

## Highlights

### **MultiBCD: A Multimodal model that simulates the human diagnostic process for automated Breast Cancer Detection**

JuntongDu,ZihanZhang,WeiyangTao

- Introducing MultiBCD, a multimodal model that emulates the human diagnostic process for the automated detection of breast cancer.
- Integrating an image classification framework with the capabilities of GPT-4 to enhance diagnostic decision-making.
- Demonstrating that MultiBCD achieves superior diagnostic accuracy and efficiency.
- Highlighting interpretability and aligning with the intuitive understanding prevalent in medical consultations.
- Emphasizing the notable practical utility of MultiBCD in the context of improving breast cancer detection methodologies.

# MultiBCD: A Multimodal model that simulates the human diagnostic process for automated Breast Cancer Detection

JuntongDu<sup>a,1</sup>, ZihanZhang<sup>b,1</sup> and WeiyangTao<sup>a,\*,2</sup>

<sup>a</sup>Department of Breast Surgery, First Affiliated Hospital of Harbin Medical University, Harbin, China

<sup>b</sup>Social Computing and Information Retrieval, Harbin Institute of Technology (HIT-SCIR), Harbin, China

## ARTICLE INFO

### Keywords:

Artificial Intelligence  
Multimodality  
Breast Cancer  
Mammogram  
CNN  
GPT-4

## ABSTRACT

To enhance the accuracy of breast cancer detection, our study introduces MultiBCD, a multimodal model that emulates the human diagnostic process for breast cancer detection. Integrating an image classifier with GPT-4, it evaluates mammographic images alongside patient complaints. The model's dual-head autoencoder efficiently processes image data, eliminating the need for manual lesion delineation, while GPT-4 extracts critical information from patient narratives.

MultiBCD demonstrates superior diagnostic accuracy and efficiency, achieving an F1 score of 86.49% and a recall rate of 94.12%, which marks an improvement over traditional methods. Furthermore, its design, emphasizing interpretability, aligns with the intuitive experience of medical consultations. The encouraging results of MultiBCD in breast cancer detection indicate its potential for application in similar diagnostic contexts.

The MultiBCD model is characterized by its compact structure, flexible and efficient coupling, and the open-sourcing of its code (<https://github.com/zhangzihan-is-good/AI-breast-cancer>), thereby enhancing the model's practical utility.

## 1. Introduction

Up to now, breast cancer has become the disease with the highest incidence rate among female malignant tumors, and the number of new cases has been increasing year after year, and at the same time, breast cancer is one of the most common cancers leading to women's mortalities [1, 2]. There are many high-risk elements for breast cancer [3, 4], including family and genetic history of breast cancer, age factor, early menarche or late menopause, early childbearing age or first delivery age greater than 30 years old, obesity, unhealthy dietary habits such as alcoholism and high-fat, and so forth [5, 6, 7].

Breast cancer is a heterogeneous disease with a heterogeneous pathogenesis [8]. The molecular typing of breast cancer varies [9] in accordance with the levels of hormone receptor (HR) and human epidermal growth factor receptor 2 (HER2) being expressed by breast cancer, as well as the diversity of therapeutic regimens [10], and the tendency for drug resistance to emerge during the treatment phase [11], which makes it a formidable challenge to cure breast cancer, to boost the survival rate of the patients, and to mitigate the rate of recurrence. Therefore, effective prevention and screening of breast cancer is of paramount importance.

Early clinical manifestations of breast cancer are atypical and need to be diagnosed with the support of imaging,

such as mammography (MG), ultrasound and magnetic resonance imaging (MRI) [12, 13]. Among them, mammography has become the primary method [14] of selection for breast screening in imaging due to its characteristics of non-invasiveness [15], convenience [16], high sensitivity [17], low radiation dose [18] and low price. Mammography can pick up microscopic foci of calcification, and is very sensitive to clusters of microcalcified spots [19].

In addition, Mammography is also very instrumental in identifying the benign and malignant nature of lumps [20]. Mammography can detect breast lumps that are beyond the reach of the doctor, which can lead to the early detection of breast cancer, thus enhancing diagnostic rate and cure rate [21, 19, 22]. The diagnostic sensitivity for breast cancer is 82% - 89% and the specificity is over 87% [23]. But while ensuring the accuracy of the machine, human interpretation and judgment of the mammogram has also become an influential factor in the breast cancer screening rate. Mammograms usually need to be read by two imaging physicians to ensure the accuracy of the judgments and to improve sensitivity [24, 25], but this ensures that the number of doctors is sufficient. Thus, the advent of artificial intelligence can be thought of as a secondary validation of the molybdenum radio-graphs, replacing one of the imaging physicians in screening for breast cancer with humans for computer-assisted detection (CAD) [26, 27].

Numerous studies utilizing mammographic imaging for breast cancer prediction have laid the groundwork for this research [28, 29, 30, 31]. However, most prior studies focused on a single modality, employing computer vision models to diagnose based on mammographic images, ultrasound, etc. [32, 33, 34, 35]. However, patient complaints are equally important. Consequently, we have developed a Multimodal

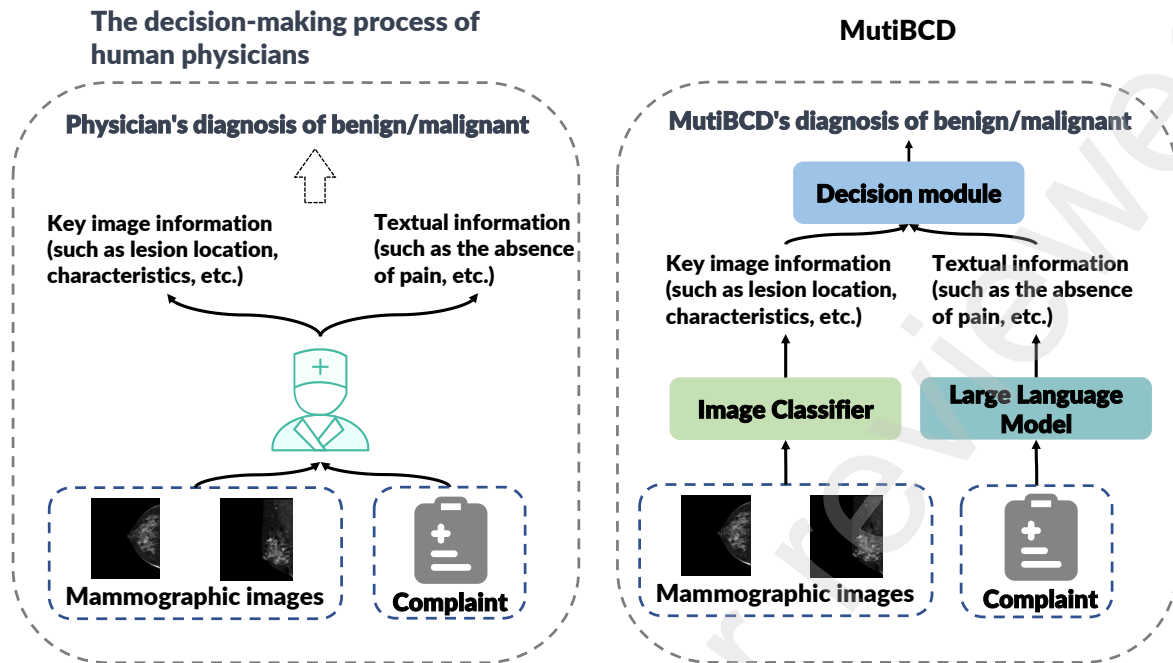
\*Corresponding author

✉ 13936610448@163.com ( JuntongDu); zihanzhang@ir.hit.edu.cn ( ZihanZhang); twysci@163.com ( WeiyangTao)

🌐 <https://github.com/DU0122> ( JuntongDu);

<https://github.com/zhangzihan-is-good> ( ZihanZhang)

ORCID(s): 0009-0001-4381-5280 ( JuntongDu); 0009-0006-7530-9127 ( ZihanZhang); 0000-0003-1107-5975 ( WeiyangTao)



**Figure 1:** Comparison of the diagnostic processes of human physicians and MultiBCD.

model that simulates the human diagnostic process for automated Breast Cancer Detection (MultiBCD). This approach not only involves the design of an innovative computer vision model for identifying features in mammographic images but also incorporate the advanced natural language model, GPT-4 [36], into breast cancer diagnosis. GPT-4, a sophisticated model trained on extensive datasets, utilizes advanced deep learning techniques to produce human-like responses to user inputs. In this study, GPT-4 is employed to process patient complaints, assisting diagnosis by extracting key information from these complaints.

Regarding the computer vision models that handle mammographic images, the better-performing models traditionally require manual demarcation of lesion locations, followed by benign-malignant classification based on radiographic features, or involved image segmentation using annotated data [37, 38, 39, 40]. This approach reduces the difficulty of learning lesion features from mammographic images but increases the workload for imaging physicians. Our study addresses this issue with a modified autoencoder structure capable of extracting features directly from mammographic images. These features are then channel-concatenated and undergo further processing and classification, enhancing diagnostic accuracy without necessitating manual lesion marking.

Specifically, the contributions of this study are as follows:

1. MultiBCD incorporates the GPT-4 into breast cancer diagnosis and utilizes key information from patient

complaints calculated by GPT-4 and Gradient Boosting Trees [41]. Our study aids mammographic image diagnosis and effectively harnesses multimodal information.

2. Automatically locating lesions in raw images has traditionally been challenging. Our study designs a model based on autoencoders with varying sizes of encoders and decoders, effectively extracting key features and pinpointing lesion locations.
3. Our model not only demonstrates superior performance, achieving an F1 score of 86.49% and a recall rate of 94.12%, approaching the performance of human imaging physicians, but also simulates the human diagnostic process, aligning with patients' intuitive experience of medical consultations and offering enhanced interpretability.
4. All data adheres to open-source protocols, and all codes have been made available<sup>1</sup> under the MIT license to facilitate replication of our results and to promote the development of future integrated systems aiding medical diagnostics. Additionally, MultiBCD is compact, with a flexible and efficient coupling, enhancing the practical applicability of the model.

<sup>1</sup><https://github.com/zhangzihan-is-good/AI-breast-cancer>

**Table 1**  
CMMD

Dataset	Benign	Malignant	Total
Training Set	2336	6864	9203
Validation Set	288	864	1152
Test Set	240	912	1152
Total	2864	8640	11504

**Table 2**  
Chinese Breast Disease Clinical Imaging Database

Dataset	Benign	Malignant	Total
Training Set	24	37	61
Test Set	10	17	27
Total	34	54	88

## 2. Materials and methods

### 2.1. Datasets

In our study, we employ two datasets, each serving a distinct purpose: The Chinese Mammography Database (CMMD) is used to train image classifiers. The Chinese Breast Disease Clinical Imaging Database is used to train and validate the final decision module. These datasets are described in more detail as follows:

- The Chinese Mammography Database (CMMD):**  
 This database was conducted on 1,775 patients from China with benign or malignant breast disease who underwent mammography examination between July 2012 and January 2016. The database consists of 3,728 mammographies from these 1,775 patients, with biopsy-confirmed types of benign or malignant tumors. For 749 of these patients (1,498 mammographies), the database also includes patients' molecular subtypes. Image data were acquired on a GE Senographe DS mammography system. The data can be obtained from this link.<sup>2</sup> We performed data augmentation and dataset partitioning on the original dataset, the processed dataset's statistical information is presented in Table 1.
- Chinese Breast Disease Clinical Imaging Database:**  
 This database includes 176 mammographic images and 84 corresponding patient complaints from 84 female breast disease patients. The data can be obtained from this link.<sup>3</sup> The dataset's information can be found in Table 2.

### 2.2. Problem formulation

The problem involves analyzing a patient's mammographic images, including the craniocaudal (CC) and mediolateral oblique (MLO) views, in conjunction with their complaints, to ascertain the presence of a malignant tumor.

<sup>2</sup><https://wiki.cancerimagingarchive.net>

<sup>3</sup><https://medbooks.ipmph.com/yx/imageLibrary/2578.html>

Define the collection of mammographic images as  $I = \{(c_i, m_i) \mid 1 \leq i \leq N\}$ , where  $c_i$  and  $m_i$  represent the craniocaudal and mediolateral oblique views, respectively. Let  $T = \{t_i \mid 1 \leq i \leq N\}$  denote the set of patient complaints, with  $N$  being the total number of patients. The mathematical model aims to optimize the function:

$$\max \sum_{i=1}^N \left[ (1 - y_i) \cdot P(y'_i = 0 \mid c_i, m_i, t_i) + y_i \cdot P(y'_i = 1 \mid c_i, m_i, t_i) \right] \quad (1)$$

where  $y_i$  and  $y'_i$  are Boolean values indicating the presence or absence of cancer in the patient.

### 2.3. Overview of MultiBCD

The diagnostic workflow of the MultiBCD model effectively replicates the process of human clinical consultation. The structural design of the MultiBCD model and its comparison to the human diagnostic process are illustrated in Figure 1.

The MultiBCD framework is comprised of three primary components: the image classifier, the LLM, and the decision-making module, as illustrated in Figure 2.

The image classifier extracts key features from mammographic images, including from the craniocaudal (CC) and mediolateral oblique (MLO) positions, such as lesion locations and characteristics. Concurrently, a large language model (LLM) gleans essential textual information beneficial for disease diagnosis from patient complaints, including symptoms like pain and nipple discharge. The decision-making module integrates information from both these modalities, enabling diagnosis of the patient's condition.

The MultiBCD model dexterously combines traditional machine learning algorithms with modern LLMs, offering not only the high performance characteristic of LLMs but also the interpretability essential for practical application in medical AI.

### 2.4. Image classifier module

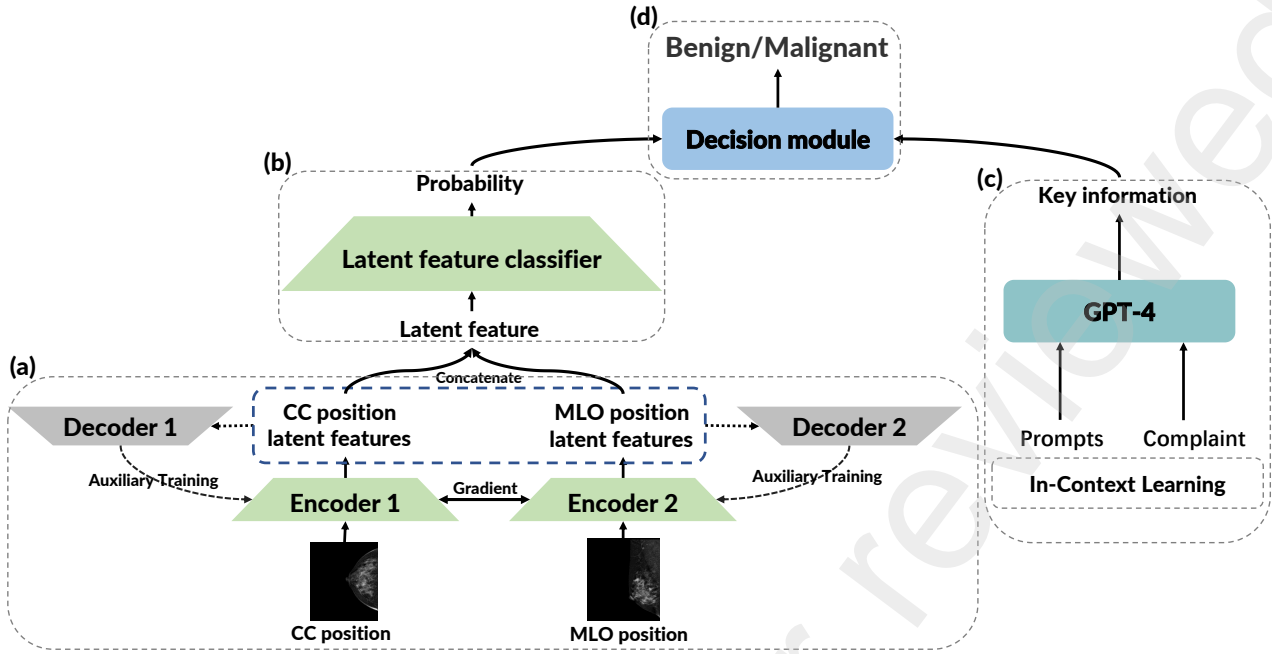
The image classifier within the MultiBCD framework is bifurcated into two key components: a dual-head autoencoder with shared parameters and a dual-head latent feature classifier. Notably, the training phase of the autoencoder also serves as the pretraining stage for the latent feature classifier.

#### 2.4.1. Dual-head autoencoder

The architecture of the dual-head autoencoder is depicted in the Figure 2 (a). In this structure, the CC position mammographic images are first encoded into CC position latent features by Encoder 1, and then subsequently decoded back to their original CC position images by Decoder 1. Similarly, the MLO position images undergo an analogous process: they are encoded into MLO position latent features and then decoded back to their original images.

This process can be formally described as follows:

$$Z_{ci} = \text{Encoder}_1(c_i) \quad (2)$$



**Figure 2:** Overview of MultiBCD. (a) Dual-head autoencoder, a part of Image classifier module, performs unsupervised learning of latent features in CC and MLO positions. (b) Dual-head latent feature classifier, also a part of Image classifier module, conducts preliminary probability predictions for benign and malignant classifications. (c) Complaint key information extraction module, extracts key information using GPT-4 and Prompts. (d) Final decision module, simulate human physicians: integrates various indicators for patient condition assessment.

$$c'_i = \text{Decoder}_1(Z_{ci}) \quad (3)$$

$$Z_{mi} = \text{Encoder}_1(m_i) \quad (4)$$

$$m'_i = \text{Decoder}_1(Z_{mi}) \quad (5)$$

The image reconstruction loss in the autoencoder is calculated using a combination of Mean Squared Error (MSE) and Structural Similarity Index Measure (SSIM). This can be represented as:

$$\text{Loss}_c = \sum_{i=1}^n (\alpha \times \text{MSE}(c_i, c'_i)) + (\beta \times \text{SSIM}(c_i, c'_i)) \quad (6)$$

$$\text{Loss}_m = \sum_{i=1}^n (\alpha \times \text{MSE}(m_i, m'_i)) + (\beta \times \text{SSIM}(m_i, m'_i)) \quad (7)$$

Considering the similarity in feature extraction methods for mammographic images at different positions, additionally, to mitigate overfitting, both encoders and decoders in our model employed a uniform parameter update strategy, as opposed to other methods that rely on weighted moving averages.

Furthermore, this encoder-decoder architecture capitalizes on the principles of compressed sensing. The dimensionality of the latent features is compressed to just 7.45% of

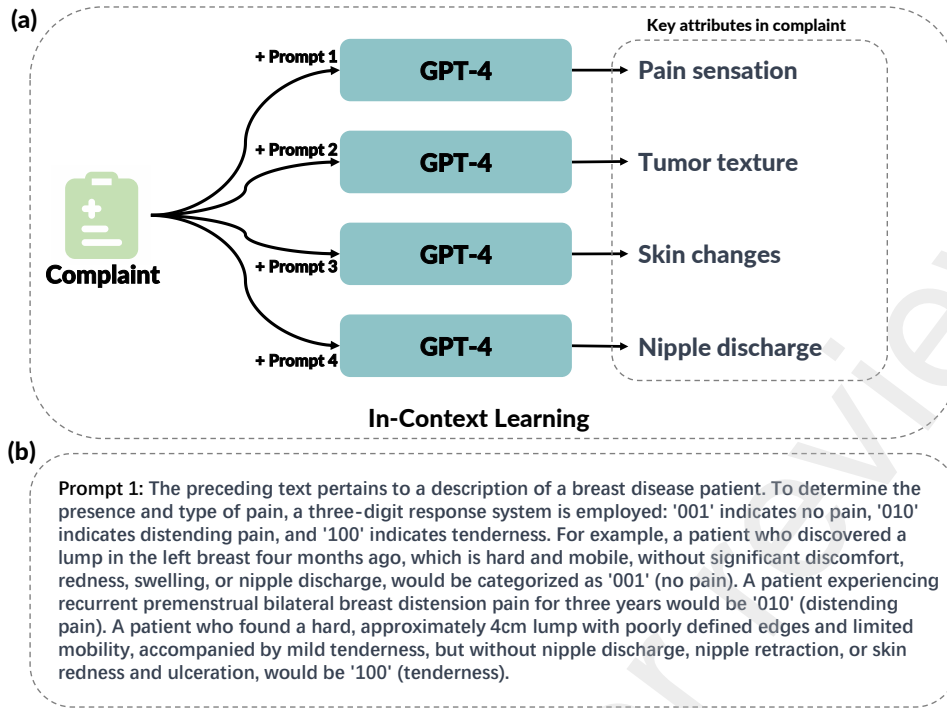
the original data size, compelling the model to distill more critical information from the image content. From an information theory perspective, the benign or malignant symbolizations of a lesion are effectively a subset of the information that the image can convey. This image reconstruction step essentially leverages unsupervised information beyond this subset.

#### 2.4.2. Dual-head latent feature classifier

The Latent Feature Classifier module, illustrated in the Figure 2 (b), utilizes latent vectors obtained from the encoder to further classify latent features. This process involves concatenating the CC and MLO position latent features along the channel dimension, which then serve as inputs to the classifier. The operation can be formalized as:

$$p_i = \text{Classifier}(Z_{ci} \oplus Z_{mi}), \quad (8)$$

where  $p_i$  denotes the preliminary probability assessment for the benignity or malignancy of a tumor based solely on imaging modalities, signifying a comprehensive summary of the image information. The patient's biopsy pathology results are used as the gold standard  $y_i$  for this assessment. The classifier's loss function is the cross-entropy loss, calculated as:



**Figure 3:** Examples of In-Context Learning: (a) An illustration of how GPT-4 extracts key information from complaint. (b) A sample of Prompt 1, with subsequent prompts following a similar methodology.

$$\text{Loss}_{\text{classifier}} = - \sum_{i=1}^n [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)], \quad (9)$$

The Dual-head autoencoder serves as pretraining step for the Dual-head latent feature classifier. In this phase, the encoder acquires mammographic image comprehension capabilities. Consequently, the overall parameters of the model are more closely aligned with the global optimum compared to those of an untrained model. This alignment significantly facilitates convergence to superior outcomes in classification tasks.

## 2.5. Complaint key information extraction module

The efficacy of the information extraction module hinges on the strategic application of well-crafted prompts and In-Context Learning (ICL) techniques to guide GPT-4 in extracting key information from patients' narratives. Specifically, multiple ingenious prompts were meticulously designed to leverage GPT-4's formidable logical reasoning capabilities. These prompts facilitate the retrieval of pertinent information for disease diagnosis from raw complaint corpora. The process can be formalized as:

$$\text{Info}_i = \text{GPT-4}(\text{Prompts}(\text{Info}_1, t_1; \text{Info}_2, t_2; \text{Info}_3, t_3), t_i), \quad (10)$$

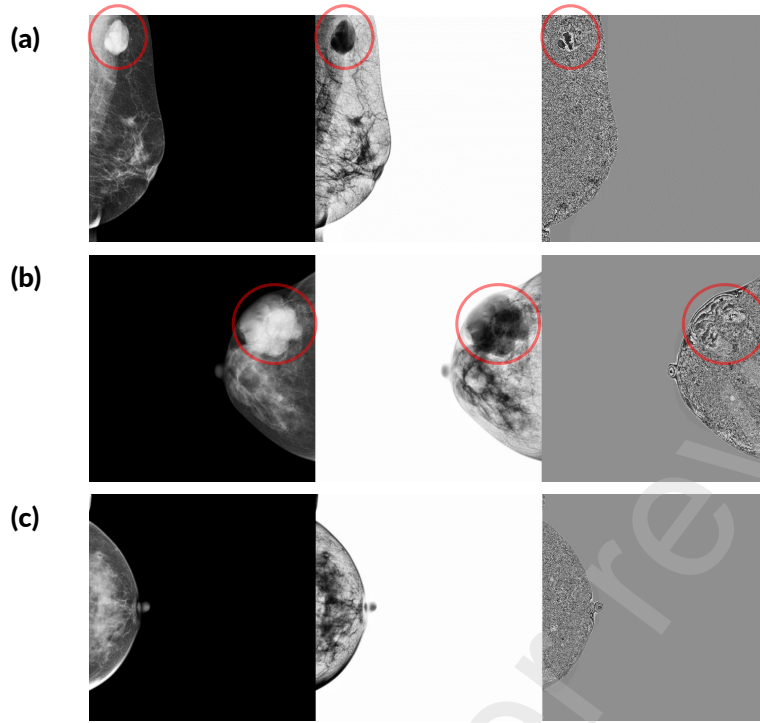
In this context,  $\text{Info}_i$  is defined as a quadruple, encapsulating four key attributes crucial for the diagnosis of breast cancer.

Figure 3 (a) illustrates the specific workflow of the Complaint Key Information Extraction module. In this process, the complaint is concatenated with various prompts and inputted into GPT-4 multiple times. This approach enables the extraction of features across different dimensions. Figure 3 (b) illustrates a specific example. We initiate with a description of the task type and objective as a part of prompt, followed by three concrete examples. This approach enables GPT-4 to extrapolate a general methodology for solving such problems based on the context of these examples.

Employing such prompts and ICL techniques ensures that the extraction of crucial information from patient narratives does not rely solely on the model's intrinsic medical knowledge. Instead, it transforms the understanding of patient complaints into a logical reasoning task. This strategy allows the effortless application of the model's logic reasoning skills, acquired in other domains, to this study without the necessity for extensive medical-related training data.

## 2.6. Final decision module

Human physicians' diagnosis of patient conditions often hinges on key features extracted from complaints and images. Decision models in Machine Learning not only effectively simulate this diagnostic process, but also have been empirically demonstrated to perform at a high level.



**Figure 4:** Three examples of mammographic images. (a) and (b) contain malignant tumors, while (c) represents a normal breast. The leftmost images are the original mammograms, the center images are the outputs from Stage 4, and the rightmost images are the outputs from Stage 6.

In our study, we have integrated the EXtreme Gradient Boosting tree (XGBoost) model. XGBoost synergizes the probabilities of benign/malignant conditions derived from images with useful diagnostic information extracted from complaints. This model discerns statistical patterns from both discrete and continuous features, thereby facilitating informed decision-making. The decision-making process can be formally represented as:

$$\max \sum_{i=1}^N \left[ (1 - y_i) \cdot P(y'_i = 0 \mid p_i, Info_i) + y_i \cdot P(y'_i = 1 \mid p_i, Info_i) \right] \quad (11)$$

In XGBoost, the relative importance of each factor in the decision-making process is transparent, thereby offering superior interpretability. Furthermore, patient complaints and mammographic images contain distinct types of information. For instance, a patient's pain cannot be discerned from mammographic images. Consequently, such decision models, in comparison to the currently popular multi-modal alignment models like CLIP, are more adept at effectively leveraging complementary information across different modalities.

### 3. Results

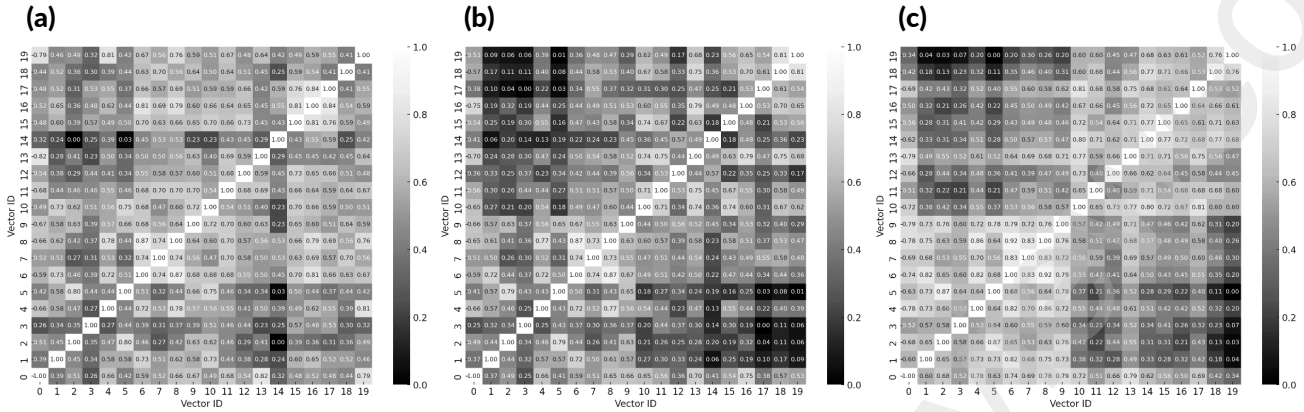
#### 3.1. Experiment settings and Performance of MultiBCD

##### 3.1.1. Image classifier module

In the dual-head autoencoder module, each encoder has a parameter count of 0.02M, while the decoder has a parameter count of 0.003M. This asymmetric parameter configuration endows the encoders with enhanced feature extraction capabilities.

During the training phase of the dual-head autoencoder module, which constitutes the first training stage of the Image Classifier module. Hyperparameter settings for the module included: an initial learning rate of 0.001 for the Adam optimizer, a random seed of 42, 20 epochs, a batch size of 8, and loss weights of  $\alpha = 1$  and  $\beta = -1$ .

In this study, the dataset underwent a series of preprocessing steps including downsampling, data augmentation, and contrast enhancement. Downsampling was employed to reduce the risk of model overfitting, while data augmentation was utilized to increase data diversity. Contrast enhancement, on the other hand, was applied to facilitate the extraction of features from grayscale images, thereby reducing the model's complexity in feature discernment. It emphasizes the detection of lesion locations. This approach mirrors the method used by human imaging physicians, who identify anomalies in density and morphology against the



**Figure 5:** Heatmap of the similarity between vectors. (a) The average similarity of  $Z_{ci}$  and  $Z_{mi}$  during the pretraining stage, (b) The similarity of  $p_i$  during the second stage of training, and (c) The similarity of the vectors from the penultimate hidden layer in the Latent Feature Classifier.

backdrop of normal breast tissue. The specific data processing procedure is outlined in Algorithm 1.

#### Algorithm 1 Image Reconstruction Process

**Require:** Original molybdenum target image

**Ensure:** Reconstructed image

**Stage 1:** Downsample a  $2294 \times 1914$  DICOM image to an 8-bit grayscale image of  $1144 \times 1144$ .

**Stage 2:** Apply 8x data augmentation (rotations at 0, 90, 180, and 270 degrees, along with their respective mirror images).

**Stage 3:** Perform CLAHE (Contrast Limited Adaptive Histogram Equalization) [42] for contrast enhancement.

**Stage 4:** Normalize the image data using the mean and variance calculated from the entire dataset, which were 22.05 and 42.99, respectively.

**Stage 5:** Encode the image into a latent vector representation using Encoder.

**Stage 6:** Decode the image using Decoder.

Three mammographic images were selected for illustration, as shown in the figure 4. Images (a) and (b) contain malignant tumors, while (c) represents a normal breast. The leftmost images are the original mammograms, the center images are the outputs from Stage 4, and the rightmost images are the outputs from Stage 6, which are the final results after processing by Algorithm 1. It is evident that the lesion features in (a) and (b) can be distinctly recognized. In contrast, the reconstructed image of (c), representing a normal breast, does not exhibit any notable abnormal features.

For the Latent Feature Classifier module, the overall architecture employs a modified ResNet34 model. This research adapted the first layer of ResNet34 to accommodate the output of the pretrained encoder.

In the second training phase of the Image Classifier module. The hyperparameters of the model are set as follows: The initial learning rate for the encoder part of the Adam optimizer is  $5e-4$ , and for the classifier part, it is  $1e-3$ . This specific setting is chosen because the encoder, having been pretrained, already exhibits parameters closer to the optimal. The random seed is set to 0, with an epoch count of 50 and a batch size of 4. The latent feature vectors are concatenated along the channel dimension, ensuring that the convolutional kernels processing the two vectors have distinct parameters.

Furthermore, in this research, 20 cases were selected from the dataset for input into the Image Classifier module, from which intermediate layer vectors were extracted. Among these, 10 benign cases were labeled 0-9, and 10 malignant cases were labeled 10-19. The similarity between each pair of these 20 vectors was calculated and illustrated in a heatmap, as shown in Figure 5. Specifically, (a) represents the average similarity of  $Z_{ci}$  and  $Z_{mi}$  during the pretraining stage, (b) depicts the similarity of  $p_i$  during the second stage of training, and (c) shows the similarity of the vectors from the penultimate hidden layer in the Latent Feature Classifier.

The method for calculating similarity is as follows:  $x_i$  and  $x_j$  represent different vectors, and  $k$  denotes the various dimensions.

$$d(x_i, x_j) = \sqrt{\sum_k (x_{ik} - x_{jk})^2} \quad (12)$$

$$D(x_i, x_j) = \frac{d(x_i, x_j)}{\max_{i,j} d(x_i, x_j)} \quad (13)$$

$$\text{Similarity} = 1 - D(x_i, x_j) \quad (14)$$

From the heatmap, it is evident that in Figures a, b, and c, the demarcation between benign and malignant cases is quite distinct. The similarity is higher among different benign cases and among different malignant cases, while it is lower between benign and malignant cases. During

**Table 3**  
Key information

Dimension	Attribute One	Attribute Two	Attribute Three
Pain sensation	No	Distending pain	Tenderness pain
Tumor texture	No	Tough	Hard
Skin changes	No	Yes	
Nipple discharge	No	Non-bloody fluid	Bloody fluid

the pretraining phase, the encoder extracted a variety of features beyond the benign-malignant distinction, aiding in image reconstruction, which resulted in a somewhat blurred demarcation between these classifications. After the second stage of training, this boundary became more pronounced. In the penultimate layer of the Latent Feature Classifier, the features distinguishing benign from malignant cases were further differentiated.

### 3.1.2. Complaint key information extraction module

Utilizing various prompts, GPT-4 is capable of extracting features from patient complaints across multiple dimensions. In this research, the number of dimensions was a priori set to four, namely pain sensation, tumor texture, skin changes, and nipple discharge. The specific configurations are detailed in the accompanying table 3.

Across four dimensions, the performance of GPT-4 closely mirrored that of human experts. Using human expert annotations as the gold standard, GPT-4 achieved accuracies of 98.86%, 100%, 100%, and 97.73%, respectively.

### 3.1.3. Final decision module

In this research, the decision model employs an XGBoost model with meticulously calibrated parameters, aiming to optimize performance and mitigate the risk of overfitting. Initially, multi-class log loss was chosen as the evaluation metric.

Regarding tree construction, a maximum depth of 4 was set, striking a balance between capturing key patterns in the data and preventing excessive complexity that could lead to overfitting. The learning rate was established at 0.01, a relatively low rate ensuring cautious adjustment of weights, thereby enhancing generalization. Furthermore, the construction of 200 trees was decided upon to facilitate thorough learning and pattern recognition. To further suppress overfitting, a random sampling technique was adopted, utilizing only 80% of the data samples in each iteration; additionally, only 20% of features were considered for splitting in each tree, increasing the model's randomness and robustness.

Lastly, the minimum loss reduction required to split a node was set to zero, indicating that even minor performance improvements would prompt further splitting of the tree.

Such configurations ensure a decision model that not only closely emulates human physicians' clinical decision-making but also exhibits robust performance.

**Table 4**

Performance of Different Models in Mammographic Image Diagnosis

Model	P	R	F1	Accuracy	AUC
MultiBCD	<b>82.29</b>	<b>89.24</b>	<b>85.62</b>	<b>76.28</b>	<b>0.63</b>
Resnet50 [43]	79.95	75.74	77.79	65.77	0.60
Googlenet [44]	80.12	86.72	83.29	72.46	0.58
VGG [45]	-	-	-	-	0.50
Alexnet [46]	-	-	-	-	0.50
VIT [47]	-	-	-	-	0.51

## 3.2. Method comparison and ablation study

### 3.2.1. Diagnosis based solely on imaging

The training of the Image Classifier module is divided into two stages. In the first stage, the Encoder is trained to autonomously learn image features through self-supervision. The second stage focuses on training the module to discern features pertinent to the benign or malignant nature of the tissue and to predict a probability associated with these characteristics. Diagnoses based solely on this probability have shown comparatively favorable classification results. To substantiate this conclusion, we also trained classical computer vision models with similar or more parameters on raw data without lesion contour annotations. The comparative results are presented in the accompanying table 4.

The outcomes reveal that, in the absence of manually delineated lesion boundaries, both VGG and AlexNet achieved an AUC (Area Under the Curve) of 0.5, indicating their ineffectiveness in extracting useful features from the original mammographic images. Additionally, the Vision Transformer (ViT) model struggled to develop classification capabilities during training, primarily due to the substantial differences between its original pretraining data and mammographic image features, and the unsuitability of ViT's architecture for such small-scale datasets.

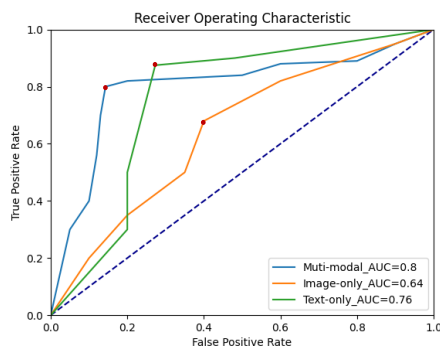
Compared to ResNet50 and GoogLeNet, our Image Classifier module exhibited superior performance in all aspects of classification.

During the second training phase of the Image Classifier module, different learning rates were assigned to the encoder and classifier. This decision was based on the fact that the encoder, having been trained in the first phase, already possessed some knowledge of feature extraction from mammographic images. However, a too high learning rate might lead to the forgetting of this pre-trained knowledge, while a too low rate could cause the model to focus excessively

**Table 5**  
Performance of MultiBCD Across Different Learning Rates

Learning Rate	P	R	F1	Accuracy	AUC
1e-3	80.08	88.69	84.17	73.59	0.57
5e-4	<b>82.29</b>	<b>89.24</b>	<b>85.62</b>	<b>76.28</b>	<b>0.63</b>
1e-4	80.66	83.32	81.97	70.98	0.60
1e-3 (No pretrain)	80.09	79.47	79.78	68.11	0.58

on irrelevant features. To validate this hypothesis, various learning rates were applied to the encoder during the second phase of training. Additionally, the impact of using a randomly initialized encoder (without a pretraining step) on image classification effectiveness is also presented, as detailed in the table 5.

**Figure 6:** Receiver Operating Characteristic(ROC) curves.

Consequently, it is evident that an appropriately balanced learning rate leads to improved outcomes.

### 3.2.2. Performance of the various components of the MultiBCD

MultiBCD is capable of processing diagnostic information from both textual and image modalities, each playing a significant role in the model's decision-making process. The table 6 illustrates the impact of text-only, image-only, and the combined text-image inputs on the model's decisions. Corresponding ROC curves are presented in the Figure 6. The values of Precision, Recall, and other metrics in Table 6 are derived from the confusion matrix at the classifier's optimal position. This optimal position is indicated by the red dot in Figure 6. The confusion matrix is detailed in Table 7.

The results indicate that the training preference of the Image Classifier module ensures a higher recall rate. The powerful inference capabilities of GPT-4, along with ingenious prompt design and the effective design of the Complaint Key Information Extraction module, enable the efficient extraction of diagnostically relevant information from complaints. The final dual-modality decision module effectively leverages the strengths of both elements.

**Table 6**  
Performance of Different Components of the MultiBCD and Human

Module	P	R	F1	Accuracy	AUC
Text-only	<b>87.50</b>	82.35	84.85	<b>81.48</b>	0.76
Image-only	68.18	88.24	76.92	66.67	0.64
Muti-modal	80.00	<b>94.12</b>	<b>86.49</b>	<b>81.48</b>	<b>0.80</b>
Human	<b>80.95</b>	<b>100.00</b>	<b>89.47</b>	<b>85.19</b>	–

**Table 7**  
Confusion Matrix of MultiBCD

Confusion matrix		Text-only		Image-only		Muti-modal	
TP	FN	14	3	15	2	16	1
FP	TN	2	8	7	3	4	6

## 4. Discuss

In tasks related to medical diagnosis, two major issues often lead to a conservative bias in models. First, datasets are frequently imbalanced, with malignant cases outnumbering benign ones. Second, in practical applications, the cost of misclassifying a malignant case as benign is significantly higher than misclassifying a benign case as malignant, and quantifying this cost is challenging. This scenario can lead to models erring on the side of caution by overly predicting malignancy, an outcome that is undesirable.

In such tasks, recall and precision are typically the most critical metrics. The final model that integrates both textual and image information, MultiBCD, achieved a commendable recall rate (94.12%) and precision (81.48%). Furthermore, an F1 score of 86.49% and an AUC of 0.80 indicate that the classifier does not resort to conservative, overly-negative classifications.

## 5. Conclusion

This study demonstrates how the integration of computer vision models with advanced large language models (LLMs) can enhance the accuracy and efficiency of breast cancer diagnosis. The MultiBCD model effectively addresses key challenges in breast cancer screening by combining an image classifier, a large language model, and a decision-making module. This multimodal approach extracts valuable diagnostic information from patients' mammographic images and complaints, offering a more precise method for breast cancer screening.

In terms of model design, we incorporate the GPT-4 into breast cancer diagnosis and employed a multi-stage training approach for the image classifier to locate and diagnose lesions from raw mammographic images. Additionally, the design of the MultiBCD model simulates the diagnostic process of human physicians, enhancing the model's interpretability. Experimental results indicate that MultiBCD excels in the automatic diagnosis of breast cancer.

The open-source code in this study not only facilitate replication and validation of our results by others but also promote the development of future integrated medical diagnostic systems. Furthermore, the compact and efficient coupling of MultiBCD enhances its practical applicability.

In summary, the results of this study provide a powerful tool for cancer screening, with the aim of achieving early diagnosis and treatment. It can provide a possibility to increase patient survival rates and reduce recurrence. Limited by the dataset, this study utilized two modalities of information: mammographic images and patient complaints. Future work could involve incorporating more modalities, such as breast ultrasound, and exploring the applicability of this method to other types of cancer diagnosis, as well as further enhancing diagnostic accuracy and efficiency.

## CRediT authorship contribution statement

**JuntongDu:** Conceptualization, Investigation, Methodology, Writing-Original draft preparation. **ZihanZhang:** Software, Writing-Original draft preparation. **WeiyangTao:** Writing-Reviewing; Validation; Supervision.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

We thank the reviewers for their constructive comments, and gratefully acknowledge the support of the First Affiliated Hospital of Harbin Medical University Fund for Distinguished Young Medical Scholars (2021J17) and the BEIJING MEDICAL AWARD FOUNDATION(YXJL-2021-0302-0287).

## References

- [1] Carol E DeSantis, Freddie Bray, Jacques Ferlay, Joannie Lortet-Tieulent, Benjamin O Anderson, and Ahmedin Jemal. International variation in female breast cancer incidence and mortality rates. *Cancer epidemiology, biomarkers & prevention*, 24(10):1495–1506, 2015.
- [2] Mahshid Ghoncheh, Zahra Pournamdar, and Hamid Salehiniya. Incidence and mortality and epidemiology of breast cancer in the world. *Asian Pacific journal of cancer prevention*, 17(S3):43–46, 2016.
- [3] Marzena Kamińska, Tomasz Ciszewski, Karolina Łopacka-Szatan, Paweł Miotła, and Elżbieta Starostawska. Breast cancer risk factors. *Menopause Review/Przegląd Menopauzalny*, 14(3):196–202, 2015.
- [4] Xiaohui Sun, Anne S Reiner, Anh Phong Tran, Gordon P Watt, Jung Hun Oh, Lene Mellemkjær, Charles F Lynch, Julia A Knight, Esther M John, Kathleen E Malone, et al. A genome-wide association study of contralateral breast cancer in the women's environmental cancer and radiation epidemiology study. *Breast Cancer Research*, 26(1):1–7, 2024.
- [5] Zohre Momenimovahed and Hamid Salehiniya. Epidemiological characteristics of and risk factors for breast cancer in the world. *Breast Cancer: Targets and Therapy*, pages 151–164, 2019.
- [6] Adam R Brentnall, Jack Cuzick, Diana SM Buist, and Erin J Aiello Bowles. Long-term accuracy of breast cancer risk assessment combining classic risk factors and breast density. *JAMA oncology*, 4(9):e180174–e180174, 2018.
- [7] Maria Escala-Garcia, Anna Morra, Sander Canisius, Jenny Chang-Claude, Siddhartha Kar, Wei Zheng, Stig E Bojesen, Doug Easton, Paul DP Pharoah, and Marjanka K Schmidt. Breast cancer risk factors and their effects on survival: a mendelian randomisation study. *BMC medicine*, 18:1–10, 2020.
- [8] Fabiana Löönd, Stefanie Tiede, and Gerhard Christofori. Breast cancer as an example of tumour heterogeneity and tumour cell plasticity during malignant progression. *British journal of cancer*, 125(2):164–175, 2021.
- [9] Gábor Cserni. Histological type and typing of breast carcinomas and the who classification changes over time. *Pathologica*, 112(1):25, 2020.
- [10] Biniyam G Demissei, Rebecca A Hubbard, Liyong Zhang, Amanda M Smith, Karyn Sheline, Caitlin McDonald, Vivek Narayan, Susan M Domchek, Angela DeMichele, Payal Shah, et al. Changes in cardiovascular biomarkers with breast cancer therapy and associations with cardiac dysfunction. *Journal of the American Heart Association*, 9(2):e014708, 2020.
- [11] Xingli Dong, Xupeng Bai, Jie Ni, Hao Zhang, Wei Duan, Peter Graham, and Yong Li. Exosomes and breast cancer drug resistance. *Cell death & disease*, 11(11):987, 2020.
- [12] Debra L Monticciolo, Mary S Newell, Linda Moy, Cindy S Lee, and Stamatia V Destounis. Breast cancer screening for women at higher-than-average risk: Updated recommendations from the acr. *Journal of the American College of Radiology*, 2023.
- [13] Nina Pötsch, Giulia Vatteroni, Paola Clauser, Thomas H Helbich, and Pascal AT Baltzer. Contrast-enhanced mammography versus contrast-enhanced breast mri: a systematic review and meta-analysis. *Radiology*, 305(1):94–103, 2022.
- [14] Stephen W Duffy, László Tabár, Amy Ming-Fang Yen, Peter B Dean, Robert A Smith, Håkan Jonsson, Sven Törnberg, Sam Li-Sheng Chen, Sherry Yueh-Hsia Chiu, Jean Ching-Yuan Fann, et al. Mammography screening reduces rates of advanced and fatal breast cancers: Results in 549,091 women. *Cancer*, 126(13):2971–2979, 2020.
- [15] Caixing Yuan, Guolin Xu, Xiangmei Zhan, Min Xie, Mingcong Luo, Lilan She, and Yunjing Xue. Molybdenum target mammography-based prediction model for metastasis of axillary sentinel lymph node in early-stage breast cancer. *Medicine*, 102(42):e35672, 2023.
- [16] Takayoshi Uematsu, Ayumi Izumori, and Woo Kyung Moon. Overcoming the limitations of screening mammography in japan and korea: a paradigm shift to personalized breast cancer screening based on ultrasonography. *Ultrasonography*, 42(4):508, 2023.
- [17] Marissa B Lawson, Savannah C Partridge, Daniel S Hippe, Habib Rahbar, Diana L Lam, Christoph I Lee, Kathryn P Lowry, John R Scheel, Sana Parsian, Isabella Li, et al. Comparative performance of contrast-enhanced mammography, abbreviated breast mri, and standard breast mri for breast cancer screening. *Radiology*, 308(2):e230576, 2023.
- [18] Nissren Tamam, H Salah, Mohammad Rabbaa, Mohammad Abduljoud, A Sulieman, M Alkhorayef, and DA Bradley. Evaluation of patients radiation dose during mammography imaging procedure. *Radiation Physics and Chemistry*, 188:109680, 2021.
- [19] Gaurav J Bansal, Lauren Emanuel, and Sesha Kanagasabai. Malignancy risk of indeterminate mammographic calcification in symptomatic breast clinics. *Postgraduate medical journal*, 99(1169):153–158, 2023.
- [20] Mindy L Yang, Chandni Bhimani, Robyn Roth, and Pauline Germaine. Contrast enhanced mammography: focus on frequently encountered benign and malignant diagnoses. *Cancer Imaging*, 23(1):10, 2023.
- [21] Muayad Sadik Croock, Saja Dhyaa Khuder, Ayad Esho Korial, and Sahar Salman Mahmood. Early detection of breast cancer using mammography images and software engineering process. *Telkomnika (Telecommunication Computing Electronics and Control)*, 18(4):

- 1784–1794, 2020.
- [22] Emanuele Neri, Vincenza Granata, Stefania Montemezzi, Paolo Belli, Daniela Bernardi, Beniamino Brancato, Francesca Caumo, Massimo Calabrese, Francesca Coppola, Elsa Cossu, et al. Structured reporting of x-ray mammography in the first diagnosis of breast cancer: A delphi consensus proposal. *La radiologia medica*, 127(5):471–483, 2022.
  - [23] Lei Yang, Shengfeng Wang, Liwen Zhang, Chao Sheng, Fengju Song, Ping Wang, and Yubei Huang. Performance of ultrasonography screening for breast cancer: a systematic review and meta-analysis. *BMC cancer*, 20:1–15, 2020.
  - [24] N Clerkin, CF Ski, PC Brennan, and Ruth Strudwick. Identification of factors associated with diagnostic performance variation in reporting of mammograms: a review. *Radiography*, 29(2):340–346, 2023.
  - [25] Rebecca Oudsema, Esther Hwang, Sharon Steinberger, Rowena Yip, and Laurie R Margolies. Screening mammography: guidelines versus clinical practice. *Journal of Breast Imaging*, 2(3):217–224, 2020.
  - [26] Jung Hyun Yoon, Kyungwha Han, Hee Jung Suh, Ji Hyun Youk, Si Eun Lee, and Eun-Kyung Kim. Artificial intelligence-based computer-assisted detection/diagnosis (ai-cad) for screening mammography: Outcomes of ai-cad in the mammographic interpretation workflow. *European Journal of Radiology Open*, 11:100509, 2023.
  - [27] Nada M Hassan, Safwat Hamad, and Khaled Mahar. Mammogram breast cancer cad systems for mass detection and classification: a review. *Multimedia Tools and Applications*, 81(14):20043–20075, 2022.
  - [28] SJS Gardezi, A Elazab, B Lei, and T Wang. Breast cancer detection and diagnosis using mammographic data: Systematic review. *Journal of Medical Internet Research*, 2019. URL <https://www.jmir.org/2019/7/e14464/>.
  - [29] Y Gu, W Xu, T Liu, X An, J Tian, H Ran, W Ren, et al. Ultrasound-based deep learning in the establishment of a breast lesion risk stratification system: a multicenter study. *European Radiology*, 2023. URL <https://link.springer.com/article/10.1007/s00330-022-09263-8>.
  - [30] I Sechopoulos, J Teuwen, and R Mann. Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art. *Seminars in Cancer Biology*, 2021. URL <https://www.sciencedirect.com/science/article/pii/S1044579X20301358>.
  - [31] RA Dar, M Rasool, and A Assad. Breast cancer detection using deep learning: Datasets, methods, and challenges ahead. *Computers in Biology and Medicine*, 2022. URL <https://www.sciencedirect.com/science/article/pii/S0010482522007818>.
  - [32] W Lotter, AR Diab, B Haslam, JG Kim, G Grisot, E Wu, et al. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nature Medicine*, 2021. URL <https://www.nature.com/articles/s41591-020-01174-9>.
  - [33] KU Rehman, J Li, Y Pei, A Yasin, S Ali, T Mahmood, et al. Computer vision-based microcalcification detection in digital mammograms using fully connected depthwise separable convolutional neural network. *Sensors*, 2021. URL <https://www.mdpi.com/1424-8220/21/14/4854>.
  - [34] B Gheflati and H Rivaz. Vision transformers for classification of breast ultrasound images. In *IEEE Conference of the IEEE Engineering in Medicine and Biology Society*, 2022. URL <https://ieeexplore.ieee.org/abstract/document/9871809/>.
  - [35] SM McKinney, M Sieniek, V Godbole, J Godwin, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577:89–94, 2020. URL <https://www.nature.com/articles/s41586-019-1799-6>.
  - [36] Brady Lund. A brief review of chatgpt: Its value and the underlying gpt technology. *University of North Texas. Project: ChatGPT and Its Impact on ...*, 2023. URL [https://www.researchgate.net/publication/366809571\\_A\\_Brief\\_Review\\_of\\_ChatGPT\\_Its\\_Value\\_and\\_the\\_Underlying\\_GPT\\_Technology](https://www.researchgate.net/publication/366809571_A_Brief_Review_of_ChatGPT_Its_Value_and_the_Underlying_GPT_Technology).
  - [37] W Ma, Y Zhao, Y Ji, X Guo, X Jian, P Liu, and S Wu. Breast cancer molecular subtype prediction by mammographic radiomic features. *Academic Radiology*, 2019. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8082943/>.
  - [38] K Dembrower, Y Liu, H Azizpour, M Eklund, K Smith, et al. Comparison of a deep learning risk score and standard mammographic density score for breast cancer risk prediction. *Radiology*, 2020. URL <https://pubs.rsna.org/doi/full/10.1148/radiol.2019190872>.
  - [39] AM Al-Hejri, RM Al-Tam, M Fazea, AH Sable, S Lee, et al. Etecadx: Ensemble self-attention transformer encoder for breast cancer diagnosis using full-field digital x-ray breast images. *Diagnostics*, 2022. URL <https://www.mdpi.com/2075-4418/13/1/89>.
  - [40] X Qu, H Lu, W Tang, S Wang, D Zheng, Y Hou, et al. A vgg attention vision transformer network for benign and malignant classification of breast ultrasound images. *Medical Physics*, 2022. URL <https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.15852>.
  - [41] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *Journal of Machine Learning Research*, 17:1–5, 2016.
  - [42] P. Singh, R. Mukundan, and R. De Ryke. Feature enhancement in medical ultrasound videos using contrast-limited adaptive histogram equalization. *Journal of Digital Imaging*, 33(1):213–230, 2020. doi: 10.1007/s10278-019-00211-5. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7064707/>.
  - [43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016.
  - [44] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2015.
  - [45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.
  - [46] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems (NeurIPS)*, 2012.
  - [47] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.