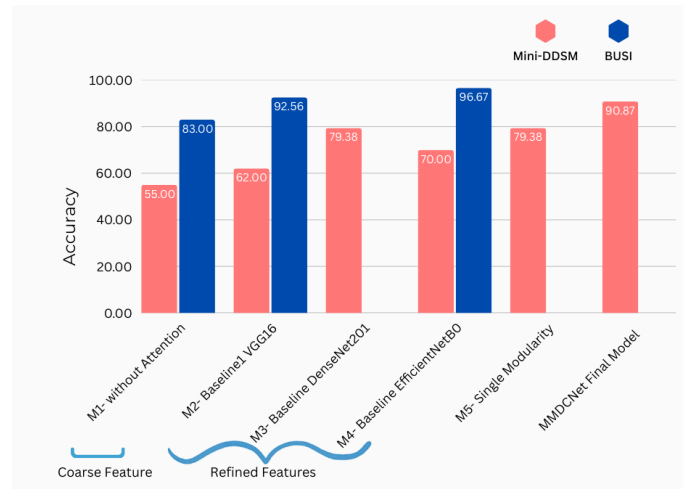

Fig. 3. Ablation studies.

Effect of Attention Mechanism and Unfreezing Layers

Model M1 processes mammogram images through a custom CNN architecture that incorporates convolutional layers, pooling layers, and Dense Blocks 1 and 2. It uses a pretrained backbone but does not include attention mechanisms or unfreeze any layers. This results in limited domain-specific adaptation and feature recalibration, which hinders its performance. In contrast, models M2, M3, and M4 integrate self-attention mechanisms and unfreeze layers for improved performance. These models leverage the same custom CNN architecture but apply attention mechanisms and unfreeze deeper layers to better adapt to the medical domain. Specifically, M2 uses VGG16 with self-attention and unfreezing of deeper layers, enhancing feature extraction for medical images. M3 uses DenseNet201 with self-attention and unfreezing of deeper layers, taking advantage of DenseNet's architecture for better gradient flow and feature reuse. M4 uses EfficientNetB0 with self-attention and unfreezing of deeper layers, offering a balance between efficiency and accuracy, which improves performance without overloading the model. These models

show significant performance improvements over M1 due to the enhanced feature representation and domain-specific adaptation provided by the attention mechanisms and unfreezing layers.

Effect of Multi-Modal Learning

In the M5 experiment, we evaluate the classification performance on the Mini-DDSM dataset using only image-based features. This model, like M1, uses the custom CNN architecture with convolutional layers, pooling, and Dense Blocks 1 and 2, but it does not integrate patient metadata, serving as a baseline for single-modality performance. The MMDCNet (Final) model takes this a step further by incorporating both mammogram image features and structured patient metadata (age and density) through a self-attention-based fusion mechanism. This model integrates a DenseNet201 backbone with unfreezing layers and a Squeeze-and-Excitation (SE) attention block to enhance visual feature extraction. As shown in Figure 3, the multi-modal approach (MMDCNet) outperforms the single-modality model, demonstrating the effectiveness of combining image and textual features for better classification accuracy. The self-attention fusion mechanism ensures effective interaction between image and textual data, allowing the model to leverage both modalities to improve classification accuracy. This multi-modal approach results in the best performance across all experiments, as it combines the rich information from both visual and patient-specific features, outperforming the image-only models.

IV. CONCLUSION

Breast cancer Classification has traditionally relied on image-based approaches, often overlooking valuable patient-specific textual data. In this study, we introduced a Multi-modal Cancer Detection Network (MMDCNet) that integrates visual and clinical data to improve diagnostic accuracy. By fine-tuning DenseNet121 with an attention mechanism on the BUSI dataset, we achieved an impressive accuracy of 97.05%. On the mini-DDSM dataset, DenseNet201 initially reached 79.38%, but the incorporation of both image and clinical data through our multi-modal approach boosted the accuracy to 90.87%. This demonstrates the effectiveness of combining visual and clinical features in addressing the complexities of breast cancer detection. The mini-DDSM dataset, being more complex, presents additional challenges, and further improvements in accuracy can be achieved by refining the model or updating the dataset with additional high-quality data. These findings underscore the importance of leveraging both visual and clinical data to provide a more comprehensive understanding of breast cancer characteristics, leading to more accurate and reliable detection.

In future, we aim to refine the model and explore ways to enhance the integration of diverse data types for even better performance. Furthermore, we plan to extend our approach to include more diverse datasets and improve its scalability for real-world clinical applications. Our results highlight the potential of multi-modal learning in advancing breast cancer

classification detection and paves the way for future AI-driven clinical decision support systems.

REFERENCES

[1] A A, P M, S Bourouis, S S Band, A Mosavi, S Agrawal, and M Hamdi. Meta-heuristic algorithm-tuned neural network for breast cancer diagnosis using ultrasound images. In *Frontiers in Oncology, Cancer Imaging and Image-directed Interventions*. frontiersin, 2022.

[2] Noor Ul Huda Shah et al. Breast cancer identification using improved darknet53 model. In *International Conference on Innovations in Bio-Inspired Computing and Applications*. Springer Nature Switzerland, 2022.

[3] S. M. McKinney, M. Sieniek, and V. et al. Godbole. International evaluation of an ai system for breast cancer screening. In *Nature 577*, page 89–94. Nature, 2020.

[4] R. Rabiei, S. M. Ayyoubzadeh, S. Sohrabei, M. Esmaeili, and A. Atashi. Prediction of breast cancer using machine learning approaches. *Journal of biomedical physics & engineering*, 12(3):297, 2022.

[5] M. D. Ganggayah, N. A. Taib, Y. C. Har, et al. Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC Med Inform Decis Mak*, 19(48), 2019.

[6] A.K. Dubey, U. Gupta, and S. Jain. Analysis of k-means clustering approach on the breast cancer wisconsin dataset. *International Journal of Computer Assisted Radiology and Surgery*, 11(2033-2047), 2016.

[7] Z Wang, R Lin, Y Li, J Zeng, Y Chen, W Ouyang, H Li, X Jia, Z Lai, Y Yu, H Yao, and W Su. Deep learning-based multi-modal data integration enhancing breast cancer disease-free survival prediction. *Precis Clin Med*, 7(2):pbae012, 2024.

[8] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29, 2016.

[9] Nasir Hayat et al. Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images. In *Machine Learning in Health Care*, pages 479–503, July 2022.

[10] H. Üzen, H. Firat, O. Atila, and A. Şengü. Swin transformer-based fork architecture for automated breast tumor classification. *Expert Systems with Applications*, 256:125009, 2024.

[11] Y. Lin, H. Wang, and J. Jiang. Krc-apm: Key region cutting and artificial prior model for breast cancer recognition in ultrasound images. *Expert Systems with Applications*, 257:125092, 2024.

[12] T. Liu, J. Huang, T. Liao, R. Pu, S. Liu, and Y. Peng. A hybrid deep learning model for predicting molecular subtypes of human breast cancer using multimodal data. *IRBM*, 41(5):293–298, 2020.

[13] C. D. Lekamlage, F. Afzal, E. Westerberg, and A. Cheddad. Mini-ddsm: Mammography-based automatic age estimation. In *Proceedings of the 3rd International Conference on Digital Medicine and Image Processing (DMIP 2020)*, pages 1–6, Kyoto, Japan, November 06-09 2020. ACM.

[14] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, Feb 2020.

[15] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[16] A. M. Omer and M. O. Elfadil. Preprocessing of digital mammogram image based on otsu's threshold. *American Scientific Research Journal for Engineering Technology and Sciences*, 37:220–229, 2017.

[17] A. Elmoufidi, K. El Fahssi, S. Jai-Andaloussi, and A. Sekkaki. Automatically density based breast segmentation for mammograms by using dynamic k-means algorithm and seed based region growing. In *2015 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings*, pages 533–538, 2015.

[18] E. M. El Houby and N. I. Yassin. Malignant and nonmalignant classification of breast lesions in mammograms using convolutional neural networks. *vol. 70*, page 102954, 2021.

[19] J. Joseph and R. Periyasamy. Biomedical signal processing and control. *Biomedical Signal Processing and Control*, 39:271–283, 2018.

[20] L. A. S. Marroquin and J. C. G. Caceres. Diagnosis of breast cancer on full mammography using the brint texture descriptor and convolutional neural networks. *Latin American Computer Conference (CLEI)*, 2024.

[21] Pablo Ramirez Amador, Dinarle Milagro Ortega, and Arnold Cesarano. Detection of pulmonary pathologies using convolutional neural networks, data augmentation, resnet50 and vision transformers. *arXiv e-prints*, pages arXiv–2409, 2024.

[22] S. Anari, S. Sadeghi, G. Sheikhi, and et al. Explainable attention based breast tumor segmentation using a combination of unet, resnet, densenet, and efficientnet models. *Scientific Reports*, 15:1027, 2025.

[23] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10076–10085, 2020.

[24] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7794–7803, 2018.

[25] W. K. Moon, Y. W. Lee, H.-H. Ke, S. H. Lee, C.-S. Huang, and R.-F. Chang. Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. *Computer Methods and Programs in Biomedicine*, 190:105361, 2020.

[26] A. Kacher, M. Merati, and S. Mahmoudi. Classification of multi-view mammogram images using a parallel pre-trained models system. *8th International Conference on Image and Signal Processing and their Applications (ISPA)*, 2024.