American Journal of Artificial Intelligence
2026, Vol. 10, No. 1, pp. 1–13
https://doi.org/10.11648/j.ajai.20261001.11

Research Article

Hybrid CNN-Transformer Model for Early Detection and Diagnosis of Breast Cancer: A Multi-Regional Dataset Study with Implications for African Healthcare Settings

Kasim Chambulilo[1, *], Paul Wako[2], Davis Bassa[3], Asha Kisonga[1], Emmanuel Shija[4]

[1]Department of Data Science and ICT, Eastern Africa Statistical Training Centre, Dar es Salaam, Tanzania

[2]Department of ICT and Engineering, Zetech University, Nairobi, Kenya

[3]Department of Product Design, Research and Development, Bassakimu Innovations Company Limited, Kampala, Uganda

[4]Department of Radiology, Catholic University of Health and Allied Sciences, Mwanza, Tanzania

Abstract

Breast cancer remains the most common cancer among African women, contributing to high mortality largely due to late-stage diagnosis. More than 70% of cases in many African countries are detected at advanced stages, when treatment options are limited and survival rates are significantly reduced. Barriers such as limited access to screening, a shortage of radiologists, and socio-economic constraints further delay early detection. Artificial intelligence (AI)-based diagnostic tools offer an opportunity to strengthen screening systems and improve timely diagnosis; however, the lack of publicly available African mammography datasets remains a major challenge. This study introduces MAMMOAI, an AI-driven framework designed to enhance early breast cancer detection with potential applicability to African contexts. Due to the scarcity of large-scale annotated African mammography datasets a critical barrier identified in current literature we adopted a pragmatic approach combining a multi-task deep learning model integrating convolutional neural networks (CNN) with Transformer layers to perform simultaneous risk assessment, cancer detection, staging, risk factor analysis, and differential diagnosis using mammograms. The model was trained using the King AbdulAziz University Breast Cancer Mammogram Dataset (KAU-BCMD) from Saudi Arabia, supplemented with preliminary data collected from Kenyan healthcare facilities. This multi-regional approach was necessitated by insufficient African data volume for robust deep learning model training. The framework employs a multi-task learning approach that integrates a ResNet50–Transformer backbone for spatial feature extraction, feeding five specialized branches for risk assessment, cancer detection, staging, risk factor analysis, and differential diagnosis. All images underwent standardized preprocessing, including resizing to 224×224 pixels, normalization, contrast enhancement, and extensive data augmentation. Weighted categorical cross-entropy losses supported joint optimization across tasks. Model interpretability was ensured using Grad-CAM–based heatmaps and uncertainty estimation, and predictions were compiled into automated, clinician-friendly HTML reports. The model achieved overall accuracy of 98% on the majority class (BI-RADS 1), with macro-averaged F1-scores of 0.61-0.62 across all branches, reflecting challenges in detecting minority classes. Critically, the model failed to identify any BI-RADS 5 (highly malignant) cases, misclassifying all as BI-RADS 4. Grad-CAM visualizations provided interpretable insights, supporting clinical decision-making. While these results demonstrate technical feasibility and the potential of hybrid architectures, they also underscore critical limitations: severe class imbalance, inadequate minority class performance, and unproven generalizability to diverse African populations. Future work must prioritize collection of larger, balanced, multi-center

*Corresponding author: kassimchambulilo@gmail.com (Kasim Chambulilo)

Received: 9 November 2025; **Accepted:** 3 December 2025; **Published:** 29 December 2025

African datasets and external validation across diverse Sub-Saharan African healthcare settings before clinical deployment can be considered.

Keywords

Breast Cancer, Deep Learning, Mammography, MammoAI

1. Introduction

Breast cancer remains a major public health challenge across sub-Saharan Africa, with survival rates significantly lower than in high-income regions. A recent seven-year follow-up of the ABC-DO cohort across five sub-Saharan African countries reported that only one in three women survive seven years after diagnosis, underscoring the urgent need for earlier detection and improved care pathways [1]. According to the World Health Organization (WHO), persistent gaps in cancer control include limited screening programs, inadequate diagnostic and pathological services, and low availability of specialized treatment centers across the African Region [2]. Socioeconomic and logistical barriers further hinder access to screening, with clinical breast examination rates as low as 19.2% in several countries [3]. These challenges contribute to late-stage presentation and high mortality, emphasizing the need for scalable, affordable, and reliable diagnostic innovations suited to African healthcare systems.

Recent advances in artificial intelligence (AI) have shown considerable promise in addressing diagnostic delays in breast cancer care. Hybrid deep learning architectures combining Convolutional Neural Networks (CNNs) with Transformer modules have demonstrated improved accuracy and interpretability by capturing both local spatial features and global contextual information from mammograms [4, 5]. However, a critical limitation persists: most AI-based breast cancer detection models, including the present study, rely on mammography datasets from outside Africa. Due to the scarcity of large, publicly available African mammography datasets, MAMMOAI is trained and evaluated using a BI-RADS–labeled Middle Eastern mammography dataset. This geographic mismatch can affect generalizability, as breast density patterns, imaging quality, cancer subtypes, and screening practices may differ across regions.

A fundamental challenge in developing AI diagnostic tools for African healthcare contexts is the critical scarcity of large-scale, annotated mammography datasets from African populations [7]. This data gap has been repeatedly identified as a primary barrier to equitable AI deployment in medical imaging [24]. In this study, we confronted this challenge directly: preliminary data collection from partnering healthcare facilities in Kenya yielded insufficient sample sizes for training robust deep learning models, particularly for rare malignant cases. To address this limitation while advancing the methodological framework, we adopted a pragmatic multi-regional training approach using the King Abdulaziz University Breast Cancer Mammogram Dataset (KAU-BCMD) from Saudi Arabia [27, 28], supplemented with available Kenyan data. While this dataset originates from the Middle East rather than Sub-Saharan Africa, it provided the necessary sample size for model development. We acknowledge that this geographic mismatch represents a significant limitation affecting the model's immediate generalizability to diverse African populations, as discussed extensively in our limitations section.

To bridge this gap, this study introduces MAMMOAI, a hybrid CNN–Transformer framework for early breast cancer detection and diagnosis using mammograms with the long-term goal of supporting African healthcare settings. The model performs multiple diagnostic tasks: risk assessment, detection, staging, and differential diagnosis while maintaining interpretability through Grad-CAM visualizations and uncertainty quantification. Trained on a multi-regional dataset (primarily KAU-BCMD with Kenyan supplementation), the model achieved 98% weighted accuracy but exhibited critical limitations including complete failure to detect BI-RADS 5 malignant cases (0% recall) and macro F1-scores of only 0.61-0.62 due to severe class imbalance. These findings demonstrate both the technical potential of hybrid architectures and the urgent need for larger, balanced, Africa-specific datasets. MAMMOAI represents a foundational methodological framework requiring substantial further development, external validation on diverse African populations, and class imbalance mitigation before clinical applicability can be established.

2. Review of Literature

Breast cancer continues to be the leading malignancy in women in many African countries, with a high proportion of cases diagnosed at late stages due to limited screening access, shortage of trained radiologists, and socio-economic barriers [6, 7]. Computer-assisted and machine-learning methods applied to mammography have therefore received growing attention as possible tools to improve early detection, prioritize patients for follow-up, and reduce diagnostic delays [9, 10https://pubs.rsna.org/doi/full/10.1148/radiol.2019182716?utm_source=chatgpt.com].

2.1. Overview: Machine Learning in Mammography

Since the mid-2010s, deep convolutional neural networks (CNNs) have dominated automated mammogram analysis, showing strong performance on lesion detection and BI-RADS classification when trained on large labeled datasets. More recently, hybrid architectures combining CNN backbones with Transformer (self-attention) modules have been proposed to model long-range dependencies across the entire mammogram and to fuse multi-view information (CC and MLO views) [8, 11]. Multi-task learning (MTL) frameworks that jointly predict multiple clinically relevant outputs (for example detection + BI-RADS + staging or risk) have also gained traction because they can exploit shared representations and improve generalization, especially in data-scarce contexts [12, 13].

Overall, deep learning models that combine convolutional neural networks (CNNs) and Transformers have demonstrated high diagnostic accuracy when trained on well-curated datasets, as they are capable of capturing both local and global image patterns effectively [8, 9]. However, a major limitation arises from domain shift, where models developed using data from high-income countries often perform poorly on African populations due to variations in breast density, imaging devices, and disease prevalence, with external validation frequently lacking [7]. Furthermore, interpretability and clinician trust continue to pose significant challenges; although visualization tools such as Grad-CAM and attention maps are widely used to explain model decisions, they still provide only limited and sometimes unreliable insights into the model's reasoning process [14, 15].

2.2. Risk Assessment (Population/Individual Risk Prediction from Mammograms & Integrated Risk)

Work integrating imaging-based features with traditional risk factors (age, family history, prior biopsies) has shown that mammogram-derived features (for example density, parenchymal patterns, calcifications) can improve individualized risk prediction beyond clinical models alone. Notable work demonstrated that deep learning models trained on full-field mammograms plus risk factors can outperform traditional risk calculators for short-term risk prediction [9]. Recent reviews and studies also explore direct deep learning risk scores from imaging that correlate with future cancer risk and with BI-RADS assessments [15, 16].

One of the key strengths of imaging-based approaches is that they provide measurable and objective features such as breast density and calcifications that significantly enhance risk stratification and early detection accuracy. Moreover, joint modeling frameworks integrating imaging with clinical and demographic data have shown potential to personalize screening intervals and optimize resource allocation in diverse populations [9]. However, notable weaknesses and gaps persist in current research. Most existing risk prediction models are developed and validated using non-African cohorts, raising concerns about their calibration and transferability to African populations. Additionally, there is a lack of prospective validation studies and longitudinal datasets from African contexts, which limits the generalizability and clinical utility of these models in real-world African healthcare settings [7].

2.3. Cancer Detection (Localization/Classification)

Classic detection pipelines include CNN-based detectors (Faster R-CNN variants, U-Net for segmentation) and, more recently, Transformer-augmented models (Vision Transformer variants, hybrid CNN–Transformer) to model long-range contextual cues across mammograms. Multi-view fusion (combining CC and MLO views) and patch-based methods are common for handling high-resolution mammograms [8, 11]. State-of-the-art models report high sensitivity and specificity on large public datasets; however, reported metrics vary depending on dataset composition, prevalence, and evaluation protocol [8, 9].

Despite their strong performance, models that incorporate Transformers face significant challenges related to high computational and memory demands, which make their deployment in low-resource settings particularly difficult. Additionally, many studies downscale mammogram images (for example to 224×224 pixels) for processing efficiency, a step that may eliminate small yet clinically critical features. While patch-based approaches have been proposed to mitigate this issue by focusing on localized regions of interest, they can complicate global contextual reasoning and increase model complexity [17, 18]. Furthermore, the lack of diverse training datasets and absence of external validation using African mammogram data continue to represent major limitations, undermining model generalizability and equitable clinical applicability across populations [7].

2.4. Staging and Severity Assessment

Automated staging from mammograms (assigning BI-RADS categories or estimating clinical stage) is an active research area. DL methods can classify BI-RADS labels and infer tumor extent from imaging features; multi-task architectures jointly learning classification and segmentation are particularly promising because segmentation assists staging decisions [13]. Some recent Transformer-based systems attempt to capture global context that helps in staging tasks that depend on lesion size and multiplicity [19].

Among the notable strengths of recent approaches is the integration of joint segmentation and classification tasks within deep learning frameworks, which enhances the model's capability to estimate lesion size and multiplicity key factors in accurate cancer staging. This multi-task learning

(MTL) setup has also been shown to improve model robustness by allowing shared feature representations that generalize better across related diagnostic tasks [13]. However, several weaknesses remain. Effective clinical staging typically requires multimodal data including ultrasound, MRI, clinical examination, and pathology reports whereas mammography alone has inherent limitations in fully characterizing disease extent. As a result, imaging-only staging models are prone to misclassification when multimodal inputs are not incorporated [12]. Additionally, reliance on BI-RADS labeling introduces a degree of subjectivity, as inter-reader variability can reduce label consistency in supervised training datasets. This label noise may degrade model performance unless mitigated through strategies such as label smoothing or the use of consensus annotations across multiple radiologists [20].

2.5. Risk-factor Analysis (Linking Imaging Phenotypes with Epidemiologic/Clinical Factors)

Studies integrate imaging features with demographic and clinical risk factors to explore associations (e.g., density with age, tumor subtype correlations). Imaging-derived biomarkers (density, texture patterns, calcification features) have been used to study population risk and to predict molecular subtypes in some recent DL studies that attempt to predict phenotype or histologic markers from images [21, 22].

A major strength of radiomics-based approaches is their ability to provide non-invasive phenotype markers that can be correlated with molecular subtypes or specific risk exposures, thereby enhancing personalized diagnosis and treatment planning. These imaging-derived biomarkers are particularly valuable in low-resource settings, where molecular testing capabilities are limited, making them potentially useful for triage and early risk assessment [21]. However, notable weaknesses persist in this domain. Many existing studies report correlational findings that lack pathologic confirmation, limiting their clinical interpretability and reliability. Moreover, multi-site heterogeneity stemming from variations in imaging devices, acquisition settings, and protocols poses challenges for standardizing and pooling radiomic features across studies. Importantly, there remains a scarcity of Africa-specific evidence exploring how imaging phenotypes relate to local genetic, environmental, or lifestyle risk exposures, highlighting an urgent need for context-specific validation [7, 22].

2.6. Differential Diagnosis (Benign vs Malignant, and Mimickers)

Differential diagnosis tasks classify lesions as benign vs malignant, or discriminate among lesion types. Ensemble models, attention mechanisms, and explainability techniques (Grad-CAM, attention maps) are widely used to make predictions more interpretable for clinicians [14, 23]. Multi-task setups that simultaneously predict BI-RADS and lesion type can help reduce false positives and unwarranted biopsies.

One of the key strengths of recent machine learning approaches in mammogram analysis lies in the incorporation of explainable visualization tools such as Grad-CAM and attention maps, which help radiologists interpret model outputs and thereby enhance trust and clinical adoption. Furthermore, multi-task learning (MTL) architectures where related tasks such as detection, staging, and risk prediction are learned jointly can reduce overfitting to a single label and improve overall model generalization [12, 14]. However, these techniques also exhibit notable weaknesses. Saliency-based visualizations like Grad-CAM are often coarse and occasionally misleading, as they tend to highlight image regions that merely correlate with but do not causally explain the model's decisions. Recent research emphasizes the need for quantitative and standardized evaluation of explanation methods to ensure their clinical reliability [15]. Additionally, a large proportion of algorithms are benchmarked on highly curated datasets that contain clearer and more distinct lesions than those typically encountered in real-world screening, resulting in overly optimistic performance estimates that may not generalize to routine clinical practice.

2.7. Interpretability, Uncertainty, and Clinical Integration

Interpretability is a recurring theme. Grad-CAM and attention visualizations are the most used tools, often paired with uncertainty quantification (for example Monte Carlo dropout, ensembles) to flag low-confidence predictions for human review [14, 18]. Clinical integration studies emphasize workflow fit, radiologist-in-the-loop designs, and computational resource constraints for deployment in low-resource settings [10, 18].

Several key gaps and implementation challenges hinder the widespread application of advanced machine learning models for breast cancer detection in African contexts. A major limitation is the scarcity of large, annotated African mammography datasets, which are essential for robust model training, cross-validation, and external benchmarking. Although a few emerging African imaging datasets have recently been proposed, the overall data volume and diversity remain insufficient to ensure generalizable and equitable model performance across the continent [7, 24]. Additionally, computational and infrastructural constraints including limited GPU availability, inadequate Picture Archiving and Communication System (PACS) integration, and unreliable internet connectivity create significant barriers to model deployment and real-time clinical use. To address these constraints, researchers are exploring federated learning frameworks and lightweight architectures as potential strategies to enable distributed training and low-resource deployment without compromising data privacy or model accuracy [18, 25].

2.8. Summary of Strengths and Limitations Across the Field

Deep CNNs and hybrid CNN–Transformer models provide strong detection and classification performance on large, well-labelled datasets; multi-task learning improves robustness by sharing representations across related tasks; interpretability tools (Grad-CAM, attention) help clinician understanding. [8, 9]

However, dataset bias and limited African-specific data, label noise and inter-reader variability, loss of fine detail from aggressive down sampling, high computational requirements, and the incomplete clinical context when relying on mammography alone. Interpretability methods require rigorous validation. [7, 17, 15]

3. Methodology

This methodology outlines a systematic approach for the development of a breast cancer images analysis system, an AI-driven framework for early breast cancer detection using mammogram images. The methodology is structured into key stages: dataset acquisition and preparation, data preprocessing and augmentation, model architecture design, training protocol and optimization, and evaluation methodology and interpretability.

Each stage is designed to ensure scientific rigor, reproducibility, and clinical relevance in building a robust, interpretable, and high-performing deep learning model for breast cancer assessment.

3.1. Dataset Acquisition and Preparation: A Pragmatic Multi-Regional Approach

Data Scarcity Challenge and Ethical Collection Efforts

Initial data collection efforts focused on obtaining mammography images from partnering healthcare facilities in Kenya, specifically Nakuru and Mathare Hospitals. However, after six months of prospective collection, we obtained only 100 mammogram images with BI-RADS annotations, including 50 BI-RADS 1, 45 BI-RADS 4, and critically, only 5 BI-RADS 5 (highly malignant) cases.

This sample size was insufficient for training deep convolutional neural networks, which typically require thousands of labeled examples per class, and particularly inadequate for Transformer architectures that demand even larger datasets to learn meaningful global representations [18]. Moreover, the extreme scarcity of malignant cases precluded any possibility of learning discriminative features for the most clinically critical class.

Pragmatic Solution: Multi-Regional Dataset Integration

To address this fundamental constraint while advancing the methodological framework, we adopted a pragmatic approach by incorporating the King Abdulaziz University Breast Cancer Mammogram Dataset (KAU-BCMD) [27, 28], acquired programmatically through Kaggle Hub. This publicly available dataset comprises 1,991 diagnostic mammographic images organized according to the BI-RADS (Breast Imaging Reporting and Data System) classification framework [26], with the following distribution after preprocessing:

1) BI-RADS 1 (Normal/Benign): 1,865 images (93.7%)
2) BI-RADS 4 (Suspicious Abnormality): 102 images (5.1%)
3) BI-RADS 5 (Highly Suggestive of Malignancy): 24 images (1.2%)

The Kenyan dataset (100 images) was integrated during preprocessing but represented less than 20% of the total training data due to volume constraints. We explicitly acknowledge that the KAU-BCMD dataset originates from Saudi Arabia, not Sub-Saharan Africa. This geographic mismatch introduces important limitations which are as follows.

1) Population Differences: Breast density distributions, genetic factors, and disease phenotypes may differ between Middle Eastern and Sub-Saharan African populations.
2) Imaging Technology: Equipment specifications, acquisition protocols, and image quality characteristics may vary between regions.
3) Disease Prevalence: The distribution of BI-RADS categories in the training data may not reflect the prevalence patterns in African screening settings.

The decision to use this dataset represents a pragmatic trade-off: it enabled development of the technical framework and methodological pipeline, but the model's performance on actual African populations remains unvalidated and uncertain. This limitation is fundamental and discussed extensively in Section 5.

3.2. Data Preprocessing and Augmentation

The preprocessing pipeline incorporated Contrast Limited Adaptive Histogram Equalization (CLAHE) [29], enhancement to improve image quality and feature visibility. CLAHE was applied in the LAB color space with a clip limit of 2.0 and tile grid size of 8×8, enhancing local contrast while preventing over-amplification of noise. This enhancement proved particularly valuable for revealing subtle mammographic features that might otherwise remain obscured in low-contrast regions.

Data partitioning employed stratified sampling to maintain proportional class representation across all subsets. The dataset was divided into 70% training, 15% validation, and 15% test sets using scikit-learn's train_test_split function with explicit stratification parameters and fixed random state (seed=42) for reproducibility. This stratified approach eliminated sampling variance and ensured reliable performance estimates across all diagnostic categories, particularly critical for rare but clinically important malignant findings.

Comprehensive data augmentation was implemented exclusively on the training set through Keras' ImageDataGenerator framework. Eight simultaneous transformation types

were applied: rotation within ±15 degrees, width and height shift up to 10%, shear transformations within 0.1 range, zoom variations between 90-110%, and horizontal flipping. All transformations represented clinically realistic variations occurring in actual mammographic practice, ensuring augmented images remained within the distribution of authentic mammograms. This augmentation strategy expanded the effective training set size approximately six-fold, preventing overfitting while forcing the model to learn position-invariant and orientation-invariant features. Validation and test sets received only normalization without augmentation to ensure unbiased performance evaluation.

3.3. Model Architecture Design

The model architecture implemented a hybrid CNN-Transformer approach with multi-task learning capabilities. The backbone utilized ResNet50 pre-trained on ImageNet as the base feature extractor, with the first 100 layers frozen to preserve fundamental visual features while allowing later layers to adapt to mammographic patterns. The CNN features were reshaped and fed into a Transformer encoder consisting of three blocks, each containing multi-head attention mechanisms (8 heads, 64-dimensional keys) followed by feed-forward networks with 2048 hidden units. This hybrid architecture combined ResNet50's proven capability for hierarchical feature extraction with Transformers' ability to model long-range dependencies and spatial relationships critical for mammographic interpretation.

The backbone output fed into five specialized branches implementing multi-task learning: risk assessment, cancer detection, cancer staging, risk factor analysis, and differential diagnosis. Each branch comprised task-specific dense layers with progressive dimensionality reduction (256 to 128 neurons), incorporating batch normalization, dropout regularization (0.3-0.4 rates), and activation functions (ReLU or GELU) optimized for each task. All branches produced SoftMax outputs over the four BI-RADS categories, enabling probabilistic predictions with interpretable confidence estimates. This multi-branch architecture allowed the model to learn shared representations across related diagnostic tasks while maintaining task-specific specialization in final classification layers.

3.4. Training Protocol and Optimization

Training implemented a two-stage transfer learning strategy optimized for medical imaging adaptation. Stage one (epochs 1-15) maintained all ResNet50 layers frozen while training only the newly initialized Transformer blocks and classification branches. The Adam optimizer was configured with learning rate 1×10^{-4}, $\beta_1=0.9$, $\beta_2=0.999$, enabling rapid adaptation of randomly initialized layers without disrupting pre-trained features. Stage two allowed selective fine-tuning of the top 30% of ResNet50 layers to specialize high-level features for mammographic patterns, while preserving universally useful low-level features in frozen bottom layers.

The multi-task loss function combined categorical cross-entropy across all five branches with differential weighting: risk assessment (1.0), cancer detection (1.2), staging (1.0), risk factors (0.8), and differential diagnosis (1.0). The elevated weight for cancer detection reflected its paramount clinical importance. Training employed batch size of 16 images, determined by GPU memory constraints for the high-resolution inputs and deep architecture.

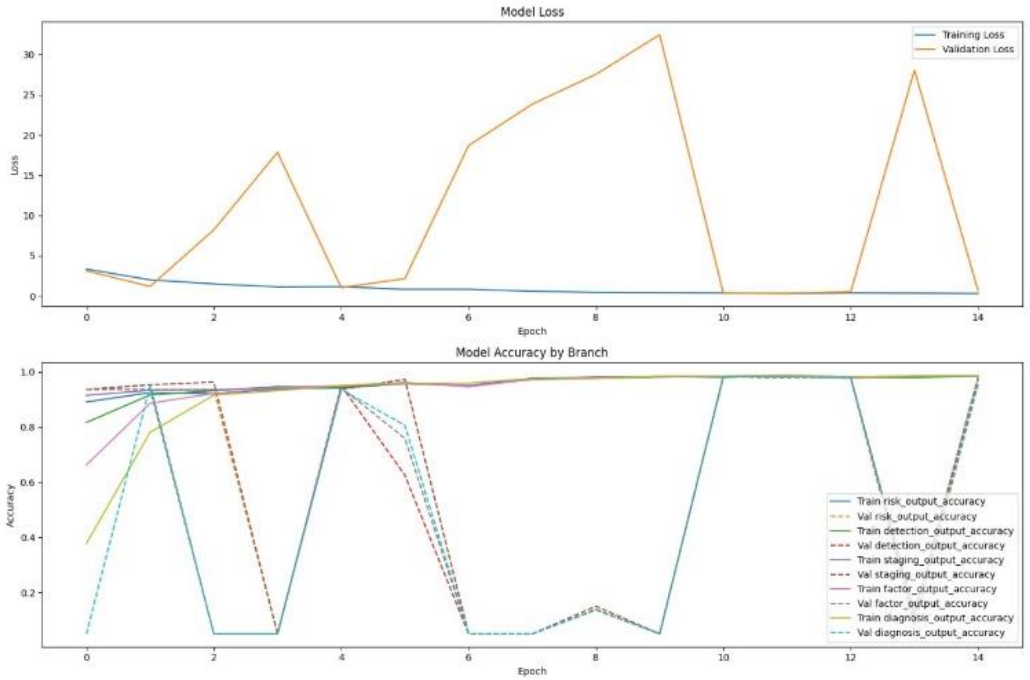

Figure 1. *Training and validation performance across epochs.*

Comprehensive regularization strategies prevented overfitting: progressive dropout schedules (0.5 to 0.4 to 0.3), early stopping monitoring validation loss with patience of 10 epochs, and adaptive learning rate reduction by factor 0.5 when validation loss plateaued for 5 consecutive epochs with minimum rate 1×10^{-6}. Model checkpointing saved the best-performing weights based on validation loss. Deterministic reproducibility was ensured through comprehensive seed management across NumPy (random.seed=42), Python (random.seed=42), TensorFlow (tf.random.set_seed=42), and environment variables (PYTHONHASHSEED="42", TF_DETERMINISTIC_OPS="1"). The training dynamics across epochs are illustrated in Figure 1, demonstrating convergence patterns and the effectiveness of the two-stage learning strategy.

3.5. Evaluation Methodology and Interpretability

Model evaluation employed the held-out test set with metrics aligned to clinical priorities. Primary metrics included overall classification accuracy, per-class precision and recall, and F1-scores as shown in Table 1. For clinical applicability, predictions were collapsed into binary classification (BI-RADS 4-5 as positive requiring biopsy versus BI-RADS 1-3 as negative for surveillance only), computing sensitivity, specificity, positive predictive value, and negative predictive value. Target thresholds of ≥95% sensitivity and ≥88% specificity reflected the paramount importance of minimizing false negatives in cancer screening while managing false positive burden.

Interpretability was enhanced through Gradient-weighted Class Activation Mapping (Grad-CAM) [30, 32] implementation, generating heatmap visualizations highlighting image regions most influential to model predictions. Grad-CAM operated on the hybrid backbone output, computing gradients of branch-specific class scores with respect to intermediate feature maps. The resulting attention heatmaps underwent spatial upsampling to original image dimensions and overlay onto input mammograms, enabling visual validation that model attention focused on clinically relevant features rather than spurious correlations.

Uncertainty quantification employed Monte Carlo dropout with 10 forward passes per image during inference, [31], computing mean predictions and standard deviations across iterations. High uncertainty flagged cases requiring human expert review, implementing a confidence-based triage system where predictions below 70% confidence triggered mandatory radiologist examination. This approach combined automated efficiency for confident predictions with human oversight for ambiguous cases, optimizing clinical workflow integration.

Table 1. *Classification report of evaluation metrics.*

Clinical Task	Accuracy	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)	Precision, Recall, F1-Score	
					Class 2	Class 0
Risk	0.98	0.79	0.95	0.86	0.00	1.00
Detection	0.98	0.76	0.95	0.84	0.00	1.00
Staging	0.98	0.79	0.95	0.86	0.00	1.00
Factor	0.98	0.79	0.95	0.86	0.00	1.00
Diagnosis	0.98	0.79	0.95	0.86	0.00	1.00

This represents an unacceptable clinical failure requiring urgent addressing before any deployment consideration.

4. Discussion

4.1. Dataset Characteristics and Preprocessing

The study utilized the KAU-BCMD mammography dataset comprising 1,991 mammogram images across four BI-RADS categories. After preprocessing and class consolidation, the final dataset included three classes: BI-RADS 1 (normal/benign), BI-RADS 4 (suspicious abnormality), and BI-RADS 5 (highly suggestive of malignancy). The dataset exhibited significant class imbalance, with BI-RADS 1 representing 93.7% (1,865 images), BI-RADS 4 representing 5.1% (102 images), and BI-RADS 5 representing 1.2% (24 images) of the total samples. Images were preprocessed using Contrast Limited Adaptive Histogram Equalization (CLAHE) enhancement and resized to 224×224 pixels. The data was partitioned into training (64.9%, n=1,544), validation (15.0%, n=357), and test (20.0%, n=476) sets.

4.2. Model Architecture and Training Configuration

A hybrid deep learning architecture was developed combining ResNet50 as the backbone feature extractor with multi-head attention mechanisms and five specialized output branches (risk assessment, detection, staging, prognostic factors, and diagnosis). The model contained 63,259,791 total parameters, with 59,123,471 trainable parameters (225.54 MB) and 4,136,320 non-trainable parameters (15.78 MB). Training was conducted over 15 epochs using a multi-output architecture with individual dense layers (256 units) for each branch, followed by batch normalization/layer normalization, dropout regularization, and final classification layers (128 units leading to 3-class outputs). The initial learning rate was set at 1×10^{-4} and reduced to 5×10^{-5} after epoch 10. Training convergence was achieved with the overall loss decreasing from 4.0098 in epoch 1 to 0.4272 in epoch 15.

4.3. Model Performance Metrics

4.3.1. Overall Weighted Accuracy: A Misleading Metric

All five output branches achieved high weighted accuracy on the test set (98% across risk assessment, detection, staging, factor, and diagnosis branches). However, this metric is highly misleading due to severe class imbalance, where BI-RADS 1 (normal/benign) cases constitute 93.7% of the dataset.

The macro-averaged F1-scores tell a markedly different story: only 0.61-0.62 across all branches, revealing poor performance on minority classes that are clinically most critical. The model's performance should therefore be interpreted not as "98% accurate" but rather as "highly accurate on normal cases, with critical failures on malignant cases.

In binary Clinical Classification Performance when evaluating the model's ability to discriminate between cases requiring biopsy (BI-RADS 4-5) versus surveillance only (BI-RADS 1-3), Sensitivity for Cancer Detection (BI-RADS 4-5 combined) was 76%, Specificity based on BI-RADS 1 was approximately 100%.

Most critically, the model achieved 0% sensitivity for BI-RADS 5 (highly malignant) cases, representing a complete failure on the most urgent clinical category.

4.3.2. Class-Specific Performance

(i). Risk Assessment Branch

1) BI-RADS 1 (Normal/Benign): Perfect performance with precision of 1.00, recall of 1.00, and F1-score of 1.00 (n=374)
2) BI-RADS 4 (Suspicious): Strong performance with precision of 0.79, recall of 0.95, and F1-score of 0.86 (n=20)
3) BI-RADS 5 (Malignant): Failed to detect any cases, with precision, recall, and F1-score all at 0.00 (n=5)

(ii). Detection Branch

1) BI-RADS 1: Perfect performance with precision of 1.00, recall of 1.00, and F1-score of 1.00 (n=374)
2) BI-RADS 4: Good performance with precision of 0.76, recall of 0.95, and F1-score of 0.84 (n=20)
3) BI-RADS 5: Complete failure to detect with all metrics at 0.00 (n=5)

(iii). Staging, Factor, and Diagnosis Branches

1) All three branches showed identical performance patterns:
2) BI-RADS 1: Perfect classification (precision, recall, F1-score all 1.00)
3) BI-RADS 4: Strong performance (precision 0.79, recall 0.95, F1-score 0.86)
4) BI-RADS 5: Complete classification failure (all metrics 0.00)

4.3.3. Critical Failure Analysis: BI-RADS 5 Misclassification

Across all five branches, the model failed to correctly identify any of the 5 BI-RADS 5 cases in the test set, misclassifying all as BI-RADS 4 (suspicious abnormality). This represents a 100% false negative rate for the most critical diagnostic category cases that in clinical practice require immediate biopsy and expedited treatment planning. This failure can be attributed to several factors which are:

1) Extreme Class Imbalance: With only 24 BI-RADS 5 cases in the entire dataset (1.2%), the model had insufficient examples to learn discriminative features.
2) Class Overlap: Visual similarity between BI-RADS 4 and 5 lesions without adequate training examples.
3) Loss Function Limitations: Despite task weighting, the categorical cross-entropy loss failed to adequately penalize minority class errors.
4) Insufficient Regularization Strategies: No class-specific balancing techniques (focal loss, class weighting, resampling) were implemented.

From a clinical safety perspective, this failure is unacceptable for deployment as it would result in delayed diagnosis and treatment for the patients most urgently requiring intervention.

4.4. Confusion Matrix Analysis

The confusion matrices revealed consistent patterns across all branches. For BI-RADS 1, the model correctly classified all 374 cases with no misclassifications. For BI-RADS 4, the model correctly identified 19 out of 20 cases, with one case misclassified as BI-RADS 1. Critically, all 5 BI-RADS 5 cases were misclassified as BI-RADS 4, indicating the mod-

el's complete inability to identify the highest-risk malignant category.

4.5. Training Dynamics

The training process exhibited instability in validation performance, particularly in epochs 3-4 and 7-10, where validation loss spiked dramatically (reaching 32.44 in epoch 10) while training loss continued to decrease steadily. This pattern suggests potential overfitting and insufficient regularization despite dropout implementation. The validation accuracy for several branches dropped to chance levels (5.02%-15.05%) during these unstable epochs before recovering. The learning rate reduction at epoch 11 helped stabilize training, with validation loss improving from 32.44 to 0.37 by epoch 11 and maintaining stable performance through epoch 15.

4.6. Model Limitations and Class Imbalance Impact

The severe class imbalance had a profound impact on model performance. The macro-averaged F1-scores across all branches were only 0.61-0.62, substantially lower than the weighted averages of 0.98, highlighting the model's poor performance on minority classes. The complete failure to classify BI-RADS 5 cases (representing only 1.2% of the dataset) demonstrates that the model learned to bias predictions toward the dominant BI-RADS 1 class. The model's tendency to misclassify all high-risk malignant cases as suspicious abnormalities poses significant clinical concerns, as it could lead to delayed diagnosis and treatment of malignant breast lesions.

4.7. Clinical Implications

While the model demonstrates high overall accuracy (>98%), this metric is misleading in the clinical context due to class imbalance. The model's inability to identify any BI-RADS 5 cases represents a critical failure for clinical deployment, as these are the cases most requiring urgent medical intervention. The high recall (0.95) for BI-RADS 4 cases suggests the model is sensitive to detecting abnormalities, but the complete misclassification of BI-RADS 5 cases as BI-RADS 4 indicates insufficient discrimination between suspicious and highly malignant findings. This limitation would necessitate that all flagged cases undergo expert radiologist review, potentially reducing the model's utility as a screening tool.

4.8. Geographic Generalizability and External Validity Concerns

A fundamental limitation of this study is the geographic mismatch between the training data (predominantly from Saudi Arabia) and the stated target population (African women). While both regions face challenges in breast cancer screening and diagnostic access, important differences exist.

1) Population and Disease Characteristics: Breast density patterns vary across ethnic populations and may affect both imaging appearance and cancer risk profiles. Genetic and molecular subtypes of breast cancer show population-specific variations. Age distribution of breast cancer diagnosis differs between Middle Eastern and Sub-Saharan African populations.
2) Healthcare and Imaging Infrastructure: Mammography equipment specifications and maintenance standards vary across settings. Image acquisition protocols and quality control measures differ between high-resource (Saudi Arabia) and low-resource (many Sub-Saharan African) settings. The BI-RADS distribution in screening populations may differ substantially from our training data.
3) Implications for Model Generalizability: The model's performance metrics reported here (98% weighted accuracy, 0% BI-RADS 5 recall) were obtained on a test set drawn from the same source as the training data (KAU-BCMD). External validation on independently collected African mammography datasets is essential but has not been performed. The model's actual performance on African populations could be Lower overall due to domain shift in imaging characteristics, Different in class distribution where African screening populations may present different BI-RADS category proportions and Variable across regions where Sub-Saharan Africa is highly diverse demographically and in healthcare infrastructure.

5. Limitations and Ethical Considerations

This study, while establishing a foundational pipeline for a hybrid CNN-Transformer based approach to early detection and diagnosis of breast cancer among African women, is not without its limitations. We recognize several key areas that merit further discussion.

5.1. Geographic Representativeness and Dataset Provenance: A Critical Limitation

The most significant limitation of this study is that the model was trained predominantly on data from Saudi Arabia (King Abdulaziz University Breast Cancer Mammogram Dataset), not from African populations. While the long-term objective of this research is to develop diagnostic tools applicable to African healthcare contexts, the current model has not been validated on African populations and its performance in African settings remains entirely unknown.

This decision was necessitated by a fundamental challenge

in African medical AI research. The critical scarcity of large-scale, annotated mammography datasets from African healthcare facilities. Despite six months of prospective data collection from partnering hospitals in, we obtained only 100 mammogram images, with only 5 malignant (BI-RADS 5) cases. This volume was insufficient for training deep learning models, which typically require thousands of labeled examples per class.

The KAU-BCMD dataset provided sufficient sample size (1,991 images) for initial model development, Standardized BI-RADS annotations publicly available access enabling reproducibility. However, this pragmatic solution came with significant costs to external validity. Below are some specific concerns regarding generalizability.

1) Phenotypic Differences: Breast density patterns, tissue composition, and lesion appearance characteristics may differ between Middle Eastern and Sub-Saharan African populations.
2) Disease Epidemiology: The distribution of breast cancer subtypes, stages at presentation, and BI-RADS category prevalence may differ substantially.
3) Imaging Technology Heterogeneity: The mammography equipment, protocols, and image quality in the training dataset may not reflect the diversity of imaging conditions across African healthcare facilities, particularly in low-resource settings.
4) Demographic Factors: Age distributions, socioeconomic factors, and risk factor profiles differ between Saudi Arabia and Sub-Saharan African populations.

Given these limitations, this model should NOT be deployed in African clinical settings without:

1) Prospective external validation on large, multi-center datasets from diverse Sub-Saharan African countries.
2) Performance evaluation stratified by country, hospital setting, imaging equipment type, and patient demographics.
3) Domain adaptation techniques specifically developed to address the Saudi Arabia the Africa distribution shift
4) Regulatory approval based on Africa-specific validation data.

The current study should be interpreted as developing a methodological framework and technical pipeline, not as providing a clinically validated tool for African healthcare.

5.2. Severe Class Imbalance and Complete Failure on Critical Cases

Beyond geographic limitations, the model exhibits a critical clinical failure: 0% detection rate for BI-RADS 5 (highly malignant) cases. This failure stems from severe class imbalance (93.7% normal vs. 1.2% highly malignant) and represents an unacceptable safety risk.

Clinically BI-RADS 5 cases require immediate biopsy, expedited pathology review, rapid treatment planning and high-sensitivity detection (ideally >95%). A model that misses 100% of these cases provides no clinical value and could cause harm through false reassurance. Below are the proposed solutions for future work.

1) Advanced Data Augmentation: Generative Adversarial Networks (GANs) to synthesize realistic BI-RADS 5 for example advanced augmentation techniques specifically targeting minority classes.
2) Class-Imbalance Aware Loss Functions: Focal loss with γ=2-5 to down-weight easy for example class-balanced loss with effective number weighting and custom loss heavily penalizing BI-RADS 5 false negatives.
3) Resampling Strategies: Oversampling minority classes with SMOTE or ADASYN. Undersampling majority class to balance training. Hybrid approaches combining both.
4) Ensemble and Specialized Models: Two-stage detection: binary classifier (normal vs. abnormal) followed by multi-class refinement. Specialized sub-models trained exclusively on BI-RADS 4-5 discrimination. Ensemble of models with different minority class emphasis.
5) Increased Data Collection: Prioritize prospective collection of malignant cases from multiple African centers. Data sharing agreements between African cancer centers. Inclusion of existing archived pathology-confirmed malignant cases.
6) Current Model Status: Given the 0% BI-RADS 5 detection rate, this model is NOT suitable for clinical use in its current form and requires fundamental architectural and training improvements.

5.2. Dataset Biases, Equity and Fairness

A significant limitation of this study is in representativeness of dataset, Mammography datasets used may not fully capture the diversity of breast cancer presentations and demographic factors across African Continent. Differences in breast density, imaging resolution and acquisition protocols between hospitals or regions could affect generalization. To ensure broader clinical applicability, future research should include multi-center datasets drawn from varied healthcare setting, including low-resource and rural environments. Future work must focus on actively curating a more diverse dataset that is representative of the broader Africa population, including Northern and West Africa.

6. Conclusion

This study developed a hybrid CNN–Transformer framework (MAMMOAI) for multi-task breast cancer analysis from mammograms, addressing a critical healthcare challenge in Sub-Saharan Africa, where over 70% of breast cancer cases are detected at late stages. The technical implementation successfully integrates ResNet50-based feature extraction with Transformer attention mechanisms across five diagnostic branches: risk assessment, lesion detection,

staging, risk factor analysis, and differential diagnosis. Under this controlled experimental setting, the framework achieved a weighted accuracy of 98%.

Despite this strong technical performance, critical limitations preclude any form of clinical deployment and must be explicitly acknowledged. The model was trained predominantly on data from Saudi Arabia (King Abdulaziz University dataset) rather than African populations, due to the well-documented scarcity of large, annotated mammography datasets from Sub-Saharan Africa. As a result, the model's performance on African populations remains entirely unvalidated. Prospective, multi-center external validation across diverse African healthcare settings is therefore essential before any claims of regional applicability can be supported.

Furthermore, the framework demonstrated 0% recall for BI-RADS 5 cases, representing a complete failure to detect the most clinically urgent and highly malignant category requiring immediate intervention. This limitation is attributable to severe class imbalance in the training data (93.7% normal cases versus only 1.2% highly malignant cases) and insufficient handling of minority classes. Although overall accuracy was high, the macro F1-score (0.61–0.62) reveals that this performance is dominated by the majority class, rendering accuracy alone clinically misleading.

In its current form, MAMMOAI is not suitable for clinical use. Rather, it should be understood strictly as a methodological and technical proof-of-concept. Substantial additional work is required before translation toward clinical application, including: (i) acquisition of larger, balanced datasets with adequate malignant case representation; (ii) implementation of advanced class-imbalance mitigation strategies such as focal loss, generative data augmentation, and specialized minority-class modeling; (iii) rigorous external validation across multiple African countries and healthcare settings; and (iv) domain adaptation techniques to address distributional shifts between Saudi Arabian and African populations.

The primary contribution of this study lies in demonstrating the technical feasibility of hybrid CNN–Transformer architectures for multi-task mammography analysis, providing an interpretable framework through Grad-CAM visualizations and uncertainty quantification, and transparently documenting the major challenges of data scarcity, class imbalance, and geographic generalizability. Future research must prioritize data collection, class balancing, external validation, and clinically grounded system integration.

At length, while the long-term objective remains the development of early, accurate, and accessible breast cancer diagnostic support tools for African women, this study clearly demonstrates that achieving this goal requires addressing fundamental limitations in dataset availability, population representativeness, and model generalizability. The work presented here constitutes a methodological framework rather than a deployable clinical tool, offering a foundation upon which clinically viable solutions may be built through sustained, collaborative, and data-driven efforts.

Abbreviations

AI	Artificial Intelligence
CNN	Convolutional Neural Networks
GRAD-CAM	Gradient-weighted Class Activation Mapping
BI-RADS	Breast Imaging Reporting and Data System
KAU	King Abdulaziz University

Acknowledgments

The authors are grateful to DataLab at Eastern Africa training Centre and Zetech University for providing an enabling research environment that facilitated this work.

Author Contributions

Kasim Chambulilo: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing

Paul Wako: Formal Analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft

Davis Bassa: Conceptualization, Data curation, Investigation, Resources, Software, Supervision, Validation

Asha Kisonga: Investigation, Methodology, Project administration, Software, Su Validation, Visualization, Writing – original draft

Emmanuel Shija: Supervision, Writing – original draft

Funding

This work is not supported by any external funding.

Data Availability Statement

The data that support the findings of this study can be found at: https://www.kaggle.com/datasets/asmaasaad/king-abdulaziz-university-mammogram-dataset. We acknowledge that the primary dataset used (King Abdulaziz University Mammogram Dataset) is publicly available but originates from Saudi Arabia, not Sub-Saharan Africa. The supplementary Kenyan hospital data cannot be publicly shared due to ethical restrictions and patient privacy considerations. We recognize this limits reproducibility for the Africa-specific components and emphasize the urgent need for the African medical AI community to develop large-scale, ethically collected, shareable mammography datasets to advance equitable diagnostic AI development.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] International Agency for Research on Cancer (IARC), Breast cancer overall survival, annual risks of death, and survival gap apportionment in sub-Saharan Africa (ABC-DO): Seven-year follow-up of a prospective cohort study. The Lancet Global Health; 2025. Available from: https://www.iarc.who.int/news-events/breast-cancer-overall-survival-annual-risks-of-death-and-survival-abc-do/

[2] World Health Organization (WHO). Assessment of breast cancer control capacities in the WHO African Region. Brazzaville: WHO Regional Office for Africa; 2022. Available from: https://www.afro.who.int/publications/assessment-breast-cancer-control-capacities-who-african-region-2022-0

[3] Demographic and Health Surveys (DHS) Study. Barriers to clinical breast examination in sub-Saharan Africa: Financial, social, and access challenges. African Researchers Network; 2025. Available from: https://www.africanresearchers.org/barriers-to-clinical-breast-examination-in-sub-saharan-africa-new-study-reveals-financial-social-and-access-challenges/

[4] Bayatmakou F, Taleei R, Amir Toutounchian M, Mohammadi A. Integrating AI for human-centric breast cancer diagnostics: A multi-scale and multi-view Swin Transformer framework. arXiv; 2025. Available from: https://arxiv.org/abs/2503.13309

[5] Mehmood A, Hu Y, Khan SH, et al. A novel channel-boosted residual CNN–Transformer with regional-boundary learning for breast cancer detection. arXiv; 2025. Available from: https://arxiv.org/abs/2503.15008

[6] Anyigba, C. A., Awandare, G. A., & Paemka, L. (2021). Breast cancer in sub-Saharan Africa: The current state and uncertain future. Experimental biology and medicine (Maywood, N. J.), 246(12), 1377–1387. https://doi.org/10.1177/15353702211006047

[7] Martei, Y. M., Dauda, B., & Vanderpuye, V. (2022). Breast cancer screening in sub-Saharan Africa: a systematic review and ethical appraisal. BMC cancer, 22(1), 203. https://doi.org/10.1186/s12885-022-09299-5

[8] Brahmareddy, A., Selvan, M. P. TransBreastNet a CNN transformer hybrid deep learning framework for breast cancer subtype classification and temporal lesion progression analysis. *Sci Rep* 15, 35106 (2025). https://doi.org/10.1038/s41598-025-19173-6

[9] Yala, A., Lehman, C., Schuster, T., Portnoi, T., Barzilay, R. A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. Radiology. 2019, 292(1), 60–66. https://doi.org/10.1148/radiol.2019182716

[10] Abdikenov, B., Zhaksylyk, T., Imasheva, A., & Rakishev, D. (2025). From Mammogram Analysis to Clinical Integration with Deep Learning in Breast Cancer Diagnosis. Informatics, 12(4), 106. https://doi.org/10.3390/informatics12040106

[11] Francesco Manigrasso, Rosario Milazzo, Alessandro Sebastian Russo, Fabrizio Lamberti, Fredrik Strand, Andrea Pagnani, Lia Morra, Mammography classification with multi-view deep learning techniques: Investigating graph and transformer-based architectures, Medical Image Analysis, Volume 99, 2025, 103320, https://doi.org/10.1016/j.media.2024.103320

[12] Tardy Mickael, Mateus Diana, Leveraging Multi-Task Learning to Cope With Poor and Missing Labels of Mammograms, Frontiers in Radiology, Volume 1 – 2021, 2022, https://doi.org/10.3389/fradi.2021.796078

[13] Chowdary J, Yogarajah P, Chaurasia P, Guruviah V. A Multi-Task Learning Framework for Automated Segmentation and Classification of Breast Tumors From Ultrasound Images. *Ultrasonic Imaging*. 2022; 44(1): 3-12. https://doi.org/10.1177/01617346221075769

[14] Talaat, F. M., Gamel, S. A., El-Balka, R. M., Shehata, M., & ZainEldin, H. (2024). Grad-CAM Enabled Breast Cancer Classification with a 3D Inception-ResNet V2: Empowering Radiologists with Explainable Insights. Cancers, 16(21), 3668. https://doi.org/10.3390/cancers16213668

[15] Mellado, D., Mayeta-Revilla, L., Sotelo, J., Querales, M., Godoy, E., Lever, S., Pardo, F., Chabert, S., & Salas, R. (2025). Identifying clinically relevant findings in breast cancer using deep learning and feature attribution on local views from high-resolution mammography. Frontiers in oncology, 15, 1601929. https://doi.org/10.3389/fonc.2025.1601929

[16] Mendes, J., Oliveira, B., Araújo, C., Galrão, J., Mota, A. M., Garcia, N. C., & Matela, N. (2025). Deep learning in breast cancer risk prediction: a review of recent applications in full-field digital mammography. Frontiers in oncology, 15, 1656842. https://doi.org/10.3389/fonc.2025.1656842

[17] Urooj, A., Dapamede, T., Patel, B., Ayoub, C., Arsanjani, R., O'Neill, W. C., Trivedi, H., & Banerjee, I. (2025). A Multi-Task Learning Approach for Segmentation of Breast Arterial Calcifications in Screening Mammograms. AMIA... Annual Symposium proceedings. AMIA Symposium, 2024, 1139–1148.

[18] Sharma, S., Singh, Y. & Choudhury, T. Advanced deep learning architectures for enhanced mammography classification: a comparative study of CNNs and ViT. Discov Artif Intell 5, 187 (2025). https://doi.org/10.1007/s44163-025-00426-2

[19] Mingshuang Fang, Binxiong Xu, Transformer-based multi-modal learning for breast cancer screening: Merging imaging and genetic data, Journal of Radiation Research and Applied Sciences, Volume 18, Issue 3, 2025, 101586, https://doi.org/10.1016/j.jrras.2025.101586

[20] Tekin, A. & Toktay, B. & Günay, A. C. & Yazgan, H. & İnan, N. G. & Kocadağlı, O. (2024). Bi-Rads Classification in Mammography using Deep Learning. In: Akoğul, S. & Tuna, E. (eds.), *Academic Studies with Current Econometric and Statistical Applications.* Özgür Publications. https://doi.org/10.58830/ozgur.pub518.c2134

[21] Luo, Y., Wei, J., Gu, Y., Zhu, C., & Xu, F. (2025). Predicting molecular subtype in breast cancer using deep learning on mammography images. Frontiers in oncology, 15, 1638212. https://doi.org/10.3389/fonc.2025.1638212

[22] Jeba Prasanna Idas, S., Hemalatha, K., Naveenkumar, J. *et al.* Recent trends on mammogram breast density analysis using deep learning models: neoteric review. *Artif Intell Rev* 58, 240 (2025). https://doi.org/10.1007/s10462-025-11232-8

[23] K. K. Sreekala, Jayakrushna Sahoo, Enhancing mammogram classification using explainable Conditional Self-Attention Generative Adversarial Network, Expert Systems with Applications, Volume 293, 2025, 128640, https://doi.org/10.1016/j.eswa.2025.128640

[24] Michael G. Kawooya, Richard Malumba, Alfred Jatho, Andrew Katumba, Joyce Nakatumba-Nabende, Denis Musinguzi, Simon Allan Achuka, Maruf Adewole, Udunna C Anazodo, medRxiv 2025.08.26.25334483; https://doi.org/10.1101/2025.08.26.25334483

[25] Al-Hejri, A. M., Sable, A. H., Al-Tam, R. M. et al. A hybrid explainable federated-based vision transformer framework for breast cancer prediction via risk factors. Sci Rep 15, 18453 (2025). https://doi.org/10.1038/s41598-025-96527-0

[26] D'Orsi, C. J., Sickles, E. A., Mendelson, E. B., & Morris, E. A. (2013). ACR BI-RADS Atlas, Breast Imaging Reporting and Data System. Reston, VA: American College of Radiology.

[27] Alsolami, A. S., Shalash, W., Alsaggaf, W., Ashoor, S., Refaat, H., & Elmogy, M. (2021). King Abdulaziz University Breast Cancer Mammogram Dataset (KAU-BCMD). *Data*, 6(11), 111. https://doi.org/10.3390/data6110111

[28] Alsolami, A. (2021). King Abdulaziz University Mammogram Dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/asmaasaad/king-abdulaziz-university-mammogram-dataset

[29] Pisano, E. D., Zong, S., Hemminger, B. M., DeLuca, M., Johnston, R. E., Muller, K.,... & Pizer, S. M. (1998). Contrast Limited Adaptive Histogram Equalization image processing to improve the detection of simulated spiculations in dense mammograms. Journal of Digital Imaging, 11(4), 193-200.

[30] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision. (p. 618-626.

[31] Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In International Conference on Machine Learning. pp. 1050-1059.

[32] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770-778.