

Article

Automated Radiological Report Generation from Breast Ultrasound Images Using Vision and Language Transformers

Shaheen Khatoon *  and Azhar Mahmood 

Department of Computer Science and Digital Technologies, School of Architecture, Computing, and Engineering,
University of East London, London E16 2RD, UK; a.mahmood3@uel.ac.uk

* Correspondence: s.khatoon@uel.ac.uk

Abstract

Breast ultrasound imaging is widely used for the detection and characterization of breast abnormalities; however, generating detailed and consistent radiological reports remains a labor-intensive and subjective process. Recent advances in deep learning have demonstrated the potential of automated report generation systems to support clinical workflows, yet most existing approaches focus on chest X-ray imaging and rely on convolutional–recurrent architectures with limited capacity to model long-range dependencies and complex clinical semantics. In this work, we propose a multimodal Transformer-based framework for automatic breast ultrasound report generation that integrates visual and textual information through cross-attention mechanisms. The proposed architecture employs a Vision Transformer (ViT) to extract rich spatial and morphological features from ultrasound images. For textual embedding, pretrained language models (BERT, BioBERT, and GPT-2) are implemented in various encoder–decoder configurations to leverage both general linguistic knowledge and domain-specific biomedical semantics. A multimodal Transformer decoder is implemented to autoregressively generate diagnostic reports by jointly attending to visual features and contextualized textual embeddings. We conducted an extensive quantitative evaluation using standard report generation metrics, including BLEU, ROUGE-L, METEOR, and CIDEr, to assess lexical accuracy, semantic alignment, and clinical relevance. Experimental results demonstrate that BioBERT-based models consistently outperform general domain counterparts in clinical specificity, while GPT-2-based decoders improve linguistic fluency.

Keywords: automatic report generation; vision transformer; BioBERT; GPT-2; multimodal learning



Academic Editor: Gabriel Cristobal

Received: 15 December 2025

Revised: 29 January 2026

Accepted: 4 February 2026

Published: 6 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Breast cancer is one of the leading causes of cancer-related mortality among women worldwide, where early detection and accurate diagnosis play a crucial role in improving patient outcomes. Breast ultrasound imaging is widely used as a complementary modality to mammography due to its safety, cost-effectiveness, and ability to visualize dense breast tissue. However, interpreting ultrasound images and generating structured radiological reports remain highly dependent on expert radiologists, making the process time-consuming and prone to inter-observer variability.

Automatic medical report generation (ARG) from imaging data has recently gained significant attention with the advancement of deep learning and Transformer-based architectures [1]. Early approaches to medical image report generation largely adapted

convolutional–recurrent (CNN–RNN) architectures originally developed for image captioning tasks [2,3]. While such models demonstrated the feasibility of generating textual descriptions from medical images, they often struggled to capture long-range dependencies and complex clinical semantics from the radiological reports. To address these limitations, subsequent studies introduced visual attention mechanisms, enabling models to focus more on fine-grained image regions associated with pathological findings [4,5]. More recently, Transformer-based architectures have been adopted to better model global contextual relationships, significantly enhancing the coherence, clinical relevance, and overall quality of generated medical reports [6].

While substantial progress has been achieved in ARG, almost all research utilizes a chest X-ray dataset published by Indiana University [7], a benchmark dataset with 7470 chest X-rays images and 3955 reports, and MIMIC-CXR [8], a large dataset with 473,057 chest X-ray images and 227,835 reports. ARG for breast ultrasound images remains underexplored. This task is particularly challenging due to the low signal-to-noise ratio of ultrasound images, variability in lesion appearance, and the need for precise clinical language to describe lesion morphology, tissue composition, and diagnostic interpretation.

To address these challenges, we propose a Transformer-based multimodal architecture for automated breast ultrasound report generation that jointly models visual features from ultrasound images and semantic features from clinical text. The proposed architecture integrates a Vision Transformer (ViT) for capturing fine-grained spatial and morphological patterns from ultrasound images, biomedical language models (BioBERT) for encoding clinical semantics, and a multimodal Transformer decoder with cross-attention to generate coherent and clinically meaningful radiological reports. In addition, GPT-2 is explored as a generative decoder to assess the trade-off between linguistic fluency and biomedical specificity.

By explicitly aligning visual and textual embeddings and employing dual cross-attention mechanisms, the proposed framework enables the effective fusion of lesion appearance, tissue composition, and clinical terminology. This design allows the model to generate diagnostically relevant reports that reflect tissue composition, lesion shape, margin characteristics, and clinical interpretation to generate an output that closely resembles radiologist-style written reports. The main contributions of this work are summarized as follows:

1. **Vision Transformer for Visual Feature Encoding:** Vision Transformer (ViT) is used to extract rich spatial and contextual visual representations from breast ultrasound images, enabling the effective modeling of lesion morphology, tissue heterogeneity, and spatial relationships that are critical for diagnostic interpretation.
2. **Biomedical-Aware Text Encoding:** We integrated BERT and BioBERT as textual encoders to capture the general linguistic structure and domain-specific biomedical terminology.
3. **Dual Cross-Attention-Based Multimodal Fusion:** We have used a Transformer decoder with dual cross-attention over both visual and textual embeddings, allowing the model to simultaneously attend to ultrasound-derived visual cues and clinically meaningful textual context during report generation.
4. **Comprehensive Encoder–Decoder Combinations and Analysis:** We evaluated multiple encoder–decoder configurations (ViT–BERT, ViT–BioBERT, ViT–GPT-2, and hybrid variants) and evaluated them with word overlap metrics to analyze their impact on report quality, clinical specificity, and linguistic coherence.

2. Related Work

Recent advancements in radiology report generation have significantly progressed, particularly using encoder–decoder architecture. Most of the work is based on image cap-

tioning, which aims to generate a text description from a given image following encoder–decoder architecture [2–5,9–12]. Many of these approaches utilize a CNN-RNN structure, where the image encoder (typically a CNN) extracts visual features from images, and a decoder (either RNN-based or Transformer-based) generates the corresponding textual descriptions.

While early work adopted CNN-RNN architecture, the recent studies have shifted towards utilizing transfer learning, Transformer-based architectures, and Large Language Models (LLMs). For instance, ref. [13] proposes a method for improving radiology report generation by incorporating multi-grained abnormality prediction. Their model uses a pretrained ResNet to extract visual features from IU X-Ray [7] and MIMIC-CXR [8] datasets. A Medical Concept Aligner predicts anatomical concepts, and a Dual-Stream Decoder integrates both disease and anatomical features using multi-head attention. This approach yields a 29.8% improvement in natural language metrics and 34.2% improvement in clinical efficacy compared to previous methods. While effective in generating accurate diagnostic reports, the model's reliance on predefined concepts may limit its ability to handle diverse reporting styles and complex cases.

The study by Singh and Karimi [14] explored the IU X-Ray to generate detailed findings and impression sections in radiology reports. They employed a deep learning architecture that utilized Google's Inception-v3 as the backbone for image processing, coupled with GRU and LSTM recurrent models within an encoder–decoder framework to generate the corresponding report text. The use of recurrent models helped capture the sequential flow of text and maintained a coherent medical context. However, recurrent networks are known to face challenges in effectively handling long-range dependencies, which may have limited their ability to accurately characterize complex anomalies spread across extended sections of the radiology reports.

In breast cancer diagnostics, Nguyen and Theodorakopoulos [15] explored the development of a hybrid model to automatically generate findings, with a particular focus on extracting BI-RADS scores from radiology reports. This approach combined text classification and natural language generation and was trained and evaluated on a dataset of 552 Dutch breast cancer radiology reports. The researchers utilized an encoder–decoder attention model alongside a BI-RADS classifier, demonstrating significant accuracy gains and achieving a ROUGE-L F1 score of 0.515. However, the model's limitations lie in its applicability to generating breast ultrasound reports, as it processes written text rather than directly analyzing ultrasound images and the corresponding findings.

Tsai and Chu [16] used a Faster R-CNN for object detection and an LSTM for report generation on gallbladder ultrasound images; their approach highlighted the limitations of solely relying on image-derived features. The incorporation of a Weight Selection Mechanism was shown to improve diagnostic reliability, but the model's performance dropped when it relied exclusively on visual cues without leveraging prior medical knowledge. This underscores the importance of incorporating domain-specific expertise and contextual information to achieve robust and accurate report generation, particularly in the complex domain of medical imaging and diagnostics.

Himel and Chowdhury [17] introduced an encoder–decoder framework for automated breast lesion segmentation and classification. The model harnessed transfer learning techniques, leveraging pretrained ResNet50V2, NASNetLarge, and EfficientNetB7 architectures to generate a segmentation mask that highlights the lesion region and extract features relevant for lesion classification. The extracted features from multiple encoder layers were then fused using a dual-stage fusion approach to predict whether the segmented lesion was benign or malignant. When evaluated on the BUSI and Thammasat datasets, the model achieved an IoU of 70.82% and a Dice score of 80.56%, demonstrating its effectiveness in

this task. Although the work does not focus on report generation, but rather the integration of Transformer architecture, it offers an insightful way forward for visual feature extraction.

Similarly, Tao and Ma [18] proposed a memory-based semantic model for generating radiology reports. The model comprises four key components: an Image Feature Extractor, a Cross-modal Memory Bank, a Cross-modal Semantic Alignment Module, and a Report Generator. The Memory Bank is used to learn disease-related representations from radiology reports. The model extracts visual features from the input image using a pretrained ResNet50 and converts the corresponding report into word embeddings. These visual and textual representations are then used to retrieve relevant information from the Memory Bank, consolidating the features. The consolidated features, along with prior knowledge retrieved from the Memory Bank and semantically aligned, are used by the Report Generator to produce the final radiology report. The model was evaluated on an MIMIC-CXR and IU X-Ray dataset. The results were compared to several other state-of-the-art models using evaluation metrics such as BLEU-1 to BLEU-4, METEOR, ROUGE-L, and CIDEr, demonstrating the effectiveness of the memory-based approach for radiology report generation on both datasets.

Chaudhury and Sau [19] use Vision Transformers pretrained with BERT and Long Short-Term Memory networks to classify breast cancer images. The ViT architecture processes image patches sequentially, enabling the model to capture long contextual dependencies within the images. The visual features are then passed to an LSTM-based recurrent neural network for classification. The model achieved an impressive 99.2% accuracy on a dataset of histopathology images and also utilized the BUSI dataset of breast ultrasound images. The work demonstrates the suitability of ViTs for effectively extracting and leveraging visual information compared to traditional convolutional neural network-based architectures.

Alqahtani and Mohsan [6] introduced CNX-B2, which combines a ResNet with a Transformer architecture. Specifically, they used a modified ConvNeXt as the image encoder to capture meaningful spatial features due to inherent convolutional biases. The visual features are then fed into a BioBERT Transformer, serving as the decoder to generate the radiological report text. Their evaluation of the IU X-Ray dataset achieved 0.479 scores for BLEU-1, 0.188 for METEOR, and 0.586 for CIDEr. The authors highlight that, while CNX-B2 achieves a competitive performance, its computational complexity presents a limitation, suggesting an area for future research in developing more lightweight approaches.

Singh and Singh [20] proposed ChestX-Transcribe, a multimodal Transformer model designed for automated radiology report generation. They used a pretrained Swin Transformer for extracting high-resolution visual features from the X-ray images and DistilGPT for generating textual reports. Evaluation on an IU X-Ray dataset demonstrated that ChestX-Transcribe outperforms previous models in generating clinically meaningful reports. While their work focuses on chest X-rays, the underlying multimodal Transformer architecture offers valuable insights for adapting similar methodologies to other medical imaging modalities, such as a breast ultrasound.

Recently, LLMs are used for radiology report generation and show promising results due to their strengths in generating syntactically and semantically coherent text. For instance, Wang and Liu [21] proposed R2GenGPT with a Swin Transformer as the Visual Encoder and the Llama2-7B as the text generator. The evaluation of MIMIC-CXR and IU X-Ray datasets achieved improved results on BLEU, METEOR, and ROUGE-L metrics compared to some existing models after extensive GPU training. However its performance on ensemble-based metrics like CIDEr lagged behind models like METransformer [22] and R2Gen [23]. The authors also acknowledge limitations related to R2GenGPT's reliance on human annotations and its lack of ensemble approaches.

Raminedi and Shridevi [24] proposed a multimodal Transformer architecture by using ViT, BEiT, and DEiT for image feature extraction and GPT-2 for report generation. The integration of a cross-attention mechanism between visual and textual features enables the model to capture complex relationships and generate more contextually relevant reports. The model is validated on an IU X-Ray dataset, and the results show improved medical image analysis and reporting—dependent on high-quality images and extensive computational resources.

The above literature highlights several research gaps. First, the majority of existing studies are evaluated on chest X-ray datasets (MIMIC-CXR and IU X-Ray) [13,14,18,20,21,24], while there is very limited work on breast cancer-specific datasets [15,16]. Second, earlier methods that rely heavily on CNN-RNN or CNN-Transformer architectures as the encoder and decoder achieve lower performance, indicating limitations in capturing long-range dependencies and complex visual–textual relationships.

Although recent Transformer-based and LLM-based approaches (e.g., Swin Transformer–DistilGPT and R2GenGPT) [20,21] show improved results, many still depend on predefined medical concepts [13] or single-modal features [16], limiting their ability to handle complex cases. While Vision Transformers and multimodal Transformers have demonstrated improved results in extracting robust visual features and aligning them with textual representations [19,24], their application to breast ultrasound report generation remains largely unexplored. This reflects the absence of standardized benchmarks and comprehensive evaluations for ultrasound-based report generation.

The proposed ViT–BioBERT–Transformer model, evaluated on the BrEaST (breast ultrasound dataset), demonstrates a strong performance across all metrics (see Section 4). It addresses these gaps by combining visual feature extraction with domain-specific textual modeling, generating accurate and clinically coherent reports.

3. Materials and Methods

The proposed architecture for generating a radiological report consists of three major components: (a) a Visual Encoder to capture the abnormalities in the ultrasound image and process them into trainable visual features, (b) a text encoder to extract findings from medical reports and process them into trainable textual embeddings, and (c) a Transformer-based decoder to train on the textual embeddings of the report with the visual features to generate comprehensive radiological reports.

This design is inspired by earlier Transformer-based approaches that have been successfully applied to chest X-ray report generation (see Section 2), where models pretrained on medical data consistently improved the quality of generated reports [20,21,24]. Breast ultrasound images present unique challenges, such as low image contrast and unclear lesion boundaries. As a result, clinical interpretation relies more on understanding the overall lesion shape, margins, and spatial relationships rather than fine visual details. To capture this broader visual context, ViT is used, as it has proven results in capturing long-range spatial patterns within an image [19].

In addition, the BrEaST dataset [25] adopted in this work contains short and structured radiological findings that include highly specialized medical terminology. To accurately capture the clinical language, pretrained language models such as BERT and BioBERT are used. These models are well adopted in the medical domain for understanding medical text and provide rich contextual representations [6]. The visual and textual features are then combined by the Transformer decoder to generate step by step radiological reports similar to the decoder adopted for X-ray chest datasets [24].

The complete architecture is illustrated in Figure 1, and the following sections describe each component in detail.

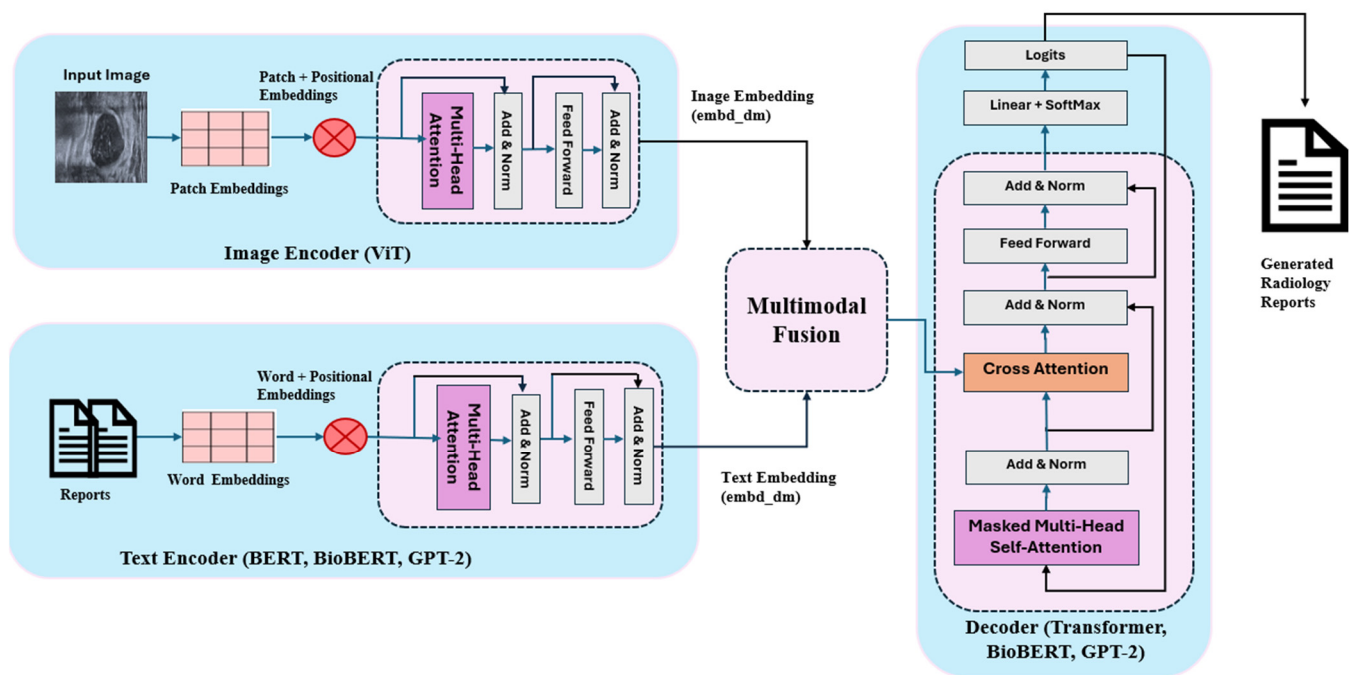


Figure 1. Abstract architecture of the proposed system.

3.1. Visual Encoder

In this study, visual feature extraction is conducted using a ViT, which represents an input ultrasound image $I \in \mathbb{R}^{H \times W \times C}$ as a sequence of non-overlapping patches $\{x_1, x_2, \dots, x_n\}$, $x_i \in \mathbb{R}^{P \times P \times C}$. The image is first divided into $N = \frac{HW}{P^2}$ patches of size $P \times P$, each flattened into a vector and projected onto a D-dimensional latent space using a learnable linear projection:

$$z_i^{(0)} = W_{px_i} + b_p; W_p \in \mathbb{R}^{D \times (P^2 \times C)} \quad (1)$$

Since Transformer do not encode spatial or sequential order, positional embeddings are added to each patch to preserve spatial information such as top-left vs. bottom-right, etc., to maintain the anatomical and spatial coherence necessary for clinical interpretation.

$$\tilde{z}_i^{(0)} = z_i^{(0)} + E_{pos}, E_{pos} \in \mathbb{R}^{(N+1) \times D} \quad (2)$$

To form a complete input, a learnable classification token embedding $z_{cls}^{(0)}$ is prepended:

$$Z^{(0)} = [z_{cls}^{(0)}; \tilde{z}_1^{(0)}; \tilde{z}_2^{(0)}; \dots; \tilde{z}_N^{(0)}] \quad (3)$$

The resulting sequence is passed through L stacked Transformer encoder layers (with $L = 12$). Each encoder layer consists of multi-head self-attention (MSA), followed by a feed-forward multilayer perceptron (MLP). At encoder layer l , the query, key, and value matrices are computed to generate self-attention weights:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V; Q = Z^{(l-1)}W_Q, K = Z^{(l-1)}W_K, V = Z^{(l-1)}W_V \quad (4)$$

where W_Q , W_K , and W_V are learnable projection matrices, and d_k is the dimensionality of the key vectors. Each encoder layer applies residual connections and layer normalization as follows:

$$\tilde{Z}^{(l)} = \text{MSA} \left(\text{LN}(Z^{(l-1)}) \right) + Z^{(l-1)}; l = 1, \dots, L \quad (5)$$

$$Z^{(l)} = +\text{MLP} \left(\tilde{Z}^{(l)} \right) + \tilde{Z}^{(l)}; l = 1, \dots, L \quad (6)$$

After passing through 12 layers of self-attention and *MLP*, the *CLS* token representation $z_{cls}^{(L)}$ is extracted as the contextualized visual embedding and normalized as:

$$V = \text{LN} \left(z_{cls}^{(L)} \right) \in \mathbb{R}^d \quad (7)$$

which serves as the input to the cross-modal Transformer for alignment with textual features. Using ViT with patch size 16×16 and removing its classification head allows the model to retain rich spatial and contextual information from ultrasound images while producing a compact feature representation for multimodal understanding.

Before feeding ViT features into the decoder, we apply $\hat{V} = W_v V, W_v \in \mathbb{R}^{768 \times 512}$, which matches the visual embedding dimension with the textual embedding space (512), enabling seamless fusion.

3.2. Textual Encoder (BERT, BioBERT, and GPT2)

For textual feature extraction, we used three large pretrained language models, BERT-base [26], BioBERT [27] and GPT-2, to generate contextual word embeddings that capture both the general semantic meaning and domain-specific biomedical knowledge from the accompanying clinical text. Both BERT and BioBERT share the same underlying Transformer architecture; however, BioBERT benefits from additional pretraining on large biomedical corpora such as PubMed abstracts and PMC full-text articles, enabling it to better encode medical terminology and clinical expressions. All models are used in their base configuration and pass through a similar sequence of blocks as follows:

Each input report is first tokenized into a subword, and special boundary tokens [CLS] and [SEP] are appended to denote the start and end of the sequence. The resulting token sequence is then converted into a combined embedding representation consisting of token embeddings, positional embeddings, and segment embeddings, ensuring the model retains both lexical and structural information.

$$H_i^{(0)} = [e_{\text{CLS}}, e_{w_1}, \dots, e_{w_M}, e_{\text{SEP}}]; e_i = e_i^{\text{token}} + e_i^{\text{pos}} + e_i^{\text{segment}}; i = 1, \dots, M+2 \quad (8)$$

The combined embedding sequence ($T = \text{CLS}, w_1, \dots, w_M, \text{SEP}$) is then passed through 12 Transformer encoder layers, each of which applies multi-head self-attention to contextualize every token with respect to the entire sentence. At each layer, multi-head self-attention is applied to contextualize each token based on every other token in the sequence, similar to ViT followed by feed-forward networks, residual connections, and layer normalization to produce deep contextual representations.

BERT-base, pretrained on BookCorpus and Wikipedia, provides strong general language understanding, while BioBERT's biomedical pretraining enables superior modeling of clinical concepts, lesion descriptors, and diagnostic phrasing essential for ultrasound report generation.

The final hidden states of these models represent each token in a 768-dimensional semantic space. To align the textual representation with the visual embedding dimension used by the decoder, a learnable projection layer maps the 768-dimensional output to a unified 512-dimensional multimodal space. This ensures that both BERT and BioBERT embeddings can be seamlessly integrated with visual features during cross-attention,

enabling the system to combine medical language understanding with image-derived visual cues during report generation.

$$H^{(L)} = [h_1, h_2, \dots, h_{M+2}], h_i \in \mathbb{R}^{768}; \tilde{H}^{(L)} = W_t H^{(L)}, W_t \in \mathbb{R}^{768 \times 512} \quad (9)$$

This unified embedding representation enables seamless cross-modal attention between image-derived visual features and clinically meaningful textual representations during report generation.

Unlike BERT and BioBERT, GPT-2 employs a unidirectional Transformer decoder architecture and is pretrained using an autoregressive language modeling objective. GPT-2 processes text sequentially from left to right and does not utilize segment embeddings or a [CLS] token. This architectural difference makes GPT-2 particularly suitable for natural language generation, as it explicitly models word-by-word dependencies during decoding. In this study, GPT-2 is used either as a standalone text encoder–decoder or as a decoder in combination with a ViT or BioBERT encoder. While GPT-2 produces fluent and coherent textual outputs, it lacks the bidirectional contextual encoding and biomedical pretraining of BioBERT, which can lead to reduced clinical specificity in generated reports.

3.3. Multimodal Fusion Through Transformer Decoder with Cross-Attention

The multimodal Transformer decoder is the generative component of the proposed architecture. Its primary role is to produce the final breast ultrasound diagnostic report autoregressively, integrating three critical sources of information: (1) previously generated textual tokens, (2) visual features extracted by the ViT, and (3) textual embeddings extracted by BERT or BioBERT.

The multimodal decoder is implemented as a stack of twelve Transformer decoder layers operating in a common 512-dimensional space, generating the breast ultrasound report autoregressively. At each decoding timestep t , the decoder receives all sequences of the previously generated tokens $y_{<t}$ and converts them into continuous vector representations. Each token is embedded through a learnable word embedding matrix W_y and positional encodings E_{pos} , which form the initial decoder representation.

$$D_t = W_y y_{<t} + E_{pos} \quad (10)$$

After the combination of the embedding vector along with positional encoding, the embedded sequence passed through various linear layers (masked multi-head self-attention) to generate queries, keys, and values, $Q = D_t W_Q$, $K = D_t W_K$, and $V = D_t W_V$, where D_t is the input sequence and W_Q , W_K , and W_V are learnable weight matrices. The masked self-attention module produces a contextualized representation of the partial report, where high attention indicates more relevance to the current token calculated, using

$$\hat{D}_t = \text{softMax} \left(\frac{Q \cdot K^T}{\sqrt{d_k}} + M \right) \cdot V \quad (11)$$

where $\hat{D}_t$ is the contextualized representation of a partial report, M is the casual mask preventing attention to future positions, and $\sqrt{d_k}$ is the dimension of the key vectors. In each decoder block, residual connection and layer normalization produce the updated decoder state to stabilize the model training, update the sequence weights, and normalize inputs into each layer.

After masked self-attention, each decoder layer performs cross-attention over both visual and textual features extracted from encoders. For visual features $\hat{V} \in \mathbb{R}^{N_v \times 512}$, in the cross-attention, the decoder uses its own queries $\hat{D}_t W_Q$ and build key (W_K), values (W_V) from visual tokens ($\hat{v}$):

$$C_t^{vis} = softmax \left(\frac{\hat{D}_t W_Q (\tilde{v} W_k)^T}{\sqrt{d_k}} \right) (\tilde{v} W_v) \quad (12)$$

This mechanism enables the decoder to focus on specific spatial regions and visual patterns in the ultrasound image relevant to the sentence currently being generated.

Similarly, textual cross-attention allows the decoder to reference clinical phrasing, terminology, and contextual cues encoded from the original report text. For textual features $\hat{T} \in \mathbb{R}^{N_t \times 512}$ the cross-attention is computed similarly:

$$C_t^{text} = softmax \left(\frac{\hat{D}_t W_Q (\hat{T} W_k)^T}{\sqrt{d_k}} \right) (\hat{T} W_v) \quad (13)$$

This dual attention mechanism ensures that the decoder can simultaneously use the learned medical semantics in the text encoder and the morphological/structural information present in the ultrasound image.

The visual and textual representations are concatenated using a fusion mechanism. In this design, both cross-modal outputs are concatenated and passed through a learnable fusion matrix (W_f), which aligns the modalities and compresses them into a unified hidden representation:

$$F_t = W_f (C_t^{vis}; C_t^{text}) \quad (14)$$

This stage integrates visual clues about breast tissue structure, biomedical semantics from the text encoder, and contextual language information from previous tokens.

After the contextual representation of visual and textual vectors, the decoder block applies a position-wise feed-forward network with a series of fully connected layers, adding nonlinear transformations and capturing the relationship between visual patterns (abnormalities in image) and textual semantics:

$$Z_t = FFN(F_t) \quad (15)$$

This helps the model capture the higher-order relationship between textual and visual features, improving semantic coherence (e.g., relating “irregular margins” to “malignant mass”). After passing through the full stack of decoder layers, the final hidden state is projected to vocabulary logits (W_o), and the next token is predicted by a softmax distribution:

$$P(y_t | y_{<t}, I, T) = softmax(W_o Z_t) \quad (16)$$

The probability distribution over the vocabulary determines the next output token. The beam search is applied to select the best token in the sequence and fed back into the decoder for the next timestep. This process continues until an end-of-sentence token is produced, completing the diagnostic report.

Besides the base-decoder, we used BioBERT and GPT-2 as the decoders in different combinations of encoder–decoder architectures. The details are given in Section 4.

4. Experimental Results

4.1. Dataset Description

This work uses the recently released Breast-Lesions-USG (BrEaST) dataset [25], which contains 256 breast ultrasound scans along with radiological findings collected in Poland from 2019 to 2022. Each scan was manually annotated and labeled by an experienced radiologist specializing in breast ultrasound examinations. Of the cases, 154 were classified as benign, 98 as malignant, and 4 as normal. For each benign and malignant case, lesions in the images were segmented, resulting in corresponding mask files added to the dataset.

In cases with multiple lesions in one ultrasound image, separate mask files were created, totaling 266 mask images. Tumors were marked using freehand annotations, and specific features were categorized according to BI-RADS, where a higher BI-RADS category reflects a greater risk of malignancy. Clinical data, confirmed through follow-up care or biopsy results, were also included in the dataset. Table 1 shows attribute descriptions and examples of clinical data released along with the images.

Table 1. Dataset attributes with description.

Attributes	Description	Example 1	Example 2
CaseID	unique identifier	4	46
Image_filename	filename of ultrasound b-mode image(s)	case004.png	case046.png
Mask_tumor_filename	filename(s) of binary mask overlaying a tumor	case004_tumor.png	case046_tumor.png
Mask_other_filename	filename(s) of binary mask overlaying other suspicious lesions in the image (e.g., cysts)	-	-
Pixel_size	pixel size of the image in centimeters	0.007812	0.009197324
Age	age of patient	43	73
Tissue_composition	breast tissue composition assessed during examination	homogeneous: fibroglandular	homogeneous: fat
Signs	objective observable abnormalities perceived by physician	no	no
Symptoms	subjective experiences reported by patient	no	not available
Shape	BI-RADS descriptor	round	irregular
Margin		circumscribed	not circumscribed—indistinct
Echogenicity		hypoechoic	hypoechoic
Posterior_features		no	no
Halo		no	yes
Calcifications		no	no
Skin_thickening		no	no
Interpretation	image interpretation for mass differentiation	duct filled with thick fluid & intraductal papilloma	suspicion of malignancy
BIRADS	BI-RADS category	4a	5
Verification	confirmed method for mass classification (follow-up care or biopsy)	confirmed by biopsy	confirmed by biopsy
Diagnosis	histological type of mass (for biopsied patients)	usual ductal hyperplasia (UDH) & pseudoangiomatous stromal hyperplasia (PASH)	invasive carcinoma of no special type (NST)
Classification	lesion type (benign or malignant)	benign	malignant

The dataset has primarily been used in the literature for breast lesion classification, benign vs malignant discrimination [28–34], and lesion segmentation [35–38], using various deep learning approaches. Existing studies mainly focus on image-level diagnosis or lesion-level prediction tasks. To the best of our knowledge, none of the existing works have explored this dataset for automated radiological report generation or multimodal image–text learning. In contrast, this work uses both ultrasound images and associated radiological findings to generate structured diagnostic reports, thereby extending the utility of the BrEaST dataset beyond cancer detection and lesion prediction.

4.2. Data Preprocessing

In the preprocessing, both the image and radiological findings were preprocessed to extract the relevant features for multi-model learning. At first, input images are resized to 224×224 pixels to match with the input dimensions of ViT. The dataset also consists of clinical reports for each image with key elements such as *Shape*, *Tissue Composition*, *Interpretation*, etc., as shown in Table 1. To generate radiological reports from the original findings, we prepared a summary of key elements by concatenating *tissue composition*, *shape*, and *interpretation* columns using the template: “The tissue is [Tissue_composition], and the lesion shows a [Shape] shape. Interpretation indicates [Interpretation].” These columns are natural language descriptions of important findings from the radiology image, and the summary serves as a ground truth. Different preprocessing techniques such as text normalization, converting the text to lowercase, and removing specified punctuation and special characters are applied on the ground truths.

As shown in Figure 2a, the resulting ground truth reports have an average length of 152 characters, with a vocabulary dominated by domain-specific clinical terms related to tissue characteristics (e.g., *homogeneous fibroglandular*, *predominantly fatty*), lesion morphology (e.g., *round*, *oval*, and *irregular*), and diagnostic interpretation (e.g., *cyst*, *malignancy*, *fibroadenoma*, and *intraductal papilloma*), as illustrated in the word cloud (see Figure 2b). This analysis is important for the model design, particularly for determining an appropriate maximum token sequence length for Transformer-based text encoders and decoders. Although the average report length is moderate, a maximum sequence length of 1024 tokens was adopted to ensure that all reports can be processed without truncation. In practice, report lengths can vary, and more detailed clinical descriptions may include additional contextual or interpretive information. Furthermore, subword tokenization used in Transformer-based models often increases the number of tokens compared to the original word count, especially for specialized radiological terminology. Using a higher token limit therefore preserves the full semantic content of clinical reports during both the training and inference stage. The word cloud further highlights the importance of accurately capturing specialized medical terminology, supporting the use of pretrained domain-specific language models (BioBERT) to accurately represent medical vocabulary and generate clinically coherent reports.

After applying basic preprocessing, each image was associated with its corresponding ground truths for training models that can learn to understand visual content and generate reports based on visual data. Figure 3 shows a preprocessed breast ultrasound image with its corresponding ground truth for sample images (case 004 and case 046).

4.3. Evaluation Metrics

To evaluate the ultrasound report generation capability of Transformer-based models' natural language generation metrics such as BELU [39], ROUGE-L [40], METEOR [41], and CIDEr [42] are used. BELU, originally developed for assessing the quality of machine translation outputs, has been widely adopted in various text generation tasks, including automatic report generation from medical images. It measures the precision of n-gram overlaps between the generated report and the reference report (ground truth), computed as:

$$BLEU = BP \times \exp\left(\frac{1}{N} \sum_{n=1}^N \log(p_n)\right) \quad (17)$$

where BP is the brevity penalty used to discourage excessively short outputs, N is the mixed n-grams order, and p_n is the precision of n-grams of length n .

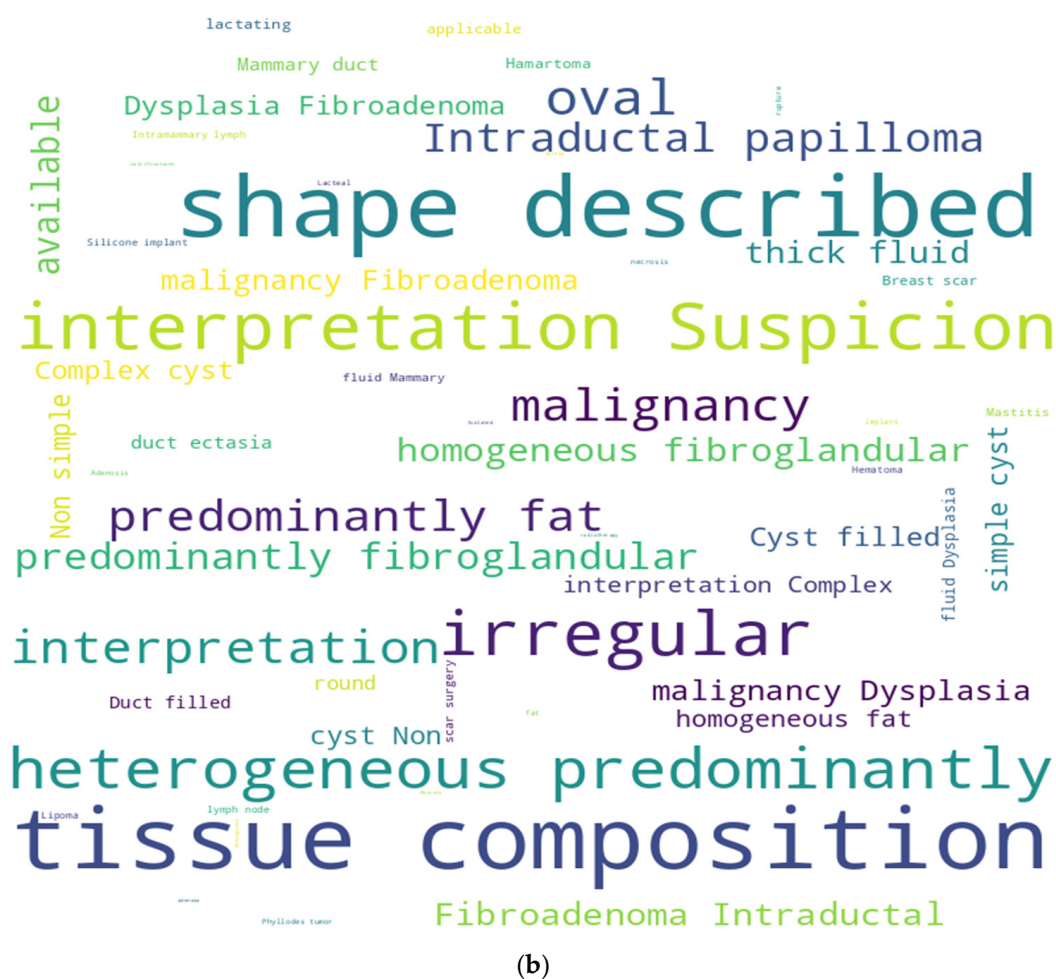
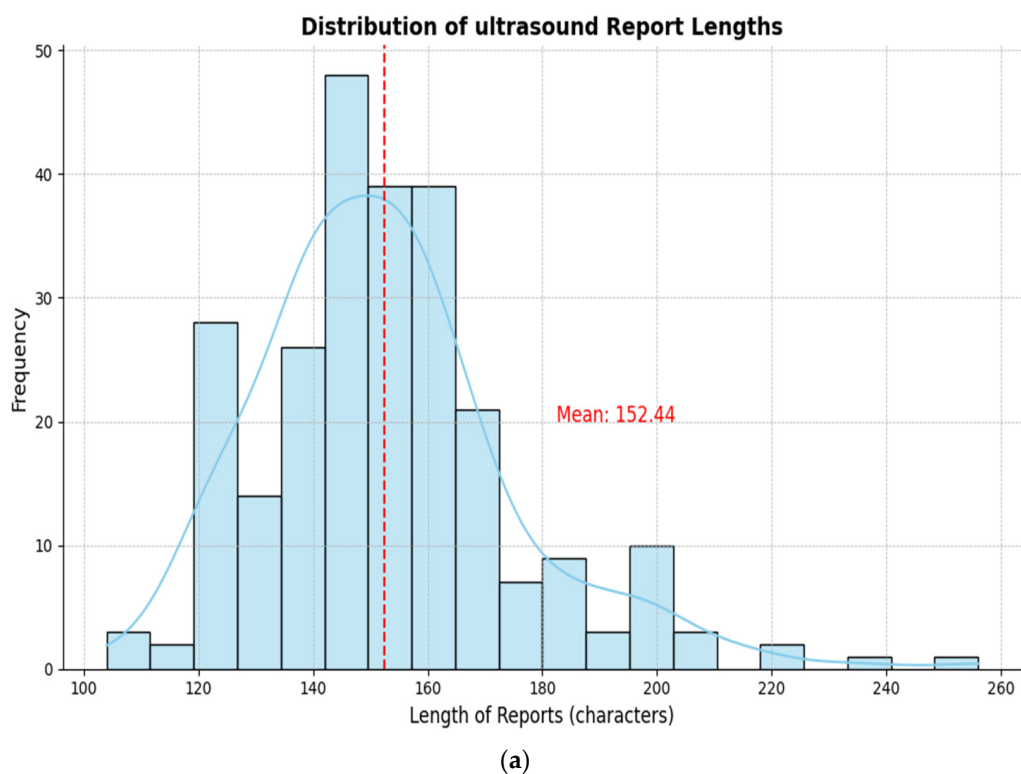


Figure 2. Average ground truth length and word cloud. (a) Ground length (red line indicates average length which is 152.44 characters), (b) Word cloud analysis.

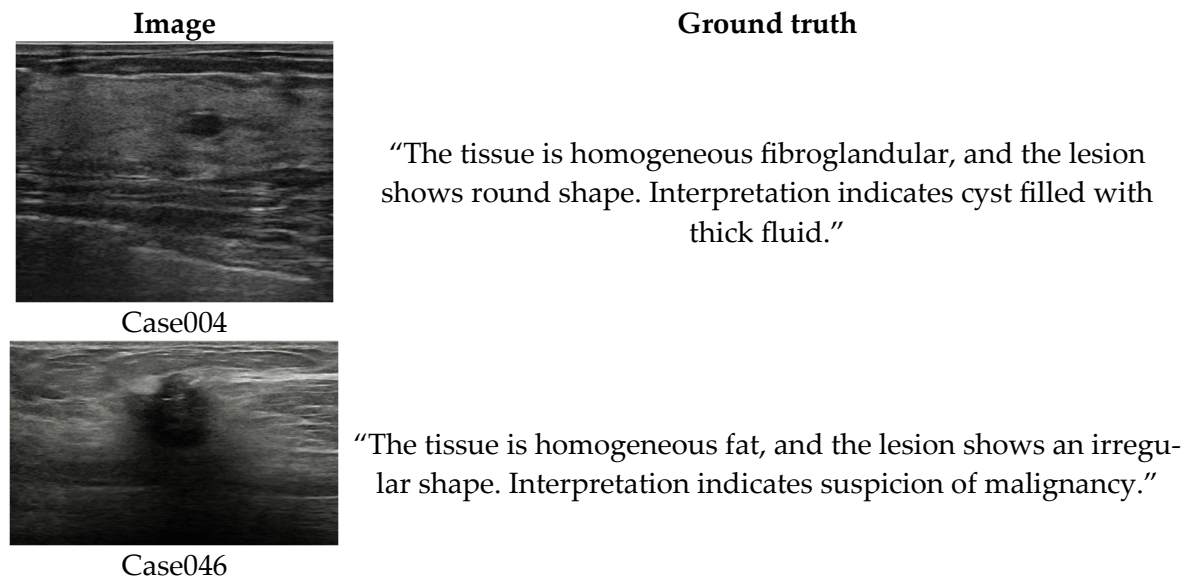


Figure 3. Sample data (breast ultrasound image with its corresponding ground truth used for model training).

ROUGE-L evaluates the longest common subsequence (LCS) between a generated report and a reference report, capturing both precision and recall without requiring consecutive matches.

$$ROUGE - L = \frac{LCS(C, R)}{R} \quad (18)$$

where $LCS(C, R)$ is the longest common subsequence of the candidate and the reference.

METEOR computes the alignment between generated and reference texts using exact matches, stemming, and synonym matching, and balances precision and recall using a harmonic mean with higher weight on recall.

$$METEOR = F_{mean} \times (1 - P_{penalty}); P_{penalty} = \gamma \left(\frac{ch}{m} \right)^\theta \quad (19)$$

where ch is the number of matched chunks, m is the total number of matched unigrams, and γ and θ are tunable parameters.

CIDEr measures the similarity between generated and reference reports using TF-IDF-weighted n-gram vectors, emphasizing the consensus across multiple references computed as:

$$CIDEr = \frac{1}{N} \sum_{n=1}^N \frac{\vec{g}_n \times \vec{r}_n}{\|\vec{g}_n\| \|\vec{r}_n\|} \quad (20)$$

where $\vec{g}_n$ and $\vec{r}_n$ are TF-IDF-weighted n-gram vectors of the generated and reference reports; N denotes the maximum n-gram length. The cosine similarity is computed for each n-gram level and averaged.

These metrics provide a comprehensive evaluation of generated ultrasound reports by capturing lexical accuracy (BLEU), sequence-level structural similarity (ROUGE-L), semantic and synonym-aware alignment (METEOR), and clinical relevance (CIDEr). This multi-metric evaluation ensures a robust assessment of both linguistic quality and clinical relevance in breast ultrasound report generation.

5. Result and Analysis

The ViT-based model was trained from scratch on breast ultrasound images to extract global visual features for report generation. To optimize performance, a series of experiments were conducted using different configurations, with depths of 9, 12, and 15 Transformer blocks, corresponding to shallow, base, and large model variants.

For all configurations, a patch size of 16×16 pixels was used, with each patch projected onto a 768-dimensional embedding, which also defined the dimensionality of the learnable positional embeddings. The Transformer layers incorporated 12 attention heads in their multi-head self-attention mechanism, allowing the model to attend to diverse spatial relationships across the patches. The dimensionality of FFN is set to capture complex spatial and contextual features. During training, all images were resized to 224×224 pixels and fed into the model in mini batches of 16 images. The model was optimized using the AdamW optimizer, with a learning rate of 1×10^{-3} and a weight decay of 0.01 and drop out of 0.1. Training was performed for 30 epochs, with the CLS token embedding from the final Transformer block extracted as the global feature vector for each image. Figure 4 shows the training loss of the ViT model with 9, 12, and 15 Transformer blocks. In the final encoder–decoder architecture ViT-12 is used due to the high computational requirements of a larger model and limited access to high-performance computing resources, such as those available through Google Colab GPU.

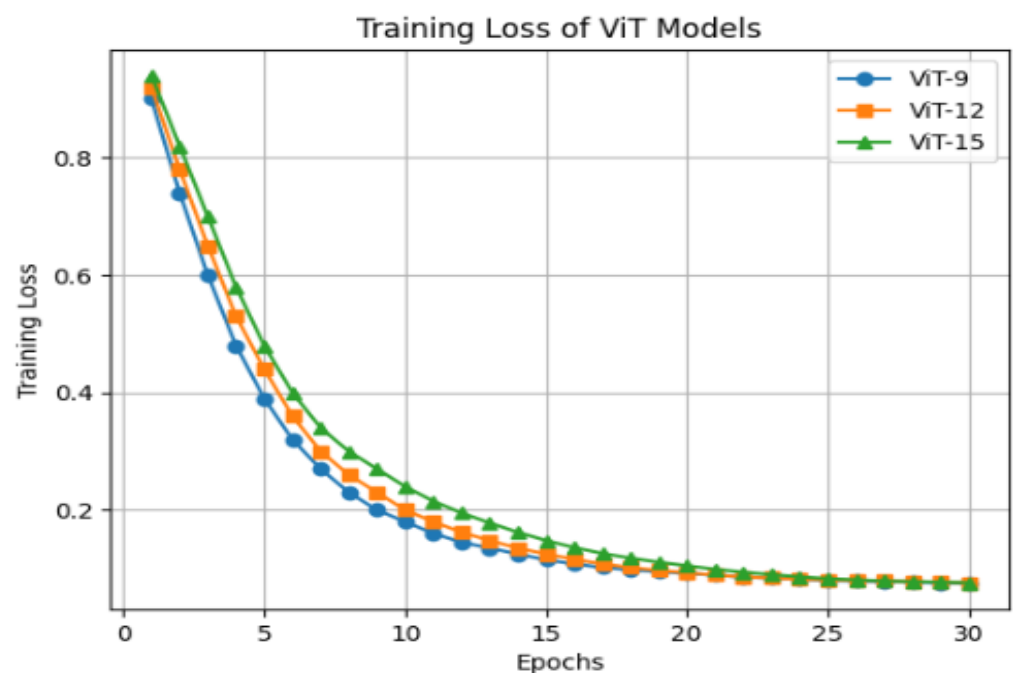


Figure 4. Training loss of ViT model with 9, 12, and 15 Transformer blocks.

For textual embedding, BERT and BioBERT are used as text encoders. These encoders were initialized with pretrained weights and further fine-tuned on image–ground truth pairs to provide better contextual embeddings. The embedding dimension set at 768 enables the representation of input tokens extracting the semantic relationship between text tokens and image embeddings. The sequence length is set at 1024 tokens to handle different medical terms in the report, which enables the model to adapt to a wide range of input text. GPT-2 was also used in some experiments to evaluate alternative text encoding strategies.

A Transformer decoder was used to generate the final reports from the fused multimodal features. It used the same 768-dimensional embedding, 12 attention heads, and a 3072-dimensional feed-forward network to ensure consistency across components and

support coherent text generation. The decoder is also set with a maximum sequence length of 1024 tokens and 0.1 dropout rate for stable generalization. GPT-2 was additionally used as a decoder to assess its generative performance in producing clinically coherent reports. All models were trained using the AdamW optimizer, with a learning rate of 1×10^{-3} , a weight decay of 0.01, a batch size of 16, and 30 training epochs, providing balanced and stable convergence throughout the multimodal architecture. During inference, beam search with a beam width of 5 was used to generate the most clinically coherent and accurate reports from the multimodal features.

5.1. Quantitative Analysis

Different combinations of encoder–decoder architectures are implemented to systematically evaluate the performance of the multimodal report generation architecture. The results shown in Table 2 indicate that the choice of text encoder and decoder significantly impacts the performance of the generated reports.

Table 2. Comparison of different configurations of encoder–decoder architectures for breast ultrasound report generation. Bold values indicate the best-performing models.

Model Configurations	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	CIDEr
ViT + BERT + Transformer Decoder	0.590	0.435	0.305	0.135	0.590	0.360	0.880
ViT + BioBERT + Transformer Decoder	0.635	0.495	0.345	0.160	0.614	0.395	1.00
ViT + BioBERT as Encoder + Decoder	0.665	0.525	0.375	0.175	0.648	0.425	1.08
ViT + BioBERT Encoder + GPT2 Decoder	0.650	0.510	0.360	0.180	0.670	0.440	1.05
ViT + GPT2 as Encoder + Decoder	0.555	0.415	0.285	0.160	0.650	0.435	0.820

- ViT + BERT + Transformer decoder, which serves as the baseline, achieves moderate BLEU (0.590) and ROUGE-L (0.590) scores, reflecting its ability to capture general visual–textual correlations, but its performance is low due to BERT’s lack of biomedical domain knowledge.
- ViT + BioBERT + Transformer decoder improves upon the baseline across all metrics. BLEU-1 scores increase to 0.635, indicating better n-gram overlaps with the ground truth, while ROUGE-L (0.614) and METEOR (0.395) improvements demonstrate enhanced fluency and semantic consistency in generated text. This confirms the advantage of incorporating biomedical-specific text embeddings for encoding clinical terminology.
- ViT + BioBERT as the encoder and decoder achieves the highest BLEU 1-3, ROUGE-L, and CIDEr scores among all tested configurations. The use of BioBERT for both encoding and decoding allows the model to fully leverage domain-specific knowledge in both text encoding and generation stages, resulting in more accurate and clinically relevant reports.
- ViT + BioBERT Encoder + GPT Decoder shows slightly lower BLEU and CIDEr scores compared to the fully BioBERT model but achieves higher ROUGE-L and METEOR scores. This indicates that GPT improves fluency and readability, producing more coherent and natural-sounding reports, albeit with a slight reduction in precise medical terminology accuracy.
- ViT + GPT as the encoder and decoder produces fluent reports with high ROUGE-L and METEOR scores but the lowest BLEU and CIDEr scores, suggesting weaker domain-specific accuracy. While the generated reports are readable, they are less precise in capturing clinical details compared to fully BioBERT-based models.

Across all metrics, the ViT- BioBERT (encoder + decoder) model achieved the highest overall scores. Models integrating GPT-2, particularly using GPT-2 as both the encoder and decoder, performed lower on all metrics, suggesting that GPT-based components are less suited for highly structured medical-domain text without domain-specific pretraining.

5.2. Qualitative Analysis

Quantitative metrics, as shown in Table 2, only provide a partial evaluation due to the imbalance and bias towards most of the cases such as benign and malignant in the dataset. To assess the overall quality of the generated reports, we performed a manual comparison between various test samples of image–ground truth pairs and the reports generated by each model and verified with a clinical expert whether the text accurately reflected visual features such as tissue echogenicity, lesion shape, boundary regularity, and structural heterogeneity. Generated words/report segments that accurately described the visual and pathological characteristics of the lesions were marked as clinically correct, while segments that misrepresented, omitted, or distorted key image features were marked as clinically incorrect.

Figure 5 shows few samples of image–report pairs for different encoder–decoder architectures along with the ground truth. It can be observed that high-performing models, particularly those using BioBERT, not only reproduce ground truth terminology but also correctly interpret visual characteristics directly observable in the ultrasound images, such as *round* or *oval shapes*, *well-circumscribed margins*, and *homogeneous fibroglandular tissue*. This demonstrates the model’s ability for effective visual understanding beyond textual memorization, whereas GPT2-based models tend to recognize lesion shape but omit secondary visual indicators, such as *margin smoothness* or *internal consistency*, resulting in shorter and less clinically informative reports.

For images with irregular lesion borders, which are also visually apparent in images, BioBERT-based models correctly associated *irregular margins* with *suspicion of malignancy* (see sample 2). These models employ radiologically appropriate language (e.g., “findings raise suspicion”), reflecting both visual irregularity and clinical reasoning.

Lower-performing models, while identifying *irregular shapes*, failed to contextualize these findings within *malignancy risk*. For images showing *heterogeneous, fat-predominant tissue with oval lesions*, the models accurately capture both *tissue heterogeneity* and *lesion geometry*, translating these visual patterns into diagnoses such as *dysplasia* and *mammary duct ectasia*. This indicates the correct handling of multi-feature visual cues and multi-label interpretation, which is a critical requirement for real-world clinical deployment.

An important observation from the visual comparison (Table 3 and Figure 5) is that models with higher quantitative scores also demonstrate strong visual–textual alignment. The correct identification of lesion shape (*round* vs. *irregular* vs. *oval*), tissue composition (*homogeneous* vs. *heterogeneous*), and associated diagnostic implications suggest that these models are learning clinically meaningful representations rather than relying solely on language priors.

These findings highlight the importance of aligning visual feature extraction with medically informed report generation. While general-purpose language models may achieve grammatical fluency, they are insufficient for capturing the diagnostic precision required in clinical practice.

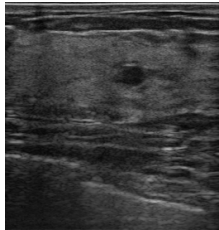
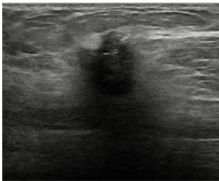
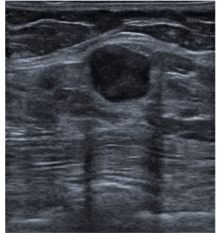
Image	Ground Truth	ViT + BERT + Transformer Decoder	ViT + BioBERT + Transformer Decoder	ViT + BioBERT as Encoder+ Decoder	ViT + BioBERT Encoder + GPT2 Decoder	ViT + GPT2 as Encoder + Decoder
	The tissue is homogeneous fibroglandular, and the lesion shows round shape. Interpretation indicates cyst filled with thick fluid.	The tissue is homogenousl y fat; lesion appears round. Cyst suspected.	Homogeneous fibroglandular tissue. Round lesion. Smooth margins, a cyst filled with thick fluid.	Homogeneous fibroglandular tissue, round, well-circumscribed lesion, cyst filled with thick fluid.	Fibroglandula r tissue with a round lesion showing smooth borders. The lesion appears cystic.	Fibroglandula r tissue with a round lesion. Cyst likely.
	The tissue is homogeneously fat, and the lesion shows irregular shape. Interpretation indicates suspicion of malignancy.	The tissue is predominantl y fat, lesion appears irregular. Malignancy suspected.	Homogeneous fatty tissue, irregular lesion, heterogeneous appearance; suspicion for malignancy.	Homogeneous fatty tissue, irregular, poorly circumscribed lesion; suspicion for malignancy.	Fatty tissue with an irregular lesion lacking smooth margins. The findings are suggestive of possible malignancy.	Fat tissue with an irregular lesion. Malignancy possible.
	The tissue is heterogeneous: predominantly fat lesion shows oval shape. Interpretation indicates dysplasia & mammary duct ectasia.	The tissue is mostly fat, heterogeneity, and lesion is oval. Dysplasia indicated.	Heterogeneous, predominantly fatty tissue, oval lesion, smooth margins. Interpretation dysplasia with associated mammary duct ectasia.	Fat-dominant heterogeneous tissue. An oval, well-defined lesion is present, consistent with dysplasia and mammary duct ectasia.	Heterogeneous fatty tissue with an oval lesion is noted. Findings suggest dysplasia.	Fatty tissue with an oval lesion. Dysplasia indicated.

Figure 5. Comparison of generated breast ultrasound reports using different encoder–decoder architectures. Green indicates clinically correct findings with ground truth, yellow indicates correct translation of visual patterns, and red indicates incorrect findings.

Table 3. Research contributions and comparative performance of ARG models for different radiological datasets.

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE-L	METEOR
IU X-Ray [7]	CNN-RNN [14]	0.374	0.224	0.252	0.109	0.359	0.307	0.163
	CNN-Transformer [13]	0.497	0.332	0.240	0.179	-	0.366	0.204
	Memory-driven ResNet50-Transformer [18]	0.489	0.325	0.238	0.184	-	0.394	0.210
	ConvNext-BioBERT [6]	0.479	0.363	0.261	0.173	0.279	0.354	0.188
	Swin Transformer-DistilGPT [20]	0.675	0.585	0.523	0.472		0.698	0.382
	R2GenGPT [21]	0.488	0.316	0.288	0.173	0.438	0.377	0.211
	ViT-GPT2 [24]	0.571	0.385	0.291	0.226		0.433	
MIMIC-CXR [8]	CNN-Transformer [13]	0.399	0.255	0.176	0.128		0.297	0.158
	ResNet50-Transformer [18]	0.379	0.230	0.153	0.109		0.284	0.149
	R2GenGPT [21]	0.411	0.267	0.186	0.134	0.269	0.297	0.160
BrEaST Dataset [25]	ViT + BioBERT as Encoder + Decoder (Proposed Model)	0.665	0.525	0.375	0.175	0.648	0.425	1.08

5.3. Comparison with Existing Methods

Most existing automated radiology report generation methods have been evaluated on chest X-ray benchmark datasets such as IU X-Ray and MIMIC-CXR, as reported in Section 2. Table 3 summarizes the reported performance of representative state-of-the-art models on these datasets. All results presented in this table are directly taken from the corresponding published studies and are evaluated using standard natural language generation metrics, including BLEU, ROUGE-L, METEOR, and CIDEr. It can be observed from the table that the earlier CNN-RNN and CNN-Transformer-based approaches achieve BLEU-1 scores ranging from approximately 0.37 to 0.50 on the IU X-Ray dataset. More recent Transformer-based architectures, such as Swin Transformer-DistilGPT, R2GenGPT, and ViT-GPT2, show noticeable improvements, with BLEU-1 scores reaching up to 0.67 and ROUGE-L values reaching 0.70. In contrast, results reported on the larger and more diverse MIMIC-CXR dataset remain comparatively lower, with BLEU-1 scores typically between 0.38 and 0.41 and ROUGE-L values below 0.30, highlighting the difficulty of large-scale chest radiology report generation.

Despite these advances, a direct quantitative comparison between the proposed method and existing approaches is not feasible due to fundamental differences in imaging modality and dataset characteristics. Chest X-ray images are generally high-contrast and anatomically structured, focusing on thoracic organs, whereas breast ultrasound images emphasize lesion appearance, boundary identification, and tissue heterogeneity under low-contrast and noisy imaging conditions. It is worth noting that none of the existing studies on radiological report generation methods have been evaluated on breast ultrasound datasets. To the best of our knowledge, no benchmark results are currently available for the BrEaST dataset used in this study; therefore, no direct comparison of the proposed Transformer-based architecture is feasible.

Although direct comparison across datasets is not appropriate, it is still informative to examine performance trends across architecturally comparable models. Our best-performing configuration, the ViT + BioBERT encoder-decoder, achieves a BLEU-1 score of 0.665 and a ROUGE-L score of 0.648 on the BrEaST dataset. These results are comparable to those reported by similar architectures, for example, the Swin Transformer-DistilGPT [20]

model evaluated on the IU X-Ray dataset, which achieves a BLEU-1 score of 0.675 and a ROUGE-L score of 0.698. Considering the increased complexity and variability of ultrasound imaging, these findings indicate that the proposed ARG system delivers a competitive and consistent performance within a more challenging and previously underexplored medical imaging domain. This further demonstrates the effectiveness and generalizability of the proposed architecture beyond chest X-ray-centric report generation models.

6. Limitations

Despite the higher performance of the proposed model, several limitations are worth noting. First, the dataset used in this study is relatively smaller and contains limited image abnormalities and report variations. This restricts the ability to generalize the model capability to other types of medical images or large-scale X-ray datasets such as the IU X-Ray and MIMIC-CXR datasets commonly used in prior studies. Expanding the dataset to include larger and more diverse breast ultrasound datasets, along with corresponding radiological reports, would significantly enhance the robustness, generalizability, and clinical applicability of the proposed approach.

Second, the model relies on large architectures such as ViT, BioBERT, and GPT-2, which require substantial computational resources for training and fine tuning. Limited access to high-performance GPUs can constrain the full training potential and lead to suboptimal model performance. Future work could explore model compression and pruning techniques to reduce computational overhead, enabling the development of resource-efficient variants accessible in low-resource settings.

Third, medical report writing is subjective and influenced by individual practitioner style, expertise, and clinical judgment. The report generated by the model may not effectively capture these variations, leading to inconsistent or potentially inaccurate outputs. Incorporating human-in-the-loop with semi-supervised strategies and expert feedback mechanisms would allow the model to adapt to these variabilities and improve both reliability and clinical trustworthiness. Additionally, integrating explainable AI (XAI) techniques could enhance transparency by linking generated textual descriptions to specific visual features in ultrasound images.

Fourth, capturing the highly specialized medical terminology is yet another challenge. Although Transformer-based architectures excel at contextual understanding, accurately reflecting the complexity of medical terminology in reports is still challenging. Additionally, the evaluation metrics currently being used focus on word overlap and may not fully reflect the clinical relevance or accuracy of the generated output. Human evaluation by medical experts is necessary for robust assessment, which was beyond the scope of this study.

Finally, real-world clinical deployment of the system poses additional challenges. Integrating the automated report generation model into existing workflows requires rigorous validation, user acceptance testing, and seamless integration with medical imaging systems and electronic health records. Future research should address these practical considerations to ensure safe and effective adoption in clinical practice.

7. Conclusions

In this study, we presented a multimodal Transformer-based framework for automated breast ultrasound report generation that effectively integrates visual and textual information to produce clinically meaningful diagnostic narratives. By leveraging ViT for visual feature extraction and pretrained language models, BERT, BioBERT, and GPT-2, for textual encoding, the proposed system captures lesion morphology, tissue composition, and diagnostic semantics essential for ultrasound interpretation.

Comprehensive quantitative evaluation using BLEU, ROUGE-L, METEOR, and CIDEr demonstrates that BioBERT-based encoder–decoder configurations achieve superior clinical specificity and semantic alignment, while GPT-2-based decoders enhance linguistic fluency. The dual cross-attention mechanism enables effective multimodal alignment, allowing the decoder to generate reports in both image-derived visual cues and biomedical language representations.

These findings highlight the importance of encoder–decoder selection and multimodal fusion strategies in medical report generation. The proposed framework establishes a strong foundation for automated breast ultrasound reporting and offers valuable insights for future clinical decision-support systems aimed at improving reporting consistency, efficiency, and diagnostic reliability. Despite the promising results, the dataset size and diversity may limit the generalizability of the proposed model across different image modalities. Expanding the dataset to include more cases and report variations would improve the generalizability and clinical applicability.

Author Contributions: Conceptualization, A.M.; methodology, S.K.; validation, S.K. and A.M.; data curation, S.K.; writing—original draft preparation, S.K.; writing—review and editing, A.M.; experimental analysis, S.K. and A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the data used were obtained from the public databases.

Informed Consent Statement: Patient consent was waived due to the data used were obtained from the public databases.

Data Availability Statement: The original data presented in the study are openly available in the “Cancer Imaging Archive (TCIA)” in the Breast Lesions Ultrasound (USG) collection available at: <https://www.cancerimagingarchive.net/collection/breast-lesions-usg/> (accessed on 15 July 2025).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nakaura, T.; Yoshida, N.; Kobayashi, N.; Shiraishi, K.; Nagayama, Y.; Uetani, H.; Kidoh, M.; Hokamura, M.; Funama, Y.; Hirai, T. Preliminary assessment of automated radiology report generation with generative pre-trained transformers: Comparing results to radiologist-generated reports. *Jpn. J. Radiol.* **2024**, *42*, 190–200. [[CrossRef](#)]
2. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
3. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
4. Cornia, M.; Stefanini, M.; Baraldi, L.; Cucchiara, R. Meshed-memory transformer for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
5. Pan, Y.; Yao, T.; Li, Y.; Mei, T. X-linear attention networks for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
6. Alqahtani, F.F.; Mohsan, M.M.; Alshamrani, K.; Zeb, J.; Alhamami, S.; Alqarni, D. CNX-B2: A novel cnn-transformer approach for chest x-ray medical report generation. *IEEE Access* **2024**, *12*, 26626–26635. [[CrossRef](#)]
7. Demner-Fushman, D.; Kohli, M.D.; Rosenman, M.B.; Shooshan, S.E.; Rodriguez, L.; Antani, S.; Thoma, G.R.; McDonald, C.J. Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Med. Inform. Assoc.* **2015**, *23*, 304–310. [[CrossRef](#)] [[PubMed](#)]
8. Johnson, A.E.; Pollard, T.J.; Greenbaum, N.R.; Lungren, M.P.; Deng, C.-Y.; Peng, Y.; Lu, Z.; Mark, R.G.; Berkowitz, S.J.; Horng, S. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv* **2019**, arXiv:1901.07042.

9. Fang, H.; Gupta, S.; Iandola, F.; Srivastava, R.K.; Deng, L.; Dollár, P.; Gao, J.; He, X.; Mitchell, M.; Platt, J.C. From captions to visual concepts and back. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
10. Wang, T.; Chen, W.; Tian, Y.; Song, Y.; Mao, Z. Improving image captioning via predicting structured concepts. *arXiv* **2023**, arXiv:2311.08223. [[CrossRef](#)]
11. Li, J.; Mao, Z.; Li, H.; Chen, W.; Zhang, Y. Exploring Visual Relationships via Transformer-based Graphs for Enhanced Image Captioning. *ACM Trans. Multimed. Comput. Commun. Appl.* **2024**, *20*, 133. [[CrossRef](#)]
12. Fu, F.; Fang, S.; Chen, W.; Mao, Z. Sentiment-oriented transformer-based variational autoencoder network for live video commenting. *ACM Trans. Multimed. Comput. Commun. Appl.* **2024**, *20*, 104. [[CrossRef](#)]
13. Jin, Y.; Chen, W.; Tian, Y.; Song, Y.; Yan, C. Improving radiology report generation with multi-grained abnormality prediction. *Neurocomputing* **2024**, *600*, 128122. [[CrossRef](#)]
14. Singh, S.; Karimi, S.; Ho-Shon, K.; Hamey, L. From chest x-rays to radiology reports: A multimodal machine learning approach. In *2019 Digital Image Computing: Techniques and Applications (DICTA)*; IEEE: New York, NY, USA, 2019.
15. Nguyen, E.; Theodorakopoulos, D.; Pathak, S.; Geerdink, J.; Vijlbrieff, O.; Van Keulen, M.; Seifert, C. A hybrid text classification and language generation model for automated summarization of dutch breast cancer radiology reports. In *2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI)*; IEEE: New York, NY, USA, 2020.
16. Tsai, M.-C.; Chu, K.-C.; Li, Y.-X. Building and Validating a Clinical Ultrasound Image Reporting Model. In *2023 IEEE 24th International Conference on Information Reuse and Integration for Data Science (IRI)*; IEEE: New York, NY, USA, 2023.
17. Himel, M.H.A.M.H.; Chowdhury, P.; Hasan, M.A.M. A robust encoder decoder based weighted segmentation and dual staged feature fusion based meta classification for breast cancer utilizing ultrasound imaging. *Intell. Syst. Appl.* **2024**, *22*, 200367. [[CrossRef](#)]
18. Tao, Y.; Ma, L.; Yu, J.; Zhang, H. Memory-based Cross-modal Semantic Alignment Network for Radiology Report Generation. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 4145–4156. [[CrossRef](#)]
19. Chaudhury, S.; Sau, K.; Shelke, N. Transforming Breast Cancer Image Classification with Vision Transformers and LSTM Integration. In *2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*; IEEE: New York, NY, USA, 2024.
20. Singh, P.; Singh, S. ChestX-Transcribe: A multimodal transformer for automated radiology report generation from chest X-rays. *Front. Digit. Health* **2025**, *7*, 1535168. [[CrossRef](#)]
21. Wang, Z.; Liu, L.; Wang, L.; Zhou, L. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiol.* **2023**, *1*, 100033. [[CrossRef](#)]
22. Wang, Z.; Liu, L.; Wang, L.; Zhou, L. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023.
23. Chen, Z.; Song, Y.; Chang, T.-H.; Wan, X. Generating radiology reports via memory-driven transformer. *arXiv* **2020**, arXiv:2010.16056.
24. Raminedi, S.; Shridevi, S.; Won, D. Multi-modal transformer architecture for medical image analysis and automated report generation. *Sci. Rep.* **2024**, *14*, 19281. [[CrossRef](#)]
25. Pawłowska, A.; Ćwierz-Pieńkowska, A.; Domalik, A.; Jaguś, D.; Kasprzak, P.; Matkowski, R.; Fura, Ł.; Nowicki, A.; Żółek, N. Curated benchmark dataset for ultrasound based breast lesion analysis. *Sci. Data* **2024**, *11*, 148. [[CrossRef](#)]
26. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019.
27. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [[CrossRef](#)]
28. Rai, H.M.; Dashkevych, S.; Yoo, J. Next-generation diagnostics: The impact of synthetic data generation on the detection of breast cancer from ultrasound imaging. *Mathematics* **2024**, *12*, 2808. [[CrossRef](#)]
29. Chen, H.; Li, Y.; Zhang, J.; Yang, L.; Sun, Y.; Chen, Y.; Zhou, S.; Li, Z.; Qian, X.; Xu, Q. An Alignment and Imputation Network (AINet) for Breast Cancer Diagnosis with Multimodal Multi-view Ultrasound Images. *IEEE Trans. Med. Imaging* **2025**, early access.
30. Rai, H.M.; Yoo, J.; Agarwal, S.; Agarwal, N. LightweightUNet: Multimodal Deep Learning with GAN-Augmented Imaging Data for Efficient Breast Cancer Detection. *Bioengineering* **2025**, *12*, 73. [[CrossRef](#)] [[PubMed](#)]
31. Wu, L.; Zhuang, Y.; Liao, G.; Han, L.; Hua, Z.; Wang, R.; Chen, K.; Lin, J. Breast tumor detection in ultrasound images with anatomical prior knowledge. *Image Vis. Comput.* **2025**, *162*, 105724. [[CrossRef](#)]
32. Pang, X.; Li, Z.; Lv, J.; Ge, R.; Wu, Z.; Gao, F. BI-RADS Boosted Breast Cancer Diagnosis with Masked Pretraining On Imbalanced Ultrasound Data. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*; IEEE: New York, NY, USA, 2025.

33. Hendry, A.; Felicia, E.; Natanael, Y.; Rumagit, R.Y. Implementation of Explainable AI on CNN-and ViT-Based Models for Classifying Breast Cancer. *Procedia Comput. Sci.* **2025**, *269*, 968–978. [[CrossRef](#)]
34. Urdaneta, A.D.; Ghosh, R. Deep Learning-Based Multi-Classification of Breast Cancer Ultrasound Images Using Convolutional Neural Networks. In *2025 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA)*; IEEE: New York, NY, USA, 2025.
35. Huang, Y.; Chang, A.; Dou, H.; Tao, X.; Zhou, X.; Cao, Y.; Huang, R.; Frangi, A.F.; Bao, L.; Yang, X. Flip Learning: Weakly supervised erase to segment nodules in breast ultrasound. *Med. Image Anal.* **2025**, *102*, 103552. [[CrossRef](#)]
36. Yu, X.; Chen, Y.; He, G.; Zeng, Q.; Qin, Y.; Liang, M.; Luo, D.; Liao, Y.; Ren, Z.; Kang, C. Simple is what you need for efficient and accurate medical image segmentation. *Expert Syst. Appl.* **2025**, *304*, 130687. [[CrossRef](#)]
37. Mallina, R.; Shareef, B. XBusNet: Text-Guided Breast Ultrasound Segmentation via Multimodal Vision–Language Learning. *Diagnostics* **2025**, *15*, 2849. [[CrossRef](#)] [[PubMed](#)]
38. Arsa, D.M.S.; Ilyas, T.; Park, S.-H.; Chua, L.; Kim, H. Efficient multi-stage feedback attention for diverse lesion in cancer image segmentation. *Comput. Med. Imaging Graph.* **2024**, *116*, 102417. [[CrossRef](#)] [[PubMed](#)]
39. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, USA, 2–12 July 2002.
40. Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*; Association for Computational Linguistics: Barcelona, Spain, 2004.
41. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, MI, USA, 29 June 2005; pp. 65–72.
42. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 7–12 June 2015.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.