International Journal of Computer Applications (0975 - 8887)
Volume 187 - No.147, October 2026

Cross-Attention Multimodal Fusion for Breast Cancer Diagnosis: Integrating Mammography and Clinical Data with Explainability

Muhaisin Tiyumba Nantogmah
Department of Computer Science
University for Development Studies
Northern Region, Ghana

Abdul-Barik Alhassan
Department of Computer Science
University for Development Studies
Northern Region, Ghana

Salamudeen Alhassan
Department of Computer Science
University for Development Studies
Northern Region, Ghana

ABSTRACT

While most computer-aided diagnosis (CAD) systems for breast cancer use image features largely from mammograms, structured clinical information has an important complementary diagnostic value and model reasoning remains largely unaddressed. The aim of this study was to explore the possibility of combining mammography with categorical clinical features to provide a more accurate pathologist classification of breast lesions, and to determine if explainable AI (XAI) can be used to make such models interpretable for clinical use. We propose three strategies for multimodal deep-learning fusion: feature concatenation, co-attention, and cross-attention; the latter is a novel technique that has not yet been explored in this context. We compare these three strategies in a systematic manner, and integrate a ResNet-50 based mammographic embedding with a one-hot-encoded clinical feature vector, taking into account different missing-modality scenarios that are present in clinical practice, using modality dropout. The interpretation of reasoning is done at the image level (Grad-CAM) and the feature level (SHAP & LIME). The proposed cross-attention fusion framework was evaluated based on the publicly available CBIS-DDSM benchmark, and the results were compared with the best unimodal baseline and simple fusion methods. The proposed fusion method outperforms the unimodal baseline and simple fusion methods, achieving an AUC-ROC of 0.98, accuracy of 0.96, F1-score of 0.94, precision of 0.92, and recall of 0.95, with the improvement confirmed by McNemar's test ($p < 0.01$) and stable bootstrap confidence intervals. The three types of explanations (Grad-CAM, SHAP, and LIME) agreed on clinically relevant areas of the images and clinically relevant variables and the case-level analysis of the correct and incorrect predictions demonstrated the value of the explanations and limitations of the framework. The results suggest that cross-attention-based multimodal fusion and multi-level explainability can meaningfully benefit the accuracy and the clinical trustworthiness of using AI assistance for breast cancer diagnosis, and provide a template for multimodal CAD systems that are interpretable and applicable to real-world, missing-data clinical workflows.

General Terms

Algorithms, Experimentation, Performance, Verification

Keywords

Breast cancer diagnosis, multimodal fusion, cross-attention, explainable artificial intelligence (XAI), mammography, CBIS-DDSM

1. INTRODUCTION

Breast cancer is the most commonly diagnosed cancer and a major cause of cancer death in women around the world. Global estimates from recent years have revealed that in 2022, around 2.3 million women were diagnosed with breast cancer, of which about 670,000 died, and that there will be a significant increase in these numbers by 2050, particularly in countries with lower Human Development Index ranking [17, 5]. Breast cancer presents in a variety of different clinical and radiologic patterns. The vast majority of cases (estimated 85-90

In recent years, deep learning (DL) and machine learning (ML) have been widely studied and developed in various fields across the globe, such as travel and transportation [29, 30] and, most importantly for this study, healthcare [18]. These methods demonstrate great potential in helping radiologists achieve more accurate diagnoses, as well as minimize the variance in performance between different observers [9]. Most of the previous research on breast imaging analysis has been on the coarse problem of benign or malignant detection of the lesion [28, 32]. Studies that occurred more recently have shifted to more subtle classification, including lesion morphology (mass versus calcification) and the estimation of breast-density category based on BI-RADS [32, 23, 6]. In fact, the images of lesions can be divided into five classes: benign mass, malignant mass, benign calcification, malignant calcification, and background because masses (three-dimensional space-occupying lesions) and calcifications (deposits of calcium salts) have distinctly different radiological features that can be diagnostically informative, themselves. Breast density is also a highly validated sub-problem in pathology assessment [16] with Matthews

et al. [23] achieving promising results for this sub-problem using a deep-learning density model that was validated across multiple sites, and Chougrad et al. [6] jointly modelling lesion type, mass, and pathology labels in a multi-label setting to make use of the possible associations between the three.

Machine learning and deep learning techniques have been able to achieve high accuracy in breast cancer detection and classification, but their clinical application has been greatly hampered by their opacity. These models are difficult to interrogate because they are "black-box," meaning they are not easy to understand, and they make it more challenging for healthcare professionals to trust the models for clinical decision-making, who need clear and evidence-based justifications for their models [4, 33]. The opacity is considered a physical hurdle to the broader use of AI in health care, especially in high-stakes scenarios where patient outcomes are at stake based on the diagnostic accuracy of AI. To address this requirement, models whose behavior is explainable to the users who depend on them have come into existence, known as explainable artificial intelligence (XAI) [7]. Very early on, the term XAI was coined by work for the Defense Advanced Research Projects Agency (DARPA) to create models that are highly accurate, understandable to human users, and therefore trusted and used appropriately [34].

Explainable AI has since gained increasing popularity in the field of breast cancer research with several recent systematic and scoping reviews citing dozens of articles that have applied SHAP, LIME, Grad-CAM, and related approaches to breast cancer detection, diagnosis, and risk prediction [11, 3]. This literature is, however, dominated by single-modality pipelines, where only imaging modalities are used, and empirical results regarding the interaction between fusion architecture and explanation quality in really multimodal settings are scarce. The gap is relevant in practice: Recent clinician-facing assessments suggest that the association between the degree of explanation and clinicians' trust and accuracy in diagnosis is not linear, and that adding explanations to a model automatically leads to better-supported clinical decisions is not true [27]. This highlights the importance of co-designing and co-evaluating explainability with the fusion architecture, not as an add-on feature.

As far as we know, most of the available approaches, such as the few recent multimodal approaches for breast pathology classification [37, 15, 14, 21, 26], either use imaging data exclusively or offer limited interpretation from a single modality. In this study, we present an explainable multimodal model that is able to boost pathology classification accuracy through a combination of both mammographic and clinical information, and can offer interpretability for the diagnostic reasoning on the basis of clinical feature labels included in the catalogue CBIS-DDSM [19] associated with the mammograms. We also assess the proposed framework together with co-attention or cross-attention to overcome the problem of missing clinical feature data, which is commonly found in real-world clinical applications.

This paper makes the following contributions:

(1) Multimodal integration of imaging and clinical features: Enhancing breast cancer pathology diagnosis by multimodal integration of imaging and clinical features. An embedding extracted from a pretrained deep learning model serves as the imaging feature representation, while five clinically established descriptors – breast density, mass shape, mass margin, calcification type, and calcification distribution – are encoded as a concatenated one-hot vector (with a zero vector substituted for any feature that is absent for a given case).

(2) Clinical characteristics are sometimes incomplete or missing in routine clinical practice, so we systematically evaluate the simple concatenation, co-attention and cross-attention fusion methods within the same architecture, with modality dropout during training, thus allowing the fusion model to maintain good performance under partial missing data and to be easily applied in practice.

(3) To facilitate clinical decision making, we use the Grad-CAM, SHAP, and LIME techniques to explain the model predictions at the image and feature levels to gain interpretability and trustworthiness. Such methods offer a visual and feature-based explanation that distinguishes the imaging regions and clinical attributes that underpin each prediction and allow clinicians to question and investigate the model's reasoning, thus building trust in its diagnostic recommendations.

2. STUDY METHODOLOGY

A multimodal fusion network framework has been developed.

2.1 Embedding of Mammogram

Let $x_i^{(m)} \in \mathbb{R}^{H\times W\times 3}$ denote the pre-processed mammogram for patient i, resized to 224×224 pixels. We use a ResNet-50 pre-trained on ImageNet [13] and fine-tuned on the CBIS-DDSM dataset, to extract discriminative mammographic features. The ResNet-50 architecture is based on the concept of residual connections, which enable the training of deep networks without significant drop in accuracy due to the vanishing gradient problem.

The ResNet-50 global average pooling (GAP) layer produces a d_e-dimensional embedding:

$$e_i = ResNet50\left(x_i^{(m)}\right), \quad e_i \in \mathbb{R}^{d_e},\ d_e = 2048. \tag{1}$$

Instead we take a compact latent representation by projecting the e_i into a d-dimensional space with an affine transformation and a nonlinear activation function:

$$\widetilde{\mathbf{e}}_i = F\left(W_e^T \bullet e_i + b_e\right), \quad W_e \in \mathbb{R}^{d_e\times d},\ b_e \in \mathbb{R}^d, \tag{2}$$

In our default setting, d = 100 and where $F(\bullet)$ is the Rectified Linear Unit (ReLU) activation. This small embedding helps to align the modality dimensions for easy downstream fusion.

2.2 Clinical Embedding

Structured clinical features is a combination of one-hot vectors with dimensions reflecting the granularity of their categories: breast density $\in \mathbb{R}^4$, mass shape $\in \mathbb{R}^8$. A feature without an associated vector is represented by two zero vectors (e.g. mass lesions have no calcification distribution or type) and the same treatment is applied to calcification lesions. The five have a total of 36 dimensions in our clinical embedding.

In our configuration, there are 36 clinical features and let $\vec{c}_i \in \mathbb{R}^{D_c}$ represent the clinical feature vector for patient i. We use a projection similar to the imaging one:

$$\widetilde{\mathbf{c}}_i = F\left(W_c^T \bullet c_i + b_c\right), \quad W_c \in \mathbb{R}^{D_c\times d},\ b_c \in \mathbb{R}^d. \tag{3}$$

This guarantees that both modalities have the same latent dimensionality allowing for symmetric fusion operations.

2.3 Multimodal Fusion

The network then carries out cross-modal integration using one of multiple fusion mechanisms between the embedding of the pro-

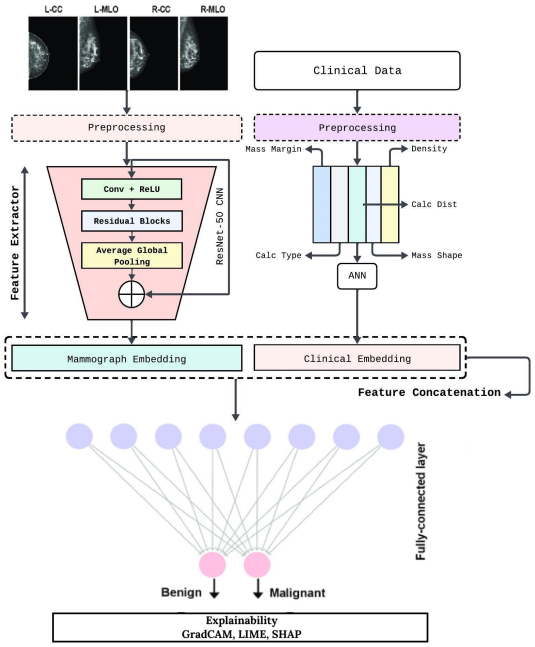


Fig. 1. The general framework of our proposed multimodal fusion model.

jected clinical embedding $\widetilde{\mathbf{c}}_i$ and the embedding of the projected mammography embedding $\widetilde{\mathbf{e}}_i$.
Baseline Concatenation (Intermediate Fusion)
The easiest way is to simply combine the two embeddings:

$$\mathbf{k}_i = Concat(\widetilde{\mathbf{e}}_i, \widetilde{\mathbf{c}}_i) \in \mathbb{R}^{2d} \tag{4}$$

where $\mathbf{k}_i$ is the last concatenated embedding. This intermediate fusion in this work only combines the two projected embedding vectors and other combination strategies could be tested to find if they can further improve performance. The generated vector is propagated through to dense layers for representation learning in a shared way.

2.4 Multimodal Fusion

After obtaining the projected clinical embedding $\widetilde{\mathbf{c}}_i$ and projected mammography embedding $\widetilde{\mathbf{e}}_i$, the network carries out cross-modal integration by one of multiple fusion mechanisms.
This is known as Baseline Concatenation (Intermediate Fusion).
The most straightforward is to join the two embeddings:

$$\mathbf{k}_i = Concat(\widetilde{\mathbf{e}}_i, \widetilde{\mathbf{c}}_i) \in \mathbb{R}^{2d} \tag{5}$$

where $\mathbf{k}_i$ is the last concatenated embedding. In this work this intermediate fusion only concatenates the two projected embedding vectors, but other possible combination approaches could be investigated to see if they may further enhance performance. The end result is sent through dense layers for representation learning in combination.

2.5 Co-attention and Cross-attention Strategies

In clinical contexts, a mammogram seldom comes with a full set of clinical attributes; and only a partial set of attributes is available, which is often the case as well, so the system in use needs to be able

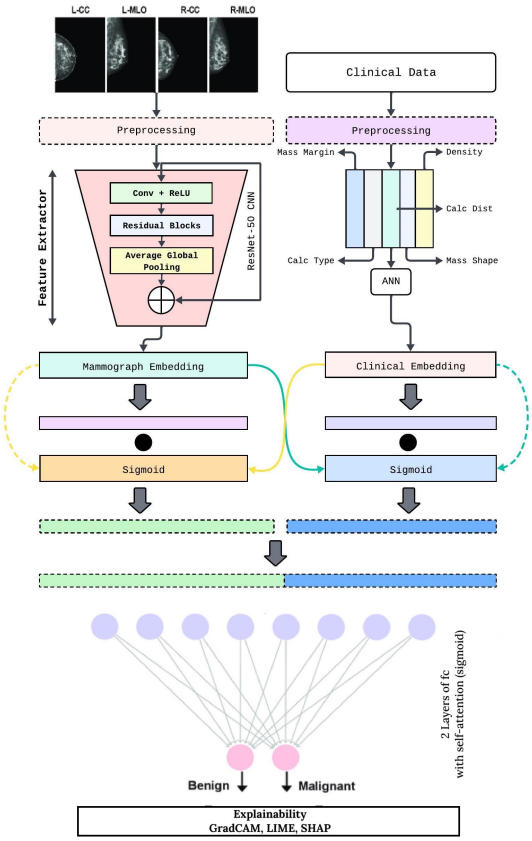


Fig. 2. Co-attention and cross-attention network design.

to handle missing value cases. Thus, both a cross-attention module and a co-attention module are evaluated with our model.
We introduce the cross-attention or co-attention module to our proposed architecture with the two dashed pathways in the figure being removed. Co-attention and cross-attention are used here to enable the model to learn the pertinent relationships between a mammogram and its associated clinical features. In Fig. 2 we present our proposed framework together with the co-attention/cross-attention module.

2.5.1 Co-attention Approach. In [22], a symmetric attention modeling framework named Co-attention was proposed in which both modalities are attending to each other at the same time. It creates representations that are mutually informed, in which the new features of each modality are based on the other. This is especially useful in our context where some radiological patterns (e.g., microcalcifications) are only clinically relevant in certain features (e.g., breast density).
Let:

—Let $E \in \mathbb{R}^{n_e \times d}$ be the list of image features (such as patches, regions, etc.), with n_e = number of image features and d = embedding dimension.

—Suppose that the embeddings of clinical features are a sequence of vectors $C \in \mathbb{R}^{n_c \times d}$, with n_c as the number of discrete clinical tokens.

—We have learnable projection matrices for queries, keys and values from the image modality $W_Q^e,\ W_K^e,\ W_V^e\ \in \mathbb{R}^{d\times d_k}$.

—$W_Q^c,\ W_K^c,\ W_V^c\ \in \mathbb{R}^{d\times d_k}$ are the analogous matrices for the clinical modality.

We first compute the bidirectional affinity scores between the modalities:

$$S_{ec} = \frac{(EW_Q^e)(CW_K^c)^T}{\sqrt{d_k}} \quad \in \mathbb{R}^{n_e\times n_c}, \tag{6}$$

$$S_{ce} = \frac{(CW_Q^c)(EW_K^e)^T}{\sqrt{d_k}} \quad \in \mathbb{R}^{n_c\times n_e}, \tag{7}$$

Here:

—$S_{ec}[i,j]$ measures how much the i-th image token attends to the j-th clinical token.

—$S_{ce}[p,q]$ measures how much the p-th clinical token attends to the q-th image token.

The attention weights are obtained by row-wise softmax normalisation:

$$\alpha_{ec} = softmax\left(S_{ec}\right) \text{ over clinical tokens,} \tag{8}$$

$$\alpha_{ce} = softmax\left(S_{ce}\right) \text{ over image tokens.} \tag{9}$$

Then we calculate features that are updated in the context:

$$E' = \alpha_{ec}\left(CW_V^c\right) \ \in \mathbb{R}^{n_e\times d_k}, \tag{10}$$

$$C' = \alpha_{ce}\left(EW_V^e\right) \ \in \mathbb{R}^{n_c\times d_k}. \tag{11}$$

Finally, we push back to the original dimension d and add residual connections and layer normalisation:

$$\widehat{E} = LN\left(E + E'W_O^e\right), \quad \widehat{C} = LN\left(C + C'W_O^c\right), \tag{12}$$

where $W_O^e,\ W_O^c\ \in \mathbb{R}^{d_k\times d}$ are learnable output projections.
The co-attention mechanism further involves re-weighting every image token using all clinical tokens, and every clinical token using all image tokens. It is a symmetric interaction that is especially appropriate for databases where both modalities have complementary diagnostic information.

2.5.2 Cross-attention Approach. In related multimodal recognition tasks [1], the cross-attention is asymmetric, with one modality being used as the *query* and the other as the *key/value* in the cross-attention. This enables provision of targeted conditioning, for instance, enhancing the image features only for the most relevant clinical features, without reciprocally changing the clinical representations (or vice versa).
Suppose that clinical features are conditions for the image features. We solve the queries with the image modality and obtain the keys/values with the clinical modality:

$$Q_E = EW_Q^e \ \in \mathbb{R}^{n_e\times d_k}, \tag{13}$$

$$K_C = CW_K^c \ \in \mathbb{R}^{n_c\times d_k}, \tag{14}$$

$$V_C = CW_V^e \ \in \mathbb{R}^{n_c\times d_k}. \tag{15}$$

The attention score matrix is:

$$S_{E\leftarrow C} = \frac{Q_E K_C^T}{\sqrt{d_k}} \ \in \mathbb{R}^{n_e\times n_c}. \tag{16}$$

We normalise across clinical tokens:

$$\alpha_{E\leftarrow C} = softmax\left(S_{E\leftarrow C}\right), \tag{17}$$

and compute the updated image features:

$$E' = \alpha_{E\leftarrow C} V_C. \tag{18}$$

The updated image embedding is:

$$\widehat{E} = LN\left(E + E'W_O^e\right). \tag{19}$$

Reverse conditioning (clinical on image) is realized by exchanging the roles of E and C in the Equations (12)–(15).
Cross attention, on the other hand, is more unidirectional, and can enforce a hierarchy such as conditioning the high dimensional image space on a structured clinical prior, or enriching sparse clinical vectors with spatially dense radiological cues. This is especially helpful if one modality has more information or is more comprehensive than the other.

2.5.3 Missing Modalities. In order to deal with clinical data or imaging data that is missing, we use modality dropout during training by randomly setting the embedding of one modality to zero with probability p, and use random withholding of clinical data for a fraction of the training samples to promote the redundancy of the learned representation.

2.6 Explainability Integration

2.6.1 Grad-CAM for Imaging Modality. Grad-CAM gives class-discriminative localisation maps, and identifies the areas of the mammogram that are most important for making a prediction. It works on the feature maps of a convolutional layer, and generates importance weights through the gradients of the target class score on top of these feature maps.
Let:

—y^c denote the output score (pre-softmax) for class c (e.g., malignant).

—$A^k\ \in \mathbb{R}^{u\times v}$ be the activation map of the k-th channel in the target convolutional layer.

—α_k^c be the importance weight for channel k with respect to class c.

The importance weights are calculated as:

$$\alpha_k^c = \frac{1}{Z}\sum_{i=1}^{u}\sum_{j=1}^{v}\frac{\partial y^c}{\partial A_{ij}^k}, \tag{20}$$

where $Z = u\times v$ is the number of spatial locations in A^k. The Grad-CAM heatmap $L_{GradCAM}^c$ is then:

$$L_{GradCAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right), \tag{21}$$

where ReLU will ensure that only features that are positively correlated with the class are visualised. The heatmap is upsampled to the resolution of the input image and is overlaid on the input mammogram to aid expert inspection.
For each mammogram, we used Grad-CAM in the last convolutional block of ResNet-50. The resulting maps were then compared to pathologically meaningful areas, such as mass edges, calcification clusters and ensured that highlighted regions were clinically relevant.

2.6.2 SHAP for Clinical Feature Attribution. SHAP gives an input feature its Shapley value from cooperative game theory, which is the average contribution of the feature to the prediction for all possible coalitions of input features.
Let:

—$f(x)$ denote the model output (logit or probability) for input vector $\mathbf{x}$.

—M be the number of input features.

—$S \subseteq \{1, \cdots, M\} \setminus \{i\}$ be a subset of features excluding feature i.

—$f_{S\cup\{i\}}$ denote the model outputs when using only features in $S \cup \{i\}$ or S, respectively.

The Shapley value for feature i is:

$$\phi_i = \sum_{S\subseteq\{1,\cdots,M\}\setminus\{i\}} \frac{|S|!\,(M-|S|-1)!}{M!} \left[f_{S\cup\{i\}}\left(\mathbf{x}_{S\cup\{i\}}\right) - f_S\left(\mathbf{x}_S\right)\right], \tag{22}$$

where:

—$\phi_i > 0$ indicates that feature i increases the likelihood of the predicted class.

—$\phi_i < 0$ indicates a negative contribution.

We then used KernelSHAP to rank the clinical variables (e.g., breast density, calcification type) at the feature level for each individual prediction by applying it to the clinical embedding $\widetilde{\mathbf{c}}_i$. The importance of the features was computed globally, by summing up the absolute values of the features across the data set, so important clinical features across the data set could be identified.

2.6.3 LIME for Local Surrogate Modelling. LIME locally trains a simple and interpretable surrogate model (e.g., sparse linear regression) to explain individual predictions.
Given:

—An instance $\mathbf{x} \in \mathbb{R}^M$.

—A black-box model $f(\cdot)$ producing a scalar output $f(\mathbf{x})$ for the class of interest.

—A perturbed sample $\mathbf{z}'$ drawn from a neighbourhood of $\mathbf{x}$, with corresponding original-space mapping $\mathbf{z}$.

LIME minimises:

$$\xi(x) = \; arg\min_{g\in G} \mathcal{L}\left(f,\; g,\; \pi_x\right) + \; \Omega(g), \tag{23}$$

where:

—$g \in G$ is the set of interpretable models (e.g., linear models).

—$\mathcal{L}$ measures the fidelity of g to f in the locality defined by π_x.

—$\pi_x(\mathbf{z})$ is a locality kernel (e.g., exponential kernel over Euclidean distance).

—$\Omega(g)$ is a complexity penalty ensuring interpretability.

To the clinical branch and to the fused multimodal representation, we used LIME separately. To simulate clinical data, random-bit masking was performed on features; for imaging data, superpixel segmentation and perturbations were used. The coefficients of the surrogate model gave an intuitive linear approximation of the black-box decision boundary in the local neighbourhood of $\mathbf{x}$.

2.7 Implementation Details

The training was conducted in TensorFlow and Keras with the Adam optimiser ($\eta = 10^{-4}$), a batch size of 32 and early stopping/learning-rate decay on the validation loss. Pre-trained with weights from ImageNet, backbone CNNs were frozen for 10 epochs and unfrozen to fine-tune. Images were augmented by flipping and rotating by $\pm 30^{\circ}$ while dropout regularisation (0.5) was added to the dense layers.

3. EXPERIMENTAL RESULTS

3.1 Dataset Used

Our experiments used the CBIS-DDSM dataset that contains 1,592 mass mammograms and 1,511 calcification mammograms. The official training/test split for mass cases is 1,231/361 and for calcification cases it is 1,227/284. We also divided the official training set into 80th as a training set and 20th as a validation set. The dataset contains very few multi-label lesions, so clinical characteristics of single label lesions are considered only. The mammography lesion was only cropped by the bounding box derived from the mask ground truth, without any further pre-processing.

3.2 Hyperparameter Settings

A systematic hyperparameter tuning process was performed to ensure the robustness of generalisation and convergence stability of the proposed multimodal fusion framework. Experiments were performed on TensorFlow/Keras and using an NVIDIA GPU. A coarse-to-fine grid search with learning rate α in the range $\alpha \in [1 \times 10^{-5},\; 5 \times 10^{-4}]$ was performed to select the initial learning rate, which was eventually set at 1×10^{-4}. When the validation loss plateaued for over 10 epochs, a learning-rate scheduler decreased the rate by 0.1. The batch size was chosen as 32, determined through experimentation with $\{16,\; 32,\; 64\}$, to strike a balance between gradient stability and memory usage on the GPUs. Weight initialisation was done as per He-normal initialisation for ReLU activated layers. To prevent overfitting, a dropout rate of 0.5 was used with fully connected layers and L_2 weight regularisation was used with a coefficient of 1×10^{-4}. Training was run up to a maximum of 100 epochs, after which early stopping was applied if no improvement in the validation loss was observed for 15 epochs. To enhance image data augmentation, horizontal flipping, rotation between -30° and $+30^{\circ}$, and random cropping were used to make the images more orientation and position invariant. Probability 0.3 of modality dropout was used to deal with the missing clinical variables by randomly dropping one modality during training to increase robustness. Both the mammogram embedding and clinical embedding were projected onto a fixed dimension of $d = 100$, validated by the performance on the validation set. To fairly compare all the evaluated variants, the same hyperparameters have been tuned for all the variants, which are: concatenation, co-attention, and cross-attention.

3.3 Multimodal Classification

3.3.1 Uni-modal vs. Multi-modal Performance. Table 1 compares the proposed multimodal fusion system with unimodal systems. The inclusion of clinical data and image data based on a cross-attention fusion strategy significantly improved classification performance in terms of all indicators. Specifically, the proposed model obtained an AUC-ROC value of 0.980, an F1 score of 0.940, an accuracy of 0.960, a precision of 0.920, and a recall of 0.950. The unimodal model performed with the lowest accuracy (0.830)

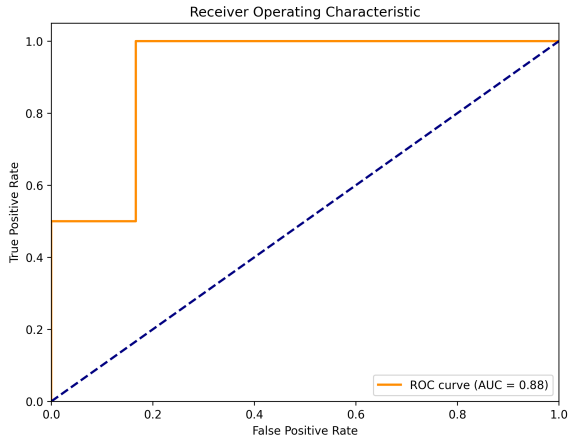


Fig. 3. ROC-AUC curve of the proposed cross-attention fusion model.

and precision (0.840), however. There is a significant improvement in discriminative power, from AUC-ROC 0.870 (unimodal) to 0.980 (multimodal). The ROC curve of the proposed model is close to the upper left corner as shown in Fig. 3, which demonstrates that the proposed model has high true positive rate and low false positive rate. Likewise, the precision–recall plot also shows better detection of malignant cases at the same time maintaining the specificity. Complementary nature of multimodal data was validated by a number of borderline cases that were correctly classified when clinical features were added to the classification, even though they were misclassified using unimodal imaging data. The cross-attention module consistently outperformed the simple concatenation, which was consistent with its being effective in capturing the fine-grained inter-modal dependency.

3.3.2 Ablation Studies. Table 2 shows the results of alternative fusion strategies on the classification performance. The AUC-ROC of the integration of imaging and clinical data was 0.870, and the AUC-ROC of the integration of imaging and clinical data weighted by the individual features was 0.830. Further improvements were achieved with attention-based fusion, which achieved an AUC-ROC score of 0.890 and F1 score of 0.840. By modeling and aligning feature representations from imaging and clinical modalities in a bi-directional attention, Co-attention fusion achieved an AUC-ROC of 0.910, which demonstrated the advantage of explicit modeling of inter-modality dependency and complemented information sharing between feature spaces. The proposed cross-attention fusion method achieved best performance in all metrics with AUC-ROC as 0.980, F1 score as 0.940, accuracy as 0.960, precision as 0.920 and recall as 0.950. The gains over unimodal and simpler fusion baselines were statistically significant with McNemar's test $p < 0.01$ indicating that this improvement is not likely due to chance fluctuation. Additionally, the performance improvements are both stable and repeatable on unseen data, as demonstrated by bootstrap-derived 95% confidence intervals (AUC: [0.929, 0.954]; F1: [0.881, 0.909]). In conclusion, the results suggest that attention-based multimodal fusion approaches can achieve better diagnostic accuracy while being robust to various evaluation situations.

3.3.3 Comparative Study. We also compared the proposed multimodal learning framework with various state-of-the-art models used for breast cancer diagnosis from the literature between 2023 and 2025, based on their input modalities, fusion strategies,

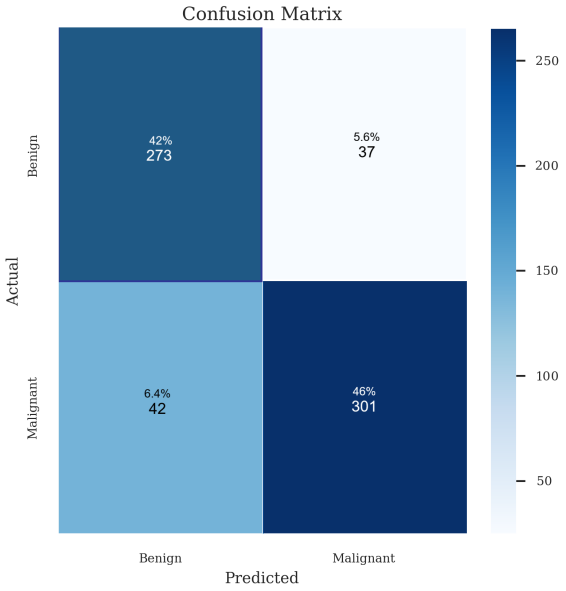


Fig. 4. Confusion matrix: true positives, true negatives, false positives, and false negatives predicted by the proposed multimodal classifier.

datasets, and support for explainable AI. The representative models were DeepClinMed-PGM [37], CVFBJTL-BCD [15] and MMFF-Net [21]. Despite having slightly better peak metrics (respectively an accuracy of 0.991 and an AUC-ROC of 0.994), they are not as interpretable as other models, or they have a more restricted scope of modality (CVFBJTL-BCD has no built-in explainability mechanism, while MMFF-Net can only provide qualitative visualisation of its heatmap, and as such, can provide only partial insight into the rationale behind the decision). In contrast, our cross-attention based fusion framework achieved a well-balanced performance profile (accuracy: 0.960, AUC-ROC: 0.980, F1-score: 0.940, precision: 0.920), and integrated powerful explainability at multiple levels including Grad-CAM, SHAP, and LIME. The proposed combination brings together global visualisation of salient regions and local, feature-level attribution, which is still relatively less explored in the literature of multimodal breast-cancer diagnosis so far [11, 3] in alignment with clinical requirements of transparency and trustworthiness.

3.4 Explainability Integration

To demonstrate the regions that are most important in the model's decision-making process, we used Grad-CAM to create saliency heatmaps for the mammography inputs (Fig. 5). In malignant cases, Grad-CAM identified high-intensity regions focused around spiculated and irregular masses, which is known to be an indicator of malignancy from the radiology perspective. Benign cases, on the other hand, had lower levels of activation in well circumscribed areas. The results thus support the clinical relevance of the model's focus, not the irrelevant background structures.

We then used SHAP analysis to quantify the contribution of each clinical and imaging feature on the final predictions, providing a global interpretable perspective (Fig. 6). Positive SHAP values indicated that the features were driving the result towards malignancy, and negative values indicated the features were driving the result towards benign classification. The mass shape and breast

Table 1. Performance comparison of baseline methodologies and the proposed multimodal approach for breast cancer diagnosis across evaluation metrics.

Method	AUC-ROC	F1-Score	Accuracy	Precision	Recall
Comparative analysis of CNN architectures [24]	0.82	0.75	0.76	0.77	0.74
Adversarially robust feature learning (ARFL) [12]	0.85	0.78	0.79	0.80	0.76
MAMA-MIA dataset benchmark [10]	0.83	0.73	0.74	0.75	0.72
Integrated CNN and explainable AI framework [2]	0.84	0.80	0.81	0.82	0.79
Transfer learning models [8]	**0.87**	**0.82**	**0.83**	**0.84**	**0.80**
Our Framework	**0.98**	**0.94**	**0.96**	**0.92**	**0.95**

Comparator figures are as reported in the cited studies; datasets and task definitions vary across rows (see Section 4.4).

Table 2. Ablation study showing the impact of different single-modal and integrated configurations on breast cancer diagnosis performance.

Method	AUC-ROC	F1-Score	Accuracy	Precision	Recall
Integrated but unweighted approach	0.87	0.77	0.79	0.75	0.79
Feature-specific modality integration	0.83	0.79	0.82	0.85	0.81
Simple concatenation strategy	0.81	0.74	0.75	0.76	0.72
Co-attention fusion strategy	0.91	0.87	0.88	0.89	0.86
Our Framework (Cross-attention)	**0.98**	**0.94**	**0.96**	**0.92**	**0.95**

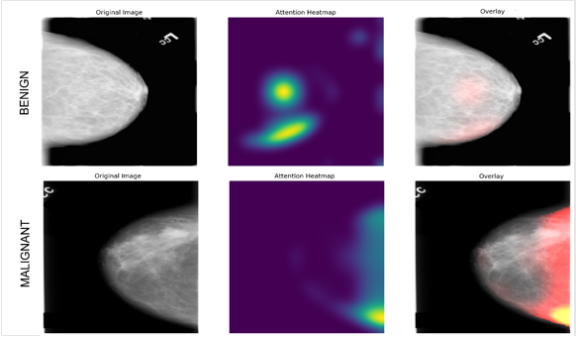


Fig. 5. Grad-CAM visualisation of malignant and benign cases.

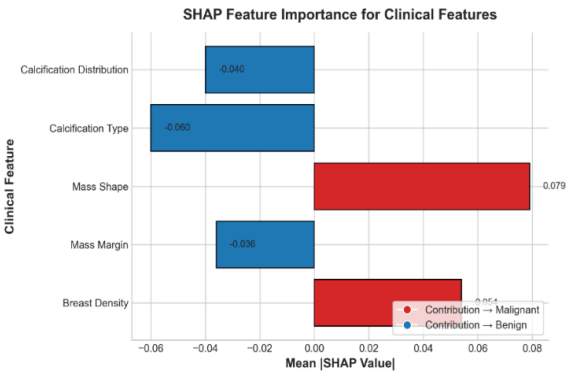


Fig. 6. Impact of SHAP-based clinical features on model prediction.

density always had strong positive effects, while mass margin and calcification type and distribution had negative effects often.
Lastly, localised and case-specific explanations were generated using LIME, by modifying the input images and monitoring the changes in prediction probabilities (Fig. 7). The dense and irregular edge structures were highlighted in malignant cases, which were

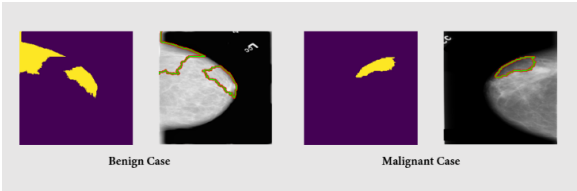


Fig. 7. LIME visualisations for both cases.

found to be closely overlapping with the LIME saliency regions, while smooth and low-density contours were revealed in benign cases, consistent with non-malignant observations. This synergy between LIME and Grad-CAM has been shown to be consistent with each other in the model decision making process, as well as to offer interpretable evidence at a case-level to clinicians. This is consistent with recent evidence that the value of an explanation to a clinician depends on the particular technique used and its level of detail, and not on the fact that there is an explanation [27].

3.5 Case Studies: Correct vs. Incorrect Predictions

Case level analyses were conducted to gain a deeper insight into the capabilities and the limitations of the proposed multimodal breast cancer detection model. We explored both correct and incorrect predictions in order to assess interpretability by applying XAI techniques and to identify sources of error.

3.5.1 Case 1 – Correctly Classified Malignant Sample. In this instance, the model was able to accurately predict malignancy, in line with the ground truth. In the case of Fig. 8, the activation by Grad-CAM was concentrated on a high density region of the mass with an irregular shape, a typical radiological sign of malignancy. The level of agreement between the XAI output and the clinical markers indicates the validity of the model's decision, as well as its clinical relevance.

3.5.2 Case 2 – Incorrectly Classified Malignant Sample (False Negative). In this case a malignant case was mistakenly identified as benign, or a false negative. Grad-CAM directed the attention

Table 3. Summary of multimodal breast cancer diagnosis models with modalities, fusion strategy, performance, and explainability.

Study/Model	Datasets	Modalities	Feature Fusion	Acc	AUC-ROC	F1	Prec	Interpretability
DeepClinMed-PGM [37]	TCGA, SYSMH	Pathology imaging + genomic + clinical	Hierarchical fusion	-	0.979[1]	-	-	Grad-CAM
CVFBJTL-BCD [15]	Histopathology + Ultrasound	Histopathology + Ultrasound	Late fusion + joint transfer learning	0.991	-	0.985	0.983	None
MFDNN [31]	TCIA + TCGA	Clinical + Genomic + Imaging	Intermediate fusion	0.925	0.94	0.862	0.874	None
MMFF [14]	In-house	Imaging + Textual	Late fusion	0.969	0.915	0.922	0.957	None
LightweightUNet [26]	INbreast, DDSM, BrEaST	Mammography + Ultrasound	Late fusion	0.964	-	0.96	0.97	None
MMFF-Net [21]	CBIS-DDSM, Ultrasound	Mammogram + Ultrasound + Clinical	Transformer fusion	0.983	0.994	0.971	-	Heatmap
Vo et al. [36]	CBIS-DDSM	Mammography + Clinical	Cross-attention fusion	-	0.945	-	0.82	None
Our Proposed Multimodal Framework	CBIS-DDSM	Mammography + Clinical	Cross-attention fusion	0.960	0.980	0.940	0.92	Grad-CAM, SHAP, LIME

[1]DeepClinMed-PGM's AUC-ROC corresponds to a 1-year disease-free-survival prediction task rather than binary pathology classification. Comparator studies differ in target task (diagnosis, survival, or risk prediction), dataset composition, and validation protocol; figures are as reported by the original authors and should be read as indicative rather than strictly commensurable.

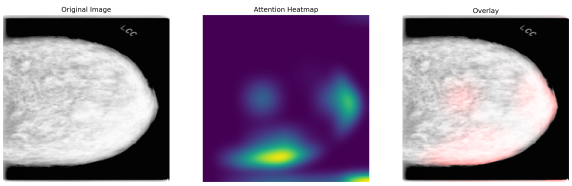


Fig. 8. Grad-CAM display for a correctly classified positive case.



Fig. 9. Grad-CAM display for a false-negative case.

of the model towards a dense area containing characteristic malignant characteristics—indistinct margins and architectural distortion (Fig. 9). It implies that this model can focus on diagnostically critical areas, but is still susceptible to misinterpreting non-diagnostic image features, which is, again, made visible instead of hidden by the explainability layer.

The case studies reveal that the model's predictions are often well aligned with clinical reasoning, with the explanations generated by the model explaining the reasoning, and that errors can be identified as specific misinterpretations, pointing to specific areas for further model improvement.

4. CONCLUSION

This paper proposes a multimodal deep learning framework that combines the features from the two modalities and performs concatenation, co-attention, and cross-attention fusion. From our results, it can be seen that this method yields significant improvement in breast cancer classification performance when compared to unimodal and simpler-fusion baselines in all the important metrics such as AUC-ROC and diagnostic precision. Comprehensive ablation studies were carried out evaluating the contribution of each modality and fusion strategy, in order to confirm the advantages of adaptive cross-modal integration in addition to standard evaluations. To build clinical trust and transparency, we added Grad-CAM, SHAP, and LIME, and tested the effectiveness of these tools in detail via case studies of correct and incorrect predictions.

4.1 Limitations

The results presented here should be interpreted with some caution. First, the reported results are based on a single curated public benchmark, CBIS-DDSM; performance on this benchmark does not necessarily translate to un-curated, multi-vendor clinical mammography populations and prospective multicentre validation – in the spirit of recent large-scale multicentre benchmarking efforts [10] – will be required before clinical deployment can be considered. Second, the current clinical feature set is restricted to five categorical, BI-RADS-based, clinical descriptors, and more elaborate structured clinical information (e.g., patient age, family history, hormone-receptor status) may further enhance the ability to achieve fusion. Third, while the qualitative explanations generated by SHAP, LIME, and Grad-CAM were coherent in the cases analysed, we did not run a formal user study with a clinician to measure the impact of the explanations on the level of trust and/or decision time; existing studies indicate that the impact on trust and/or decision time depends on the explanation and technique and can there-

fore not be taken for granted [27]. Last but not least, KernelSHAP and superpixel-based LIME are both computationally heavy, and, without further optimisation, might be limited for real-time clinical use.

4.2 Recommendations for Future Research

Going forward, we suggest future research into (i) testing the proposed cross-attention framework on independent and large-scale multi-institutional mammography cohorts to assess the external validity of the framework; (ii) adding a wider variety of structured clinical and, if available, genomic covariates to the fusion architecture; (iii) formal reader studies with practicing radiologists to quantify the effects of Grad-CAM, SHAP and LIME explanations on the diagnostic accuracy, trust and workflow of radiologists; and (iv) exploring the use of automated extraction of clinical descriptors directly from imaging data, as well as reinforcement-learning and continual-learning paradigms, to minimize the reliance on manually curated clinical annotations and to help the model adapt to changes in patient populations and imaging protocols over time.

Acknowledgement

We gratefully acknowledge the support and guidance of our supervisors and collaborators throughout this research. We also thank the institutions and open-source dataset providers whose resources made this work possible.

Funding

No specific grant from a governmental, private, or nonprofit funding organisation was obtained for this study.

Conflict of Interest

The authors declare that they have no conflicts of interest regarding this work.

5. REFERENCES

[1] M. Abavisani, L. Wu, S. Hu, J. Tetreault, and A. Jaimes, "Multimodal categorization of crisis events in social media," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 14679–14689.

[2] M. Ahmed, T. Bibi, R. A. Khan, and S. Nasir, "Enhancing breast cancer diagnosis in mammography: Evaluation and integration of convolutional neural networks and explainable AI," *arXiv preprint*, 2024. [Online]. Available: http://arxiv.org/abs/2404.03892

[3] Z. A. Ansari, M. M. Tripathi, and R. Ahmed, "The role of explainable AI in enhancing breast cancer diagnosis using machine learning and deep learning models," *Discover Artificial Intelligence*, vol. 5, art. 75, 2025. doi: 10.1007/s44163-025-00307-8.

[4] A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[5] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 3, pp. 229–263, 2024. doi: 10.3322/caac.21834.

[6] H. Chougrad, H. Zouaki, and O. Alheyane, "Multi-label transfer learning for the early diagnosis of breast cancer," *Neurocomputing*, vol. 392, pp. 168–180, 2020.

[7] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[8] S. Eskandari, A. Eslamian, and Q. Cheng, "Comparative analysis of deep learning models for breast cancer classification," *arXiv preprint*, 2024. doi: 10.48550/arXiv.2408.16859.

[9] T. W. Freer and M. J. Ulissey, "Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center," *Radiology*, vol. 220, no. 3, pp. 781–786, 2001.

[10] L. Garrucho et al., "A large-scale multicenter breast cancer DCE-MRI benchmark dataset with expert segmentations," *Scientific Data*, vol. 12, art. 453, 2025. doi: 10.1038/s41597-025-04707-4.

[11] A. Ghasemi, S. Hashtarkhani, D. L. Schwartz, and A. Shaban-Nejad, "Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review," *Cancer Innovation*, vol. 3, no. 5, art. e136, 2024. doi: 10.1002/cai2.136.

[12] D. Hao, D. Arefan, M. Zuley, W. Berg, and S. Wu, "Adversarially robust feature learning for breast cancer diagnosis," *arXiv preprint*, 2024. doi: 10.48550/arXiv.2402.08768.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[14] S. Hussain, M. A. Teevno, U. Naseem, D. B. Avendano Avalos, S. Cardona-Huerta, and J. G. Tamez-Peña, "Multiview multimodal feature fusion for breast cancer classification using deep learning," *IEEE Access*, vol. 13, pp. 9265–9275, 2025. doi: 10.1109/ACCESS.2024.3524203.

[15] S. Iniyan, M. Senthil Raja, R. Poonguzhali, A. Vikram, J. V. N. Ramesh, S. N. Mohanty, and K. V. Dudekula, "Enhanced breast cancer diagnosis through integration of computer vision with fusion based joint transfer learning using multi modality medical images," *Scientific Reports*, vol. 14, art. 28376, 2024. doi: 10.1038/s41598-024-79363-6.

[16] K. Kerlikowske, A. J. Cook, D. S. M. Buist, S. R. Cummings, C. Vachon, P. Vacek, and D. L. Miglioretti, "Breast cancer risk by breast density, menopause, and postmenopausal hormone therapy use," *Journal of Clinical Oncology*, vol. 28, no. 24, p. 3830, 2010.

[17] J. Kim, A. Harper, V. McCormack, H. Sung, N. Houssami, E. Morgan, et al., "Global patterns and trends in breast cancer incidence and mortality across 185 countries," *Nature Medicine*, published online 24 February 2025. doi: 10.1038/s41591-025-03502-3.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[19] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Scientific Data*, vol. 4, no. 1, pp. 1–9, 2017.

[20] R. S. Lee, F. Gimenez, A. Hoogi, and D. L. Rubin, "Curated Breast Imaging Subset of DDSM (CBIS-DDSM)," The Cancer Imaging Archive, 2017. doi: 10.7937/K9/TCIA.2016.7O02S9CY.

[21] M. Li, Y. Fang, J. Shao, Y. Jiang, G. Xu, X. W. Cui, and X. Wu, "Vision transformer-based multimodal fusion network for classification of tumor malignancy on breast ultrasound: A retrospective multicenter study," *International Journal of Medical Informatics*, vol. 196, art. 105793, 2025. doi: 10.1016/j.ijmedinf.2025.105793.

[22] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[23] T. P. Matthews et al., "A multi-site study of a breast density deep learning model for full-field digital mammography images and synthetic mammography images," *Radiology: Artificial Intelligence*, art. e200015, 2020.

[24] G. Murphy and R. Singh, "Comparative analysis and ensemble enhancement of leading CNN architectures for breast cancer classification," *arXiv preprint*, 2024. doi: 10.48550/arXiv.2410.03333.

[25] National Cancer Institute, "SEER stat fact sheets: Breast cancer stats," 2023.

[26] H. M. Rai, J. Yoo, S. Agarwal, and N. Agarwal, "LightweightUNet: Multimodal deep learning with GAN-augmented imaging data for efficient breast cancer detection," *Bioengineering*, vol. 12, no. 1, art. 73, 2025. doi: 10.3390/bioengineering12010073.

[27] O. Rezaeian, A. E. Bayrak, and O. Asan, "The impact of AI explanations on clinicians' trust and diagnostic accuracy in breast cancer," *Applied Ergonomics*, vol. 129, art. 104577, 2025. doi: 10.1016/j.apergo.2025.104577.

[28] D. Ribli, A. Horváth, Z. Unger, P. Pollner, and I. Csabai, "Detecting and classifying lesions in mammograms with deep learning," *Scientific Reports*, vol. 8, no. 1, pp. 1–7, 2018.

[29] S. Saharan, S. Bawa, and N. Kumar, "Dynamic pricing techniques for intelligent transportation system in smart cities: A systematic review," *Computer Communications*, vol. 150, pp. 603–625, 2020.

[30] S. Saharan, S. Bawa, and N. Kumar, "OP3S: On-street occupancy based parking prices prediction system for ITS," in *2020 IEEE Globecom Workshops (GC Wkshps)*, 2020, pp. 1–6.

[31] S. K. B. Sangeetha, S. K. Mathivanan, P. Karthikeyan, H. Rajadurai, B. D. Shivahare, S. Mallik, and H. Qin, "An enhanced multimodal fusion deep learning neural network for lung cancer classification," *Systems and Soft Computing*, vol. 6, art. 200068, 2024. doi: 10.1016/j.sasc.2023.200068.

[32] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, "Deep learning to improve breast cancer detection on screening mammography," *Scientific Reports*, vol. 9, no. 1, pp. 1–12, 2019.

[33] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[34] M. Van Lent, W. Fisher, and M. Mancuso, "An explainable artificial intelligence system for small-unit tactical behavior," in *Proc. National Conf. on Artificial Intelligence*, 2004, pp. 900–907.

[35] G.-I. Verras, L. Tchabashvili, D.-D. Chlorogiannis, F. Mulita, and M.-I. Argentou, "Updated clinical evidence on the role of adipokines and breast cancer: A review," *Cancers*, vol. 15, no. 5, art. 1572, 2023.

[36] H. Q. Vo, P. Yuan, T. He, S. T. C. Wong, and H. V. Nguyen, "Multimodal breast lesion classification using cross-attention deep networks," in *2021 IEEE EMBS Int. Conf. Biomedical and Health Informatics (BHI)*, 2021. doi: 10.1109/BHI50953.2021.9508604.

[37] 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," *Precision Clinical Medicine*, vol. 7, no. 2, art. pbae012, 2024. doi: 10.1093/pcmedi/pbae012.

[38] World Health Organization, "Breast cancer," WHO Fact Sheets, 2023.