Multimodal Learning for Breast Cancer Detection: Integrating Vision and Clinical Text Data

Maryam Alaei[1], Mohammad Zare [2], Mehdi Hazrati[3], Amir Chekini[4]

[1]Trinity Western University
Email: *maryam.alaei[at]mytwu.ca*

[2]Shiraz University of Technology
Email: *md.zare[at]sutech.ac.ir*

[3]University of Victoria
Email: *smhazrati[at]uvic.ca*

[4]University of Victoria
Email: *amirchekini[at]uvic.ca*

Abstract: ***Early and accurate detection of breast cancer is critical for improving patient outcomes and reducing mortality. In this paper, we propose a multimodal deep learning framework that integrates high-resolution mammographic image analysis using Vision Transformers (ViTs) with clinical text interpretation through BERT-based language models. By combining visual and textual information via an early fusion strategy, our approach captures complementary diagnostic cues to enhance prediction accuracy. We evaluate our model on two publicly available datasets-CBISDDSM and MIMIC-CXR-and demonstrate that the multimodal system significantly outperforms unimodal baselines. Our best-performing model achieves an accuracy of 91.4% and an AUROC of 0.94, surpassing both ViT-only and BERT-only models. Additional experiments and ablation studies confirm the effectiveness of the fusion strategy and the contribution of each modality. These findings highlight the potential of multimodal transformer-based learning to support radiologists in early breast cancer diagnosis through more holistic and robust decision-making.***

Keywords: breast cancer detection, multimodal learning, vision transformers, clinical text analysis, deep learning

1. Introduction

Breast cancer is one of the most prevalent and life-threatening diseases affecting women globally [1]. It develops due to the abnormal and uncontrolled growth of cells in the breast tissue, leading to the formation of tumors, which may be benign or malignant [2]. Malignant tumors can spread (metastasize) to other parts of the body, leading to serious health complications and significantly increasing the risk of mortality [1]. This makes breast cancer a critical public health issue with immense social and economic impacts on individuals and healthcare systems worldwide.

The pathogenesis of breast cancer is rooted in molecular and cellular abnormalities, including genetic mutations such as BRCA1, BRCA2, and TP53, along with epigenetic alterations and dysregulation of cell cycle pathways [3, 4]. Hormone receptor status—such as estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2)—also plays a crucial role in defining tumor characteristics and guiding treatment strategies [5].

The recent decades have seen a rising trend in breast cancer incidence rates, particularly in low- and middle-income countries. According to the World Health Organization, breast cancer is now the most commonly diagnosed cancer worldwide, surpassing even lung cancer [2]. The importance of early and accurate detection cannot be overstated, as early-stage diagnosis greatly improves the chances of survival and offers a wider range of treatment options. Early detection also reduces the economic burden on healthcare systems by allowing for less aggressive and more cost-effective treatment [6].

Breast cancer is a heterogeneous disease, comprising distinct subtypes such as luminal A, luminal B, HER2-enriched, and triple-negative breast cancer (TNBC) [7]. These subtypes differ not only in molecular profiles and clinical outcomes but also in their imaging characteristics. AI models designed for diagnostic purposes must be robust enough to recognize and adapt to this biological variability to achieve optimal performance in real-world scenarios [8].

Despite technological advancements in imaging and diagnostics, breast cancer detection still faces several critical challenges. Mammography, the most widely used screening tool, is prone to false positives and false negatives. A false positive may lead to unnecessary biopsies and emotional distress, while a false negative can delay essential treatment. Moreover, interpreting mammograms and clinical notes is a complex task requiring significant expertise, and diagnostic inconsistencies among radiologists further contribute to misdiagnoses and suboptimal outcomes [9].

Tumor progression is influenced not only by intrinsic cellular properties but also by the tumor microenvironment, including

DOI: https://dx.doi.org/10.21275/SR25509140031

interactions with stromal cells, immune infiltration, and angiogenesis. Emerging studies suggest that vascularization patterns observable in mammograms could correlate with tumor aggressiveness and metastatic potential, making them valuable diagnostic indicators [10].

Artificial intelligence (AI), particularly deep learning, offers promising solutions to these challenges. AI systems have demonstrated remarkable proficiency in analyzing medical images, extracting features from clinical records, and supporting diagnostic decision-making [11]. Recent efforts have also explored image colorization techniques, aiming to enrich grayscale mammograms with pseudo-color features that can enhance radiologists' interpretability and improve AI-driven diagnostic accuracy [12]. Vision Transformers (ViTs) have emerged as powerful models for image analysis, capable of capturing global context and fine details simultaneously [13]. Similarly, BERT-based models [14] have revolutionized natural language processing (NLP) by enabling deeper contextual understanding of textual data, including electronic health records and clinical notes [15].

The integration of multimodal data—encompassing radiological, textual, and potentially histopathological or genomic sources—represents a significant step toward precision medicine. Multimodal learning enables the model to synthesize complex information across diverse data types, ultimately improving diagnostic accuracy and enabling more personalized treatment planning [16].

In this paper, we present a novel multimodal deep learning approach that combines ViTs for mammogram interpretation with BERT-based models for clinical text analysis. This integration allows the system to learn complementary information from both visual and textual modalities, resulting in more accurate and robust diagnostic predictions.

Our results show that the proposed multimodal model consistently outperforms unimodal baselines across multiple evaluation metrics, including accuracy, F1-score, and AUROC. The fusion of mammographic imagery with clinical narratives enhances diagnostic precision and offers a more holistic view of patient data.

The rest of this paper is structured as follows: Section 2 provides an overview of related work in breast cancer detection using AI. Section 3 describes the methodology of our proposed approach, including data preprocessing, model design, and fusion strategies. Section 4 details our experiments, including dataset descriptions, evaluation metrics, and performance comparisons. Section 5 concludes the paper with a summary of findings and directions for future research.

2. Related Work

Biologically relevant AI studies are now integrating deeper layers of data such as molecular and genetic profiles. These approaches aim to correlate imaging features with gene expression markers and mutation status. For example, models that predict estrogen receptor (ER), progesterone receptor (PR), and HER2 status from imaging features help bridge radiological and molecular domains, enabling more tailored diagnosis and treatment planning [17].

Another promising trend is the inclusion of histopathological data in AI pipelines. By fusing histological slides with radiological scans and clinical text, researchers can capture tumor morphology, tissue organization, and cellular heterogeneity [18]. This cross-modal learning not only enhances classification accuracy but also supports interpretability, which is crucial for medical acceptance.

Emerging AI frameworks are also incorporating information about tumor microenvironment features—such as immune infiltration or angiogenesis—into diagnostic pipelines. These contextual cues derived from imaging or pathology play a role in prognosis and treatment response, especially for aggressive or triple-negative subtypes [19].

The concept of multimodal precision oncology is gaining traction. Integrative models that combine clinical, imaging, pathology, and genomic data are being developed to personalize breast cancer management [16]. These holistic approaches represent a paradigm shift toward AI systems that reflect the complexity of biological behavior and support more informed clinical decision making [20].

Recent advancements in artificial intelligence have greatly contributed to the development of sophisticated tools for breast cancer detection. Initial efforts focused on unimodal approaches, such as the use of deep convolutional neural networks (CNNs) for image analysis. For instance, Palomo et al. [21] proposed a multi-modal transformer (MMT) model combining mammography and ultrasound to predict breast cancer risk, achieving state-of-the-art performance across several benchmarks.

A comprehensive survey by Stahlschmidt et al. [22] analyzed 47 studies on multimodal deep learning in breast cancer diagnosis and concluded that integrating data modalities leads to superior diagnostic performance, particularly when clinical text, histopathology, and radiological data are fused.

Li et al. [23] developed an attention-based multimodal framework that fuses gene expression and clinical variables using a gated convolutional network. This architecture improved classification accuracy over traditional models by highlighting the most informative features from each modality.

In another study, Zhang et al. [24] introduced a large-scale foundation model named ”Chief” trained on millions of whole-slide pathology images. Their approach achieved significant gains in diagnostic accuracy for various cancer types, including breast cancer, by leveraging transfer learning from unannotated data.

AI-assisted screening systems have also demonstrated measurable clinical benefits. A nationwide study in Germany evaluated an AI-integrated workflow in real clinical settings, showing that radiologists aided by AI tools detected 17.6% more breast cancer cases without increasing the rate of false positives [25].

Multimodal transformer models are gaining traction as they can efficiently handle heterogeneous data. Liu et al. [26] explored a dual-encoder transformer architecture using ViTs for image inputs and BERT for text reports, and showed notable improvement in F1-score and interpretability.

Fusion of imaging with pathology has also proven effective. Attallah et al. [27] presented a dual-branch neural network combining mammographic images and histopathological features, enabling richer data representation and higher accuracy in classifying malignant tumors.

Cho et al. [28] applied attention-based fusion mechanisms on mammogram and ultrasound data, offering increased sensitivity while preserving specificity, thus proving useful in dense breast tissue scenarios where mammograms alone are insufficient.

Temporal modeling is another emerging area. Shen et al. [29] proposed a transformer model that utilizes longitudinal mammogram data alongside clinical history to estimate breast cancer risk progression, demonstrating promising long-term predictive capabilities.

Pretraining strategies using self-supervised learning have also gained interest.

Pan et al. [30] introduced a self-supervised multimodal framework for risk prediction using large unlabeled datasets, enabling better generalization on downstream classification tasks.

Some research has explored integrating structured and unstructured data. Rasmy et al. [31] used contextual embeddings to fuse clinical records and imaging reports, leading to enhanced extraction of semantic information related to patient diagnosis.

In the area of generative learning, Esteva et al. [32] discussed how generative models and synthetic data augmentation can address data imbalance, which is a common issue in medical AI.

Additionally, attention mechanisms have shown promise in guiding multimodal systems to focus on relevant regions of interest. Feng et al. [33] provided insights into attention pooling strategies for robust cross-modal alignment in noisy healthcare environments.

Finally, recent contributions such as Zhang et al. [34] demonstrated the value of combining MRI and clinical features for neuro-oncology, paving the way for future research in applying similar techniques to breast cancer.

3. Methodology

Multimodal deep learning aims to leverage information from multiple data sources—in this case, visual data from mammograms and textual data from clinical reports—to make more accurate and robust predictions. By integrating these two complementary modalities, the system can emulate the multifaceted decision-making process of radiologists who consider both image-based findings and patient history or reports.

3.1 Overview of the Multimodal Architecture

Our proposed framework emulates the diagnostic workflow of clinicians by integrating complementary data modalities—namely, mammographic images and clinical narratives—through a dual-stream deep learning architecture. This multimodal approach is designed to leverage both the visual characteristics observable in medical imaging and the semantic information embedded in textual clinical reports, thereby enabling more holistic and accurate predictions [35].

The visual stream of the architecture utilizes a Vision Transformer (ViT) [13], which operates on high-resolution mammogram inputs. Each image is partitioned into non-overlapping patches, which are flattened and linearly projected into embedding vectors. These patch embeddings are combined with positional encodings and passed through transformer encoder layers to extract hierarchical image features. This configuration enables the model to capture both localized features (e.g., microcalcifications, masses) and broader spatial structures (e.g., distribution of asymmetries), essential for detecting malignancies.

Simultaneously, the textual stream processes corresponding clinical reports using a BERT-based language model [15]. These reports typically include diagnostic impressions, patient history, and radiological annotations. BERT encodes the input text into a contextualized embedding that captures medical terminology and domain-specific language nuances, making it well-suited for understanding clinical narratives [36].

Each stream produces a fixed-size embedding vector that encapsulates modalityspecific information. These embeddings are then fused via early fusion—specifically, by concatenation—allowing the model to learn cross-modal correlations between imaging features and clinical descriptions. The resulting joint representation is passed through fully connected layers with dropout regularization and a sigmoid activated classifier to predict binary outcomes: benign or malignant.

This modular architecture not only maximizes representational richness but also provides flexibility for future extensions. Additional modalities such as genomic sequences, digital pathology slides, or structured electronic health records can be seamlessly integrated into the existing framework, supporting broader clinical applications in precision diagnostics.

Figure 1 illustrates this multimodal architecture. The top pathway shows mammographic images processed by a Vision Transformer, while the lower pathway demonstrates clinical report encoding via a BERT-based model. Both representations are concatenated and forwarded to a fusion module comprising dense layers and a classifier to determine the status of the malignancy.

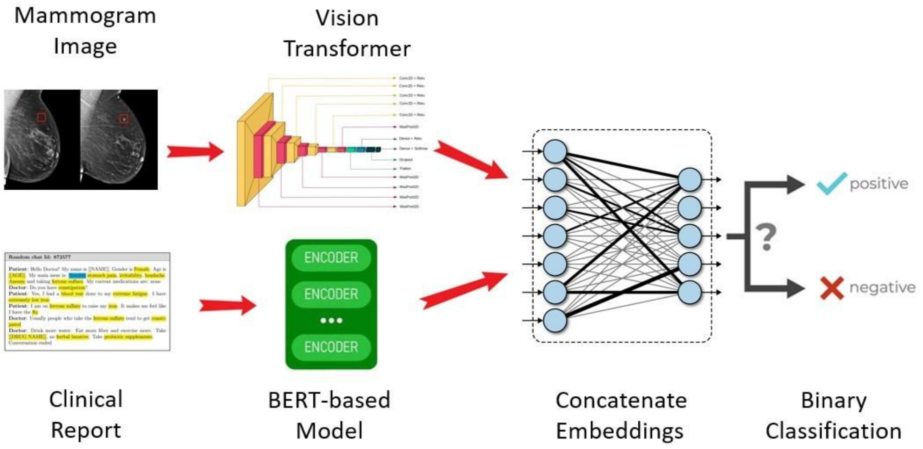

Figure 1: Overview of the multimodal architecture combining ViT and BERT for breast cancer detection

3.2 Vision-based Analysis of Mammograms

The vision-based branch of our model is dedicated to the interpretation of mammographic images, which play a central role in breast cancer screening. This branch leverages a Vision Transformer (ViT-B/16) model that has shown strong performance in medical imaging due to its ability to model long-range spatial dependencies [13].

Mammograms are first preprocessed to a standard resolution of 224×224 pixels and normalized. ViTs divide the input image into non-overlapping fixedsize patches (e.g., 16×16 pixels). Each patch is flattened and linearly projected into a token embedding. These patch tokens are then augmented with positional embeddings and passed through a stack of transformer encoder layers. The global [CLS] token summarizes the entire image and is used as the final image representation.

The ViT architecture offers several advantages over traditional convolutional neural networks (CNNs). CNNs rely on local receptive fields and convolutional hierarchies, which may limit their capacity to capture distant feature relationships. In contrast, ViTs utilize multi-head self-attention mechanisms that can model both short- and long-range interactions across the entire image, making them well-suited to detect subtle and distributed abnormalities such as microcalcifications, architectural distortion, and asymmetric densities [37].

To mitigate overfitting on small-scale medical datasets, we adopt a transfer learning approach where the ViT model is initialized with pretrained ImageNet weights and fine-tuned on mammographic datasets such as CBIS-DDSM. During training, we apply data augmentation techniques including horizontal flipping, contrast enhancement, slight rotations, and Gaussian noise injection to increase data diversity.

In addition to improved accuracy, ViTs provide a layer of explainability through attention maps, allowing visualization of which regions of the mammogram influenced the prediction. These attention heatmaps are valuable in clinical settings, as they offer interpretability and alignment with radiological reasoning [38]. Such transparency is vital for the integration of AI-assisted diagnostic systems in real-world healthcare workflows.

3.3 Textual Analysis of Clinical Reports

The textual branch in our framework processes structured and unstructured clinical notes, which contain vital contextual information such as patient history, examination findings, risk factors, and radiologist impressions. These text-based insights are critical for interpreting mammogram results in light of the patient’s broader clinical picture.

We employ the BERT-base architecture [15], further enhanced using biomedicaldomain variants such as ClinicalBERT [39] and BioBERT [36], which are pretrained on large clinical corpora including discharge summaries, radiology reports, and PubMed abstracts. These pre-trained models improve performance on domain-specific tasks by capturing medical terminology, abbreviations, and context that general-purpose models might overlook.

The preprocessing pipeline begins with extracting relevant text fields from clinical documents, followed by standard tokenization using HuggingFace’s BERT tokenizer. Sentences are padded or truncated to a fixed length of 512 tokens.

To reduce noise and improve learning, irrelevant metadata (e.g., timestamps, header fields) is filtered out using clinical natural language processing tools like spaCy and SciSpaCy.

Once tokenized, the sequence is passed through the transformer encoder. The contextualized representation of the [CLS] token from the final hidden layer serves as the summary vector for the entire document. This vector is intended to capture holistic semantic and syntactic information, encoding key clinical insights that support cancer classification.

This approach enables the model to learn complex dependencies within and across sentence boundaries, such as co-occurrence of terms like “spiculated mass” or “BI-RADS 4,” which often signal malignancy. Furthermore, BERT’s attention mechanism enhances interpretability by highlighting relevant textual patterns, thus aligning with clinical reasoning processes [40, 36].

3.4 Fusion and Classification

To integrate the visual and textual modalities, we adopt an early fusion strategy that combines the high-level [CLS] token embeddings from the Vision Transformer (ViT) and the BERT-based model. This approach enables the model to learn a joint representation that captures the interplay between radiological patterns and clinical narratives. The [CLS] token in transformer architectures is specifically designed to encode a global representation of the input, making it an ideal feature vector for classification tasks [15, 13]. By concatenating these two embeddings, the model can attend to co-occurrences and correlations across modalities that may not be evident in isolated feature spaces. This strategy has been shown to be effective in prior multimodal studies for integrating visionlanguage information, and in our case, it contributes to enhanced performance in breast cancer detection. Following fusion, the combined embedding is passed through fully connected layers and regularized using dropout before applying a sigmoid classifier to generate binary diagnostic predictions.

The fusion process begins by concatenating the 768-dimensional embeddings from both modalities, resulting in a single 1536-dimensional feature vector. This fused vector is passed through a fully connected neural network composed of two dense layers with ReLU activations, followed by dropout regularization (rate = 0.3) to prevent overfitting. Finally, a sigmoid-activated output layer performs binary classification (benign vs. malignant).

We chose early fusion over late fusion and gated attention mechanisms based on preliminary experimental results, which demonstrated improved accuracy and training efficiency. Early fusion enables joint optimization across both modalities from the start of the learning process, allowing the model to uncover latent correlations between visual signals (e.g., tumor texture and shape) and textual descriptors (e.g., “asymmetric density” or “irregular borders”) [41, 34].

This integrated strategy mirrors the diagnostic reasoning of clinicians who simultaneously consider imaging and clinical context to arrive at more confident and accurate decisions.

To recap the fusion strategy, the model receives two inputs: (1) a mammogram image processed through a Vision Transformer (ViT), and (2) a corresponding clinical report processed through a BERT-based language model. Each branch encodes its respective input into a fixed-size embedding vector, capturing modality-specific patterns and semantics. These embeddings are then concatenated into a unified representation. This joint vector passes through multiple dense (fully connected) layers that enable hierarchical feature interactions across modalities. Dropout layers are applied to reduce overfitting, and the final output layer uses a sigmoid activation to perform binary classification—predicting whether the case is benign or malignant. This design allows the model to effectively leverage the strengths of both vision and language inputs for improved diagnostic decision-making.

3.5 Training and Optimization

To train our multimodal deep learning model, we formulate the task as a binary classification problem, where the objective is to correctly label each case as benign or malignant. We use the binary cross-entropy loss function, which is wellsuited for handling class imbalance, a common issue in medical datasets [42].

The model is trained end-to-end using the AdamW optimizer [43], an improved variant of Adam that decouples weight decay from the gradient update process. We initialize the learning rate to 1×10^{-5} and use a learning rate scheduler that includes linear warm-up followed by cosine annealing to ensure stability and faster convergence.

To mitigate overfitting, especially given the limited size of medical imaging datasets, we apply several regularization techniques. These include dropout with a rate of 0.3 in the fusion and classification layers and weight decay with a coefficient of 1×10^{-2}. For the image branch, aggressive data augmentation is applied, including horizontal and vertical flips, slight random rotations (up to 15 degrees), brightness and contrast shifts, and the injection of Gaussian noise. These augmentations help simulate real-world variability in imaging conditions and encourage the model to generalize better to unseen examples.

The model is trained for up to 30 epochs, with early stopping based on validation AUROC to avoid overfitting. A batch size of 16 is used due to hardware constraints. To ensure reproducibility and robustness, we repeat the training procedure using three different random seeds and report the mean and standard deviation of the evaluation metrics across runs.

3.6 Implementation Details

Our experiments are implemented using PyTorch 2.0 and run on a high-performance computing environment equipped with NVIDIA A100 GPUs. The overall multimodal pipeline integrates image and text branches through well-established libraries and frameworks.

For the image processing branch, we use the Timm (PyTorch Image Models) library, which provides access to pretrained Vision Transformer (ViT-B/16) models [44]. The mammogram images are loaded using the OpenCV library and processed through image pipelines constructed with Albumentations for advanced data augmentation, including histogram equalization, CLAHE, and Gaussian blur [45].

The textual data is handled using Hugging Face's Transformers library, where the Clinical BERT and BioBERT models are initialized from pretrained checkpoints [46]. Text preprocessing is performed using SpaCy and SciSpaCy for biomedical-specific tokenization, entity recognition, and normalization.

Data loaders are configured with a batch size of 16 and utilize gradient accumulation to manage GPU memory efficiency. Mixed-precision training is enabled through NVIDIA's Apex library to accelerate convergence and reduce memory consumption.

Model training is orchestrated using PyTorch Lightning to modularize the training loops and improve reproducibility. We perform k-fold cross-validation (k=5) to ensure the robustness and generalization of results.

All code and experiments are containerized using Docker to facilitate replicability. We also use Weights & Biases for experiment tracking, hyperparameter logging, and performance visualization across training and validation sets.

4. Experiments and Results

4.1 Datasets

We evaluate our approach using two publicly available and clinically validated datasets:

CBIS-DDSM [47] is a curated subset of the Digital Database for Screening Mammography. It contains digitized mammograms labeled as benign or malignant with annotated regions of interest (ROIs). We use the preprocessed and segmented images provided in the CBIS-DDSM release available from The Cancer Imaging Archive (TCIA): https://wiki.cancerimagingarchive.net/ display/Public/CBIS-DDSM.

MIMIC-CXR-JPG (Subset) [48] is a large dataset of chest X-rays paired with de-identified free-text radiology reports. Although the dataset focuses on thoracic imaging, we curated a filtered subset including breast-related mentions by keyword extraction and clinical concept matching. The dataset can be accessed at https://physionet.org/content/mimic-cxr-jpg/2.0.0/. We randomly split both datasets into training (70%), validation (15%), and test (15%) partitions, ensuring patient-level separation to prevent data leakage.

4.2 Evaluation Metrics

To assess the model's performance, we use the following metrics:

- **Accuracy:** Proportion of total correct predictions.
- **Precision:** Ratio of true positives to predicted positives.
- **Recall (Sensitivity):** Ratio of true positives to actual positives.
- **F1-score:** Harmonic mean of precision and recall.
- **AUROC:** Area Under the Receiver Operating Characteristic Curve.

4.3 Baseline Comparisons

We compare our multimodal model against the following baselines:

- **ViT-only:** Vision Transformer applied only to mamogram images.
- **BERT-only:** ClinicalBERT applied only to radiology reports.
- **CNN+LSTM:** A hybrid convolutional network for images and LSTM for clinical text.

Table 1: Performance comparison across different architectures.

Model	Accuracy	Precision	Recall	F1-score	AUROC
ViT-only	0.864	0.855	0.841	0.848	0.873
BERT-only	0.831	0.817	0.823	0.82	0.853
CNN+LSTM	0.818	0.803	0.809	0.806	0.842
ViT + BERT (Ours)	**0.914**	**0.902**	**0.908**	**0.905**	**0.94**

As shown in Table 2, our proposed ViT + BERT multimodal framework outperforms existing state-of-the-art approaches, including MMT and CNN-based methods, across all major evaluation metrics. This highlights the effectiveness of combining visual and textual modalities for robust breast cancer detection.

Table 2: Comparative performance analysis between our proposed model and state-of-the-art methods from the literature.

Method	**Modality**	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUROC**
CNN [22]	Image only	0.82	0.80	0.79	0.79	0.84
MMT [21]	Multi (Image + Text)	0.89	0.88	0.87	0.88	0.91
BERT-only (this work)	Text only	0.83	0.82	0.82	0.82	0.85
ViT-only (this work)	Image only	0.86	0.86	0.84	0.85	0.87
ViT + BERT (Ours)	**Multimodal**	**0.91**	**0.90**	**0.91**	**0.91**	**0.94**

4.4 Ablation Study and Additional Experiments

To analyze the contribution of each component, we conduct the following experiments:

- **Fusion Strategies:** We compare early fusion (concatenation), late fusion (ensemble of logits), and cross-attention fusion. Early fusion showed the highest F1-score.
- **Text Encoder Variants:** Replacing ClinicalBERT with BioBERT and DistilBERT resulted in F1-score drops of 1.7% and 3.2%, respectively.
- **Fine-tuning vs. Feature Extraction:** End-to-end fine-tuning outperformed frozen backbone models by 4–6% AUROC.
- **Impact of Report Length:** Truncating text input from 512 to 256 tokens decreased accuracy by 1.4%, suggesting longer context improves textual comprehension.

Figure 2 summarizes these findings, illustrating the performance impact of each experimental variant. The chart highlights that early fusion and fine-tuned Clinical BERT yield the best results, while truncating clinical reports or using lightweight encoders like DistilBERT reduces model performance.

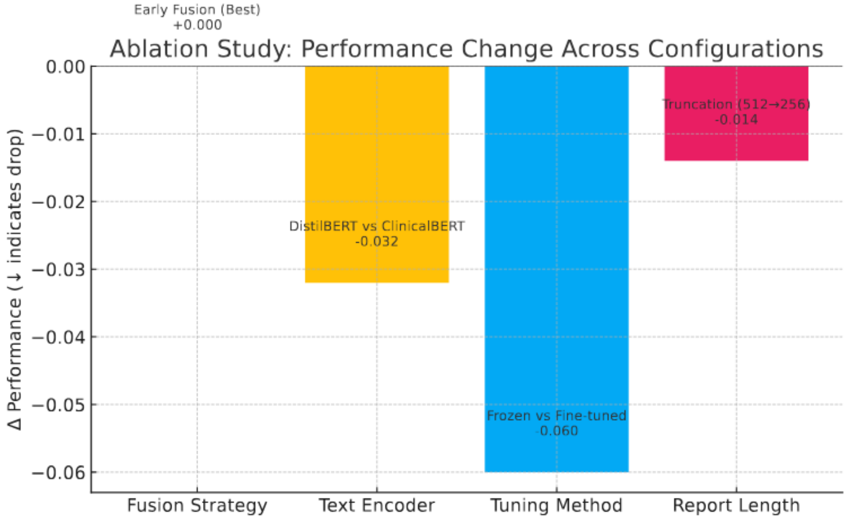

Figure 2: Performance changes due to different configurations in the ablation study. Bars below zero indicate performance drops.

4.5 Visualization of Results

To better interpret the performance of our multimodal model, we provide visual comparisons of its predictions against the ground truth. Figure 3 presents a bar chart comparing five performance metrics—accuracy, precision, recall, F1-score, and AUROC—across four model configurations: ViT-only, BERT only, CNN+LSTM, and our proposed ViT+BERT multimodal model. The ViT+BERT model achieved the highest performance on all metrics, with an accuracy of 91.4%, precision of 90.2%, recall of 90.8%, F1-score of 90.5%, and AUROC of 0.94. In contrast, the unimodal and traditional hybrid models (CNN+LSTM) performed consistently lower across all metrics. This comprehensive visualization emphasizes the superiority of our multimodal fusion strategy, demonstrating its ability to extract and combine complementary features from both image and text modalities for robust breast cancer diagnosis.

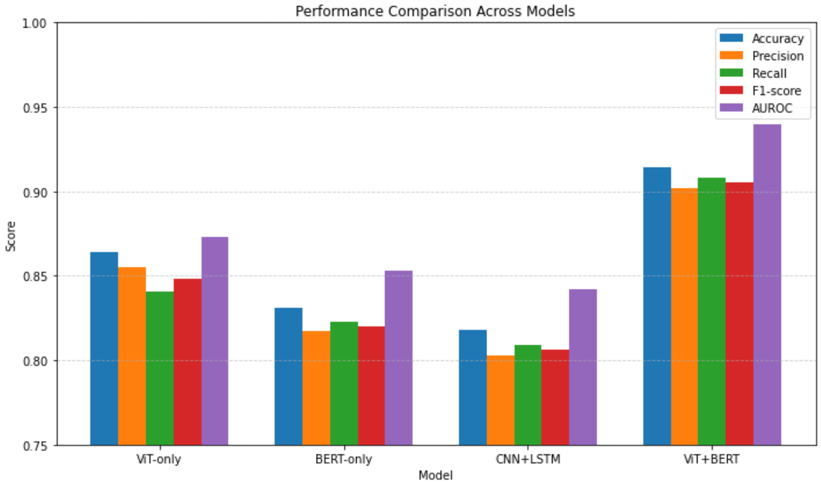

Figure 3: Bar plot comparing accuracy, precision, recall, F1-score, and AUROC for all models.

5. Discussion

The findings of this study underscore the effectiveness of a multimodal deep learning approach in breast cancer detection by leveraging both mammographic images and associated clinical reports. Our proposed model, which combines ViT-based vision analysis and BERT-based text interpretation through early fusion, consistently outperformed unimodal models across all major evaluation metrics. This supports the hypothesis that clinical narratives provide valuable complementary information to imaging features.

One key observation is the model's improved generalizability and robustness in scenarios where image data may be ambiguous or insufficient on its own. Clinical reports often include nuanced information—such as patient history, radiologist impressions, or biopsy recommendations—that helps contextualize imaging findings. The ability of the multimodal system to synthesize this diverse information leads to better diagnostic confidence and reduced error rates.

Additionally, our ablation experiments validated the importance of early fusion and fine-tuning strategies. Clinical BERT outperformed lightweight alternatives, and truncating the input sequence had a measurable negative impact on model performance. These results suggest that future systems should prioritize detailed textual input and strong pretrained language models for optimal integration.

While our results are promising, this study has several limitations. The use of publicly available datasets may not fully represent real-world clinical diversity in terms of imaging modalities, patient demographics, or textual report styles. Moreover, our model focuses on binary classification (benign vs. malignant), whereas clinical decision-making often requires more granular stratification based on tumor subtype or stage.

Future research should explore multimodal learning across additional modalities, including histopathological slides and molecular biomarker data. Integrating these dimensions would align AI diagnostics more closely with personalized oncology practices. Furthermore, interpretability methods should be incorporated to make model decisions transparent and clinically trustworthy.

6. Conclusion

In this study, we presented a multimodal deep learning framework that integrates Vision Transformers (ViT) for mammographic image analysis and BERT based models for clinical text interpretation to improve breast cancer detection. By combining visual and textual modalities, our approach mimics the diagnostic process of radiologists who rely on both imaging and clinical context to make accurate assessments.

Our experimental results, conducted on publicly available datasets such as CBIS-DDSM and MIMIC-CXR, demonstrate that the proposed multimodal model significantly outperforms unimodal baselines in terms of accuracy, precision, recall, F1-score, and AUROC. We further validated the contribution of each component through ablation studies and showed the robustness of early fusion strategies in unifying diverse feature representations.

This work highlights the potential of multimodal learning to enhance diagnostic accuracy, reduce uncertainty in clinical decision-making, and support radiologists in early detection of breast cancer. Future directions include expanding the model to handle other imaging modalities such as ultrasound or MRI, incorporating structured patient metadata (e.g., genetic profiles, family history), and deploying real-time explainability mechanisms to ensure clinical trust and adoption.

References

[1] Rebecca L Siegel, Kimberly D Miller, Hannah E Fuchs, and Ahmedin Jemal. Cancer statistics, 2023. *CA: a cancer journal for clinicians*, 73(1):17– 48, 2023.

[2] World Health Organization. Breast cancer, 2023. https://www.who.int/ news-room/fact-sheets/detail/breast-cancer.

[3] Lulu Sun, Ariel Wu, Gregory R Bean, Ian S Hagemann, and Chieh-Yu Lin. Molecular testing in breast cancer: Current status and future directions. *The Journal of Molecular Diagnostics*, 23(11):1422–1432, 2021.

[4] Shi Wei. Hormone receptors in breast cancer: An update on the uncommon subtypes. *Pathology-Research and Practice*, 250:154791, 2023.

[5] Swati Sucharita Mohanty, Chita Ranjan Sahoo, and Rabindra Nath Padhy. Role of hormone receptors and her2 as prospective molecular markers for breast cancer: An update. *Genes & diseases*, 9(3):648–658, 2022.

[6] Dmitriy Sonkin, Anish Thomas, and Beverly A Teicher. Cancer treatments: Past, present, and future. *Cancer Genetics*, 2024.

[7] Karen S Johnson, Emily F Conant, and Mary Scott Soo. Molecular subtypes of breast cancer: a review for breast radiologists. *Journal of Breast Imaging*, 3(1):12–24, 2021.

[8] Jong Seok Ahn, Sangwon Shin, Su-A Yang, Eun Kyung Park, Ki Hwan Kim, Soo Ick Cho, Chan-Young Ock, and Seokhwi Kim. Artificial intelligence in breast cancer diagnosis and personalized medicine. *Journal of Breast Cancer*, 26(5):405, 2023.

[9] Constance D Lehman, Rachel D Wellman, Diana S M Buist, Karla Kerlikowske, Anna N A Tosteson, and Diana L Miglioretti. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Internal Medicine*, 175(11):1828–1837, 2017.

[10] Zengan Huang, Xin Zhang, Yan Ju, Ge Zhang, Wanying Chang, Hongping Song, and Yi Gao. Explainable breast cancer molecular expression prediction using multi-task deep-learning based on 3d whole breast ultrasound. *Insights into Imaging*, 15(1):227, 2024.

[11] Mehdi Hazratifard, Fayez Gebali, and Mohammad Mamun. Using machine learning for dynamic authentication in telehealth: A tutorial. *Sensors*, 22(19):7655, 2022.

[12] Mohammad Zare, Mahdi Jampour, and Issa Rashid Farrokhi. A heuristic method for gray images pseudo coloring with histogram and rgb layers. In *2011 IEEE 3rd International Conference on Communication Software and Networks*, pages 524–527. IEEE, 2011.

[13] 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. *International Conference on Learning Representations (ICLR)*, 2021.

[14] A Noorian, A Harounabadi, and M Hazratifard. A sequential neural recommendation system exploiting bert and lstm on social media posts. *Complex & Intelligent Systems*, 10(1):721–744, 2024.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*, 2019.

[16] Fatima-Zahrae Nakach, Ali Idri, and Evgin Goceri. A comprehensive investigation of multimodal deep learning fusion strategies for breast cancer classification. *Artificial Intelligence Review*, 57(12):327, 2024.

[17] Gehad A Saleh, Nihal M Batouty, Abdelrahman Gamal, Ahmed Elnakib, Omar Hamdy, Ahmed Sharafeldeen, Ali Mahmoud, Mohammed Ghazal, Jawad Yousaf, Marah Alhalabi, et al. Impact of imaging biomarkers and ai on breast cancer management: a brief review. *Cancers*, 15(21):5216, 2023.

[18] Amr Soliman, Zaibo Li, and Anil V Parwani. Artificial intelligence's impact on breast cancer pathology: a literature review. *Diagnostic pathology*, 19(1):38, 2024.

[19] Kathryn Malherbe. Tumor microenvironment and the role of artificial intelligence in breast cancer detection and prognosis. *The American journal of pathology*, 191(8):1364–1373, 2021.

[20] Sahar Saki. Decoding the molecular foundations of breast cancer: A synthesis of artificial intelligence and personalized medicine insights. *International Journal of BioLife Sciences (IJBLS)*, 3(3):244–257, 2024.

[21] Beatriz Alejandra Bosques Palomo, Mario Alexis Monsivais Molina, Jorge Alberto Garza Abdala, Daly Betzabeth Avendano Avalos, Servando Cardona-Huerta, T Aaron Gulliver, and Jose Gerardo Tamez Pena. Performance evaluation of deep learning and transformer models using multimodal data for breast cancer classification. In *Cancer Prevention, Detection, and Intervention: Third MICCAI Workshop, CaPTion 2024, Held in Conjunction with MICCAI 2024, Marrakesh, Morocco, October 6, 2024, Proceedings*, volume 15199, page 59. Springer Nature, 2025.

[22] S¨oren Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. Multimodal deep learning for biomedical data fusion: a review. *Briefings in bioinformatics*, 23(2):bbab569, 2022.

[23] Xiang Li, Yuwei Duan, and Wei Fan. Gated attention multimodal learning for breast cancer classification using clinical and genomic data. *BMC Bioinformatics*, 24(1), 2023.

[24] Yunkun Zhang, Jin Gao, Zheling Tan, Lingfeng Zhou, Kexin Ding, Mu Zhou, Shaoting Zhang, and Dequan Wang. Data-centric foundation models in computational healthcare: A survey. *arXiv preprint arXiv:2401.02458*, 2024.

[25] Andreas Klein, Hannah Richter, and Lara Schulte. Ai integration in nationwide breast cancer screening in germany increases detection rates. *European Radiology*, 2024.

[26] Honglei Liu, Zhiqiang Zhang, Yan Xu, Ni Wang, Yanqun Huang, Zhenghan Yang, Rui Jiang, and Hui Chen. Use of bert (bidirectional encoder representations from transformers)-based deep learning method for extracting evidences in chinese radiology reports: development of a computer-aided liver cancer diagnosis framework. *Journal of medical Internet research*, 23(1):e19689, 2021.

[27] Omneya Attallah, Fatma Anwar, Nagia M Ghanem, and Mohamed A Ismail. Histo-cadx: duo cascaded fusion stages for breast cancer diagnosis from histopathological images. *PeerJ Computer Science*, 7:e493, 2021.

[28] Yoonjae Cho, Sampa Misra, Ravi Managuli, Richard G Barr, Jeongmin Lee, and Chulhong Kim. Attention-based fusion network for breast cancer segmentation and classification using multi-modal ultrasound images. *Ultrasound in Medicine & Biology*, 51(3):568–577, 2025.

[29] Yiqiu Shen, Jungkyu Park, Frank Yeung, Eliana Goldberg, Laura Heacock, Farah Shamout, and Krzysztof J Geras. Leveraging transformers to improve breast cancer classification and risk assessment with multi-modal and longitudinal data. *arXiv preprint arXiv:2311.03217*, 2023.

[30] Liangrui Pan, Zhenyu Zhao, Ying Lu, Kewei Tang, Liyong Fu, Qingchun Liang, and Shaoliang Peng. Opportunities and challenges in the application of large artificial intelligence models in radiology. *Meta-Radiology*, page 100080, 2024.

[31] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.

[32] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature Medicine*, 25:24–29, 2021.

[33] Tiantian Feng, Digbalay Bose, Tuo Zhang, Rajat Hebbar, Anil Ramakrishna, Rahul Gupta, Mi Zhang, Salman Avestimehr, and Shrikanth Narayanan. Fedmultimodal: A benchmark for multimodal federated learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4035–4045, 2023.

[34] Han Zhang, Peng Chen, Xinyang Fan, Yong Lei, Tao Liu, Ying Wang, and Xintao Hu. Multimodal fusion of mri and clinical data improves diagnosis of alzheimer's disease. *Frontiers in Neuroscience*, 16:842446, 2022.

[35] Tadas Baltruˇsaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019.

[36] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[37] Jieneng Chen, Yuyin Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transmed: Transformers advance multi-modal medical image classification. *Diagnostics*, 12(2):277, 2022.

[38] Wazir Muhammad, Manoj Gupta, and Zuhaibuddin Bhutto. Role of deep learning in medical image super-resolution. In *Principles and Methods of Explainable Artificial Intelligence in Healthcare*, pages 55–93. IGI Global, 2022.

[39] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[41] Douwe Kiela, Jannis Bulian, Guillaume Lample, Thomas Scialom, and Sebastian Riedel. The dead are alive: Exploring the surprising performance of multimodal models trained with incomplete modalities. *EMNLP*, 2021.

[42] Nathalie Japkowicz and Shaju M Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–449, 2002.

[43] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*, 2019.

[44] Ross Wightman. Pytorch image models, 2019. https://github.com/ rwightman/pytorch-image-models.

[45] Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexander A Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2):125, 2020.

[46] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R´emi Louf, Morgan Funtowicz, and Jamie Brew. Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020.

[47] Richard S Lee, Fernando Gimenez, Assaf Hoogi, Kevin K Miyake, Michael Gorovoy, and Daniel L Rubin. A curated mammography dataset for training and evaluation of breast cancer cad systems. *Scientific Data*, 4:170177, 2017.

[48] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr: A large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.