Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: jeeemi.org; Vol. 7, No. 4, October 2025, pp: 1320-1339 e-ISSN: 2656-8632

RESEARCH ARTICLE OPEN ACCESS

Vision Language Transformer Framework for Efficient Cancer Diagnosis through Multimodal Integration

Bala Gangadhara Gutam and Sunil Kumar Malchi

Department of Computer Science and Engineering, Mohan Babu University, Tirupati, India

Corresponding author: Bala Gangadhara Gutam (e-mail: balabgangadhar@gmail.com), Author(s): Sunil Kumar Malchi (e-mail: sunilmalchi1@gmail.com)

Abstract **Finding and treating cancer as soon as possible help patients get better outcomes. Patients requiring imaging or biopsy tests sometimes find it challenging to access them because these procedures are often limited by their high cost and availability in clinical settings. Recent AI methods, particularly those involving deep learning, can address these problems and significantly enhance the process for detecting cancer, offering greater efficiency and scalability. In this context, LLMs and VLMs are considered leading solutions for trying to make sense of multimodal variables within AI-driven healthcare systems. Although LLMs are strong at working with unstructured, clinically related text data, they have not often been used for patient assessment beyond descriptive or summarization tasks, by combining images and descriptions, along with both structured and unstructured data. The VLMs allow doctors and medical researchers to catch cancer symptoms from multiple angles. In this work, we study both LLMs and VLMs in cancer detection, analyzing their architectures, learning mechanisms, and performance on various datasets, and identifying directions for expanding multimodal AI in healthcare. Our results indicate that combining these two data types enhances how accurately we are able to diagnose patients across different types of cancer. Our studies in MIMIC-III, MIMIC-IV, TCGA, and CAMELYON 16/17 datasets revealed that multimodal transformer models significantly improve the accuracy of diagnosing biopsy results. In particular, BioViL achieves an AUC-ROC of 0.92 for detecting lung cancer, whereas CLIP Fine-tuned achieves a similar result of 0.91 for colon cancer detection.**

Keywords **Cancer detection; Vision-Language Models; Large Language Models; Transformers; Clinical data, Histopathology; Medical Imaging; Multimodal AI.**

I. Introduction

Cancer remains one of the leading causes of mortality worldwide, with an estimated 19.3 million new cases and almost 10 million deaths in 2020 (as per the International Agency for Research on Cancer (IARC) and the World Health Organization (WHO) [1]. It is estimated that these numbers will be increased by nearly half in 2040, which not only make cancer one of the most important health-related issues but also a significant economic and social burden. Cancer care is estimated to cost the economy USD 1.16 trillion each year, with inconsistent effects on the low- and middle-income nations, where resources to aid diagnosis are limited [2]. Timely and accurate diagnosis is one of the most effective measures to increase survival rates, optimize treatment courses, and decrease the number of financial and emotional costs imposed on patients and medical services. Conventional diagnostic methods remain incomplete despite decades of investigation. Radiological imaging, like CT, MRI, PET, and mammography, is capable of essential and practical evaluation of tumors but may not work in the case of early tumor detection, and also highly relies on office interpretation [4]. The efficient standard is considered to be histopathological biopsies, which are aggressive and prone to inter-observer variation [5]. Biomarker testing provides molecular information that is not globally applicable to all cancer subclasses [6]. The challenges make it apparent that the pressing need in the diagnostic solutions is for them to be scalable, available, and capable of combining various sources of clinical data. Artificial Intelligence (AI) is a recent strategy in the healthcare sector, as it suggests advanced approaches to computerize and improve cancer diagnostics [39]. Although the initial applications of Deep Learning (DL) show encouraging outcomes in unimodal tasks, such as skin cancer and histopathology slide analysis [7] [8], the systems are unimodal and do not provide the broadened capabilities of AI due to their limited scalability on large and heterogeneous datasets [9]. These innovations create the possibility of combining medical text and images, among other forms of data, in structures for cancer detection.

Manuscript received July 8, 2025; Revised October 20, 2025; Accepted October 25, 2025; date of publication October 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i2.652

Though, regardless of these developments, there are still a number of challenges. First, the generalization of most multimodal AI systems is incorrect because cancer data in different institutions, patient groups, and imaging modalities are heterogeneous [23]. Second, the affected data silos are because medical datasets are disjointed and constrained by privacy laws, which hamper large-scale training [24]. Third, a significant obstacle is interpretability; models are very diagnostic, but clinical practice normally requires clear decision-making mechanisms to develop trust and implement them in clinical settings [25]. Additionally, scalability is also an essential issue, as transformer-based models [21] can be very resource-intensive and challenging to deploy in healthcare systems with limited resources [26]. Finally, such ethical and regulatory concerns as bias, fairness, and patient privacy are still in the process of clinical integration, and the question of accountability with safety in practical execution [27, 28].

This paper makes the following contributions:

1. Evaluating transformer-based LLMs for cancer detection in clinical texts.
2. Evaluating VLMs for multimodal (text + image) cancer diagnosis.
3. Analyzing the diagnostic of Vision Language integration compared to unimodal methods.
4. Discussing challenges such as model interpretability, multimodal fusion, and clinical deployment.

In particular, we examine representative models including BioGPT-VL, LLaVA-Med, and RadFM, demonstrate their capabilities for cancer classification, report generation, and early disease detection [5]. We also highlight that these models overcome some limitations of traditional methods while acknowledging open challenges such as data heterogeneity, generalization, and clinical standardization [10]. This paper intends to prove the transformer power of multimodal AI systems in the diagnosis of cancer by critically analyzing the state of the art LLMs and VLMs. The work does not only benchmark the existing models but also outlines the existing gaps in the work, which need to be filled in to achieve adequate, safe, reliable, and fair implementation of these technologies in the real-life oncology environment.

II. Related Work

The most current developments in AI technology have created a new horizon in cancer detection, especially in the case of the combined use of VLMs [23] and transformer-based LLMs [1]. The conventional diagnosis methods, although they are significant, are usually limited in terms of their sensitivity, reliability, and applicability. Joint processing of medical images and clinical text, with the help of multimodal AI systems, allows for a more effective and timely diagnosis of cancer [36]. VLMs bridge the gap between visual information and textual expertise, while LLMs enhance the interpretation of complex medical narratives. In combination, the technologies transform the game of oncological diagnostics with powerful, situation-aware, and automated decision support.

A. Traditional Cancer Detection Methods

Medical imaging, histopathology, and biomarker tests form the foundation of cancer diagnosis, which is inherently limited. Imaging modalities, such as CT, MRI, PET, and mammography, provide invaluable spatial information but often fail in the early detection of tumors, requiring expert interpretation [30]. The culture histopathological method is regarded as an important standard. However, it is invasive, time-consuming, and prone to inter-observer variability [13] [28]. The test of biomarkers (e.g., PSA, HER2, and CA-125) also provides information at the molecular level but is not universally applicable across different types of cancer [30]. The initial deep learning work, such as that by Esteva et al. [15], had already shown that dermatologist-level accuracy in skin cancer classification is possible with images, and Coudray et al. [12] and Campanella et al. [13] had provided evidence that CNNs can be used with histopathology slides. Nevertheless, these unimodal approaches cannot describe the multifaceted nature of cancer data.

1. Imaging-Based Diagnostics

Today, visual tests called imaging are the main tool health professionals use to identify cancer. To identify irregular cell growths and measure the progress of a disease, doctors utilize various imaging technologies, including CT, MRI, ultrasound, PET, and mammography. Doctors and specialists use CT scans and mammograms to find lung cancer and colorectal cancer and to identify breast cancer [30]. Despite requiring radiologists' knowledge, these techniques commonly have difficulty finding tumors that are just starting or are small.

2. Histopathology

A biopsy is necessary for pathologists to examine the tissue under a microscope as part of regular histopathological analysis. Though direct microscopy, doctors can check for cell problems at the same time as they spot, classify, and identify the main forms of each type of tumor. Although it takes time and is considered the most reliable method, cancer diagnosis [1] [3] through biopsy relies heavily on the expertise of pathologists. The differences in decisions between experts make it hard to reach consistent assessments and outcomes [4] [5] in many cases.

3. Biomarker Testing

Doctors can identify cancer and track its advancement through biomarker examination of blood tests along with urine collection or tissue samples using proteins

Manuscript received July 8, 2025; Revised October 20, 2025; Accepted October 25, 2025; date of publication October 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i2.652

and hormones together with genetic mutations. Three widespread biomarkers used to detect cancer exist within the medical field: PSA (Prostate-Specific Antigen) for prostatic malignancies and CA-125 for ovarian cancer, together with HER2 expression analysis for breast cancer pathology [1]. Many useful biomarkers fail to produce either precise detection methods or sensitive detection capability which results in incorrect positive readings and early cancer identification failures. Not all cancers possess established biomarkers that receive universal acceptance by the medical community. As shown in the Fig. 1, the pipeline of the end-to-end deployment of pre-trained medical VLMs in clinical decision support. It is provoked by the pre-training on the multimodal clinical data (images, text, and patient records) and proceeds with the assessment of performance on specific datasets. This is followed by optimization, contrastive learning [9], and modular fine-tuning strategies so as to achieve the best accuracy that the model can attain. Model evaluation is a combination of human expert evaluation and quantitative evaluation (e.g., BLEU, ROUGE) [31]. Finally, the model would be deployed into the clinical pathways, generating actionable outputs tailored to patient context and physician orders.

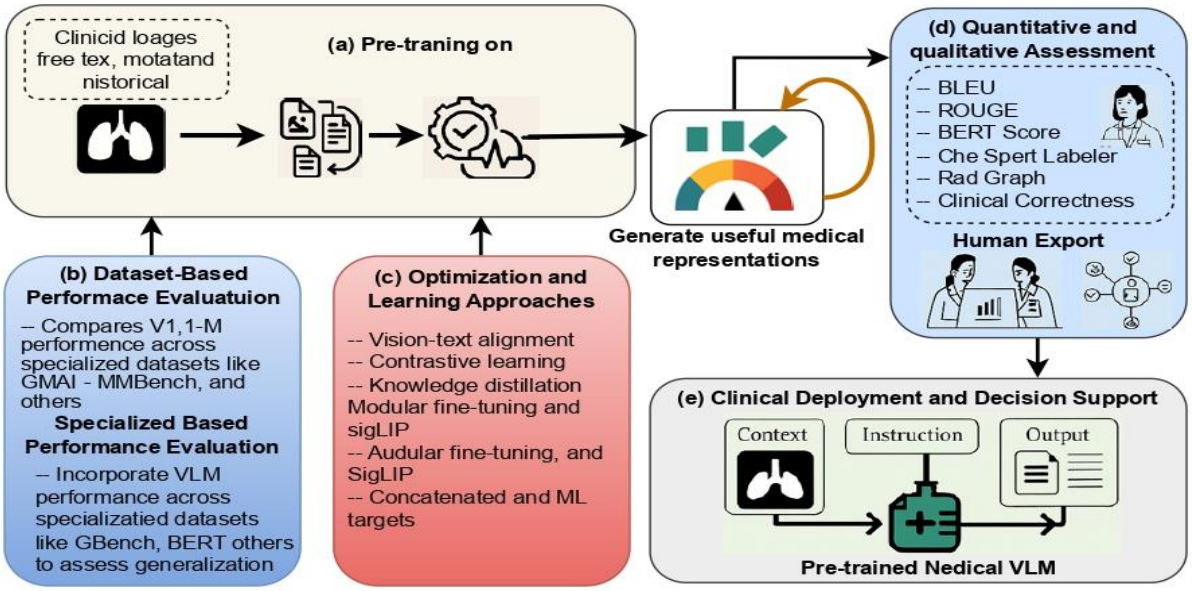


Fig. 1. Conceptual overview of the VLM pipeline for cancer diagnosis. (a) Pre-training using multimodal clinical inputs. (b) Evaluation on benchmark datasets (e.g., TCGA, MIMIC). (c) Model optimization strategies including contrastive and supervised learning. (d) Quantitative and qualitative performance evaluation (e.g., AUC, Grad-CAM maps). (e) Clinical deployment for diagnostic support.

B. Natural Language Processing in Healthcare

Electronic Health Records (EHRs), discharge summaries, and pathology notes have been analyzed to a wide extent using Natural Language Processing (NLP). The transformer-based models, such as BioBERT [2], ClinicalBERT [3], and PubMedBERT, largely improved text mining in biomedicine by relying on domain-specific corpora such as PubMed and MIMIC datasets [16] [17]. Peng et al. [4] showed that transfer learning is effective in biomedical NLP. These models help in extracting, detecting relations in, and automated coding of data, and are used in oncology, such as cancer report type and adverse event detection. Recent reviewers [29] highlight the application of NLP in the pipeline of multimodal oncology, especially in conjunction with imaging data.

1. Applications of NLP in Clinical Settings

In the healthcare field, NLP assists in supporting various operations through its diverse applications. It enables the identification of important medical entities, such as diseases, medications, procedures, and symptoms, from clinical narratives. Then provide important patient information that suggests diagnoses and indicates prohibited use cases for medical personnel. In addition to its clinical function, NLP can also be used to process free-text medical reports into administrative ICD or SNOMED codes, which are necessary for reimbursement, along with administrative processing. Through the history of patients, the health specialists evaluate the risks of disease development chances and detect possible healthcare complications. Moreover, a hospital report integrated with the public health surveillance system can support the real-time detection of diseases, as well as adverse drug reactions. Then improving the care of patients and overall awareness of the healthcare system.

2. Advancements in Contextual NLP

Manuscript received July 8, 2025; Revised October 20, 2025; Accepted October 25, 2025; date of publication October 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i2.652

Clinical NLP underwent diverse transformation through deep learning [12] and contextual embedding, which led to the development of word2vec, followed by GloVe and transformer-based models that include BERT [1]. These models read semantic meaning from text context therefore they function at a higher performance level with complex clinical texts. The pre-training process of medical literature and Electronic Health Records on BioBERT [2], ClinicalBERT [3], and Blue-BERT leads to improved domain accuracy. Modern healthcare models show effectiveness in recognizing medical conditions and drug-illness relationships in addition to their ability to condense detailed clinical documentation. Interpretation across the entire healthcare system becomes accessible because enhanced medical terminology understanding links ontologies between UMLS, RxNorm and SNOMED CT.

C. Large Language Models in Medicine

GPT-4, BioGPT, PubMedBERT, and MedAlpaca are other LLMs that transform the field of text in medicine by allowing zero-shot and few-shot reasoning mechanisms [5] [24] [25]. BioGPT (2022) demonstrated its current performance in medical QA and cancer-specific entity extraction, while the Med-PaLM (2023) platform expanded instruction tuning to meet medical reasoning standards. The MedAlpaca (2023) system became an open-source and lightweight medical LLM for clinical tasks. The scalable DL on EHRs introduced by Rajkomar et al. [5] earlier formed the basis of the present-day applications of LLM. Gupta and Lin [34] also raised difficulties, and it was demonstrated that even when presented with false assumptions when asked a question by a patient, LLMs do not produce accurate reactions. Although LLMs are powerful in medical text analysis, they can, by design, only be unimodal, which requires multimodal extensions. The dedical training of these models requires broad biomedical data retrieved from PubMed sources combined with MIMIC-III [16] records and clinical trial data until they develop specialization for particular medical applications. As a result, they excel in following medical texts, enabling us to find diagnostic and therapeutic agents, genetic information, and pharmaceutical substances in clinical texts through Named Entity Recognition (NER). In addition, relation extraction methods reveal meaningful relationships, such as drug-disease, gene-disease, or treatment response relationships, and enhance clinical awareness.

The NLP system is used to triage patient messages with categorized diagnostic regions in written records and automated, summarized medical reports, and answer questions to address patient needs. Additionally, medical system agents utilize past information processing with NLP algorithms to calculate the likelihood of disease comorbidity, including the association between sepsis, diabetes, and the risk of cancer, which can provide predictive analysis and inform proactive care. GPT-4 achieves effective processing of extensive medical documents to help doctors generate diagnostic information and treatment sequences using the few-shot learning approach. Specialized training for general purpose LLMs becomes necessary because healthcare language includes complex structures and high risk operational zones. The implementation of GPT-4 for clinical use faces three main implementation hurdles because of its hallucinatory behavior together with regulatory requirements and explainable system expectations. Fig. 1 is a Comprehensive Multimodal cancer Detection [32] Pipeline whereby a text and image [29] encoder (e.g., BioBERT, ClinicalBERT, and ResNet-50) is incorporated to produce joined representations.

Fig. 1 task of training such an embedding is to categorize the cancer with contrastive and cross-entropy loss. It also shows the key clinical issues, lying in the heterogeneity of information, interpretability, and the necessity to get models that could be explained, which could be further used as Grad-CAM, to help in early detection and evidence-based clinical decision making. The system fuses image features (e.g., histopathology or radiographs) and textual reports (e.g., pathology or clinical notes) through a dual-encoder architecture followed by joint reasoning via cross-attention or contrastive [40] learning.

D. Vision Language Models In Medicine

Vision-Language Models (VLMs) are a type of text-visual model that allows joint reasoning on radiology images, histopathology slides, and patient notes (Fig. 2). Computer-assisted image-text CLIP [6] has been adapted into medical practices. Its domain extension, MedCLIP [8], enhanced zero-shot classification and image report retrieval error. GLoRIA [7] employed hierarchical multimodal learning by mapping textual clinical concepts to local representative image regions, whereas BioViL [11] maximized radiology-specific vision-language pre-training. More recently, CHIEF (2024) reached state-of-the-art pan-cancer detection with an AUC ~0.94 and showed that multimodal fusion can be used to achieve state-of-the-art [24]. A transformer framework utilizing multimodal oncology, in addition to imaging and text, was introduced by Cai et al. [24], known as DeePathNet, emphasizing the significance of multimodal data integration with pathways in cancer. However, challenges persist. Vision-language systems are computationally expensive and are usually trained on biased datasets. Multimodal oncology frameworks were reviewed by Yang et al. [28] and Waqas et al. [29], with the focus on

Manuscript received July 8, 2025; Revised October 20, 2025; Accepted October 25, 2025; date of publication October 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i2.652

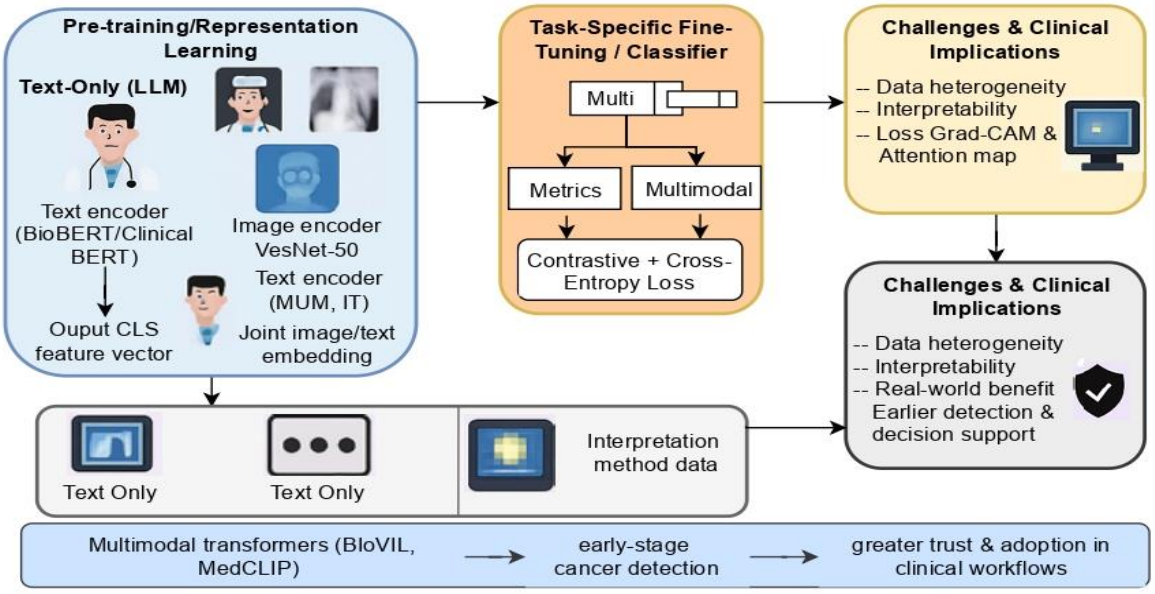


Fig. 2. Overview of the multimodal cancer detection system based on vision-language models (VLMs).

The heterogeneity of the data, its interpretability, and equity. Yang et al. [36] identified the problem of bias, as they discovered that VLM outputs had demographic differences, whereas Clusmann et al. [35] established that oncology-focused VLM had prompt injection vulnerability. These articles point to the fact that VLMs can be highly diagnostic, but such aspects as robustness, explainability, and clinical safety are still problematic.

1. Popular VLMs in medicine

Trained initially on large-scale natural image to text pairs, Contrastive Language Image Pre-training (CLIP) [6] has been reused for clinical work, but it has interpretability challenges. GLoRIA [7] uses hierarchical VLM multimodal learning to connect textual clinical entities to attend to specific image regions and provide a localization in radiology. BioViL [11] has radiology-specific vision language pre-trained models designed and trained with contrastive and masked language objectives. MedCLIP [8] is used in the medical field as an extension of CLIP with strong retrieval and classification in multiple modalities, with high computational expense. Together, these models highlight the advantages and limitations of multimodal AI in the medical field, particularly in terms of generation, interpretability, and training efficiency.

2. These models currently maintain the best available performance levels

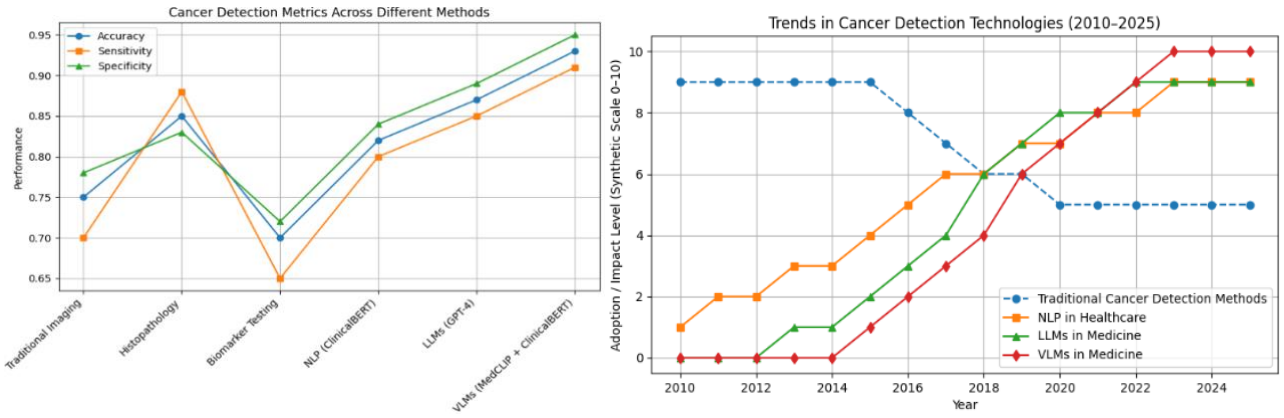
The image report retrieval is one of the tasks, as clinical images are compared with the most suitable textual reports and vice versa. Zero-shot classification is another key feature that enables the diagnosis of disease types by using a few training samples with the help of textual prompts. VLMs can also help in localization activities using weakly labeled data to identify areas of interest, such as tumors or lesions. Finally, with the advantage of classification and localization, these models automatically produce natural language reports or summaries directly on medical images, which help clinicians provide quick, accurate, and readable records. While these models highlight progress in Vision Language Learning for medicine, they are most important when evaluated in isolation, without direct comparison across cancer-specific tasks. CLIP, though powerful in zero-shot learning, struggles in clinical interpretability. BioViL achieves strong alignment but remains restricted to radiology. MedCLIP generalizes better across domains but demands extensive medical pre-training, while GLoRIA improves region-level grounding but is computationally expensive. More recent multimodal frameworks such as CHIEF (2024) extend beyond radiology to pan-cancer detection, achieving ~0.94 AUC, but challenges in interpretability and validation remain. As shown in Fig. 3 (a) and (b), the advancement in detection capabilities and the shift of classic single modality techniques to the modern multimodal paradigms are illustrated.

E. Cancer detection with NLP and VLMs

Recent surveys provide methodical insights into multimodal AI in oncology. Nakach et al. [30] reported on deep learning fusion methods in breast cancer, whereas Gao et al. [31] introduced an explainable framework for fusion methods in predicting therapy response. Patel et al. [33] have focused on cross-attention transformers to detect anomalies in medical imaging, but Waqas et al. [29] and Yang et al. [28] have

Manuscript received July 8, 2025; Revised October 20, 2025; Accepted October 25, 2025; date of publication October 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i2.652

placed emphasis on multimodal data integration in oncology settings. These works taken together create a solid base, yet they also show the necessity to make comparative benchmarking of VLMs and LLMs across the types of cancer, which inspires the current study. Text-based NLP processing analyzes pathology reports to detect tumor grades and extract biomarkers, while the medical imaging CNN [12] identifies suspicious regions in mammograms and histopathology slides. To jointly analyze the clinical narratives and imaging data, multimodal models achieve stronger diagnostic performance. For instance, the mammograms paired with radiology reports then improve breast cancer localization. The CT scans with patient notes enable earlier lung cancer detection, and colonoscopy or histology images with textual annotations enhance colorectal and prostate cancer classification. The pipelines, such as MedCLIP + ClinicalNERT, have demonstrated superior sensitivity and specificity in Table 1 compared to single-mode baselines. Nevertheless, the challenges remain around interpretability, dataset generalization, and real-world clinical validation, which must be addressed before widespread deployment.



(a) (b)

Fig. 3. (a) Comparative AUC-ROC performance across different cancer detection methods (e.g., TF-IDF, BioBERT, BioViL, CLIP). (b) Evolution of cancer detection technologies from traditional single-modality methods to modern multimodal transformer-based systems (2010–2025).

III. Methodology

In this section, the mathematical context of the intended framework to identify cancer with the help of LLMs and VLMs will be outlined. To formulate the defined task, both unimodal and multimodal representations are presented, and the fusion strategy defines the learning goals and assessment measures.

A. Problem definition

This study addresses two fundamental challenges in cancer detection that directly affect clinical decision-making design requirements. Formally, cancer diagnosis is a multimodal classification problem, with the medical images and clinical text being mapped to a cancer class in Eq. (1) [21] [22].

$$f = (X_{img}, X_{txt}) \rightarrow \mathcal{Y} \quad (1)$$

Where X_{img} an input is medical images (CT, MRI, histopathology), X_{txt} is clinical textual data (EHRs, pathology/radiology reports), $\mathcal{Y} \in \{0, 1, \ldots, K$ is cancer class label.

1. Text-only Cancer Detection

The electronic health records, pathology report, and radiology outcomes are clinical descriptions that provide fundamental diagnostic data. However, the texts are typically ambiguous, inconsistent, counterintuitive, and filled with unnecessary data. The task is to categorize specific clinical notes as a cancer diagnosis if they identify the specific cancer type, including lung, breast, or colon. Radiology records that contain a detected quantity or a hazy object are typically a sign of lung cancer. The reports of pathology, which state that the carcinoma is HER2 positive, are usually related to breast cancer, whereas the discharge summaries referring to adenocarcinoma of the colon point toward colorectal cancer [41]. We formulate this as either a binary (cancer vs. non-cancer) or a multi-class classification problem. The model processes the input sequence of tokens $T = \{t_1, t_{2,..} t_n\}$, the transformer encoder produces contextual embeddings, as defined in Eq. (2) [1].

$$H = Transformer(T) \in \mathbb{R}^{n \times d} \quad (2)$$

Where H is the output matrix of the transformer, $\mathbb{R}$ is set of real number, n and d are number of tokens in input sequence and the embedding dimension, $\mathbb{R}^{n \times d}$ is a matrix with n rows and d columns. In order to get the global text representation, the [CLS] token embeddings are used to aggregates the clinical report in Eq. (3) [2] [3].

$$h_{txt} = H_{[CLS]} \quad (3)$$

Manuscript received July 8, 2025; Revised October 20, 2025; Accepted October 25, 2025; date of publication October 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i2.652

where h_{txt} is a text feature vector representation of the whole clinical report, then embedding with $H_{[CLS]}$ token in the transformer model used whole sequence. It applies a classification head with softmax to generate class probabilities. This approach reduces variability in interpretation and provides standardized outputs across diverse clinical texts. The global embedding is offered by a softmax layer through a classification head that generates class probabilities, as defined in Eq. (4) [2] [4].

radiology reports connected to the chest X-rays increased the interpretability and consistency of lung cancer screening. This task requires VLMs capable of learning cross-modal relationships [22] and exploiting complementary signals. This framework employs contrastive loss to maximize alignment between correctly paired image and text embedding while penalizing mismatches. This ensures that, for example, a lung CT scan is most similar to its associated radiology note, improving diagnostic robustness. The

Table 1. Comparison of Foundational Vision Language Models and Recent Advances (2020–2024).

Model	Modality Focus	Key Application	Reported Performance	Main Limitation
CLIP	General images + text	Zero-shot retrieval, prompt-based classification	Strong in zero-shot (~80% retrieval accuracy on generic tasks)	Poor clinical interpretability; not trained on medical data
BioViL	Radiology images + reports	Report-image alignment, chest X-ray retrieval	~0.86 AUC (radiology retrieval tasks)	Limited to radiology, lacks histopathology/generalization
MedCLIP	Medical images + reports	Image-text retrieval, few-shot diagnosis	Competitive retrieval, stronger than CLIP in medicine	Requires large-scale medical pre-training, limited explainability
GLoRIA	Entity–region grounding	Localizing clinical terms in images	Accurate grounding of pathology terms	Computationally intensive; small datasets
CHIEF (2024)	Multimodal (pathology + clinical text + radiology)	Pan-cancer detection, subtype classification	~0.94 AUC across multiple cancer types	Early-stage model; limited clinical deployment and interpretability

$$P\left((\mathcal{Y}|X_{txt})\right) = Softmax(W_t.h_{txt} + b_t) \quad (4)$$

Where h_{txt} is [CLS] token embedding from BioBERT/ ClinicalBERT, W_t is the weight matrix of the classification head, b_t is a bias vector of the classification head, and $\mathcal{Y}$ is a probability distribution over cancer classes.

2. Multimodal Cancer Detection

Text alone captures semantic descriptions, but imaging modalities such as histopathology slides, CT scans, and radiographs provide spatial and morphological details critical for cancer subtype [15] identification. The important integration of both modalities is crucial for ensuring decent diagnostic performance for various medical tasks. In the case of early tumor detection using lung CT scans with corresponding textual descriptions, more accurate nodules can be identified. The integration of breast cancer histopathology slides and pathology notes enhances the evaluation of HER2 status, resulting in improved efficiency in subtype classification. Similarly, during the workflow alignment, image and text multimodal embeddings are aligned with the help of the contrastive loss function defined in Eq. (5) [19].

$$H_{img} = ViT(I) \in \mathbb{R}^{m \times d} \quad (5)$$

Where H_{img} is a matrix, where each row is the embedding of one image patch, m and d number of patches and embedding dimension, I is input medical image, $ViT(I)$ is breaks image into small patches and encodes them. To establish effective alignment between the embeddings of paired images and texts, which use InfoNCE contrastive loss, is able to maximize the similarity among correctly paired embeddings and to minimize similarity between unmatched pairs, which improves multimodal discrimination ability. The image and text multimodal embeddings are further optimized with contrastive loss function defined in Eq. (6) [19] [40].

$$\mathcal{L}_{InfoNCE} = -\sum_{i=1}^{N} \log \frac{\left(sim(z_i,t_i)/_{\mathcal{T}}\right)}{\sum_{j=1}^{N} exp\left(sim(z_i,t_j)/_{\mathcal{T}}\right)} \quad (6)$$

Manuscript received July 8, 2025; Revised October 20, 2025; Accepted October 25, 2025; date of publication October 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i2.652

where z_i and t_i are image and text embeddings, $sim(.)$ is cosine similarity, and $\mathcal{T}$ is a temperature scaling factor.

B. Datasets and Preprocessing

To build and validate both the unimodal (text-only, image-only) and multimodal (image + text) models for diagnostic purposes in cancer, we curated and pre-processed data sets from many publicly available and institutionally derived repositories. The datasets range through structured clinical reports and raw medical imaging to paired multimodal entries [18]. This section describes their sources, the way they are structured, and preprocessing procedures, as well as their complications. The multiple multimodal cancer detection benchmark datasets. In order to make it reproducible, all preprocessing steps are fully explained in Table 2, MIMIC-III/IV (EHR and Clinical Notes) discharge summaries, pathology notes, and radiology reports are available. Text data were pre-processed by performing token normalization with UMLS, negation identification with NegEx, and section division (e.g., introduction, finding, impression, history of present illness). TCGA (The Cancer Genome Atlas) whole Slide Images (WSIs) were divided into overlapping 256 - 256 patches (20 percent), and the color variation was reduced by application of Macenko stain normalization.

The class imbalance between the tumor subtypes was addressed by patch-wise augmentation on the minority class towards rotation, flipping operations, and Gaussian noise addition. CAMELYON16/17 contains metastasis WSIs of lymph nodes with slide-level and region-level annotations. Severe class imbalance was compensated with the Synthetic Minority Oversampling Technique (SMOTE) at the patch level to overcome the underrepresentation of cancer-positive samples. MIMIC-CXR (Radiology Reports + Images) multimodal with radiology reports alongside the associated chest radiographs. All the images were resized to the same dimension of 224 x 224, and metadata (such as patient age, sex, and admission details) were also standardized into the tabular form to enable downstream modeling. In this case, histopathology slides, we applied the Macenko normalization of the stain to minimize the color variation across slides. This transformation standardizes an intensity and contrast in relation to images, ensuring consistent representation during training, and is expressed in Eq. (7) [20].

$$I_{norm} = \alpha.(I - \mu).\Sigma^{-1} + \beta \quad (7)$$

Where I is the raw histopathology image, μ, Σ are the mean and covariance of the stain vectors, and β control contrast and intensity alignment.

1. Clinical text datasets

To obtain accurate multimodal cancer results, we brought together a variety of both public and our private image and text datasets. MIMIC-III [16] and MIMIC-IV [17] were the main sources I used for finding clinical texts that contained discharge summaries and notes from radiology and pathology. They were also supported by artificial reports and EHRs that included only selected information, as well as EHRs from hospitals approved by the IRB. The note listed parts of the illness, including results from biopsy, CT, or MRI, and current medical history. For cancer-related interests, the experts selected the data using ICD-9/ICD-10 codes and, where necessary, also checked them by manually annotating the same information. Cancer and non-cancer can be classified using different schemas, as well as different types of cancer. To verify and review all issues where texts are vague, different terms are included, negative statements appear, and timing references are included. To do this, normalize UMLS, seek out negations with NegEx, and leverage the token rules from SciBERT and ClinicalBERT for all medical-related information.

2. Image datasets

The Cancer Genome Atlas (TCGA) [24], CAMELYON16/17, and data collected in clinical studies were among the sources of imaging data. The images included histopathology slides, CT tests, MRI scans, and PET scans. Since whole slide images (WSI) [13] are huge and reach gigapixel sizes, flow cytometry involves preprocessing the images using Otsu threshold, collecting patches that overlap to preserve continuity in the image, and applying the Macenko or Reinhard methods to lessen color differences. Often, all annotations were at the level of a whole slide or just a segmentation, apart from limited spatial labels. Images from radiology were standardized and resized to allow their use with ResNet and ViT models.

3. Multimodal datasets

MIMIC-CXR and TCGA, together with their clinical and diagnostic details, were integrated to enable multimodal learning. Again, the methods they used were single tests that led to reports as well as whole groups of reports from samples with similar patterns. To the pipeline, all the recording pairs were organized and tagged accurately, and any missing recordings were eliminated. I tried different approaches to decrease overfitting and boost the model's results when presented with fresh data. I had to translate many pieces into other languages and look for synonyms to use throughout the writing, along with other changes. The new features I integrated are options to resize, control brightness, include artifacts like noise, and apply different effects often found in standardized histopathology, like scattering of normal tissue and blurring. Ensuring the data was preprocessed and augmented made it possible to create a multimodal

Manuscript received July 8, 2025; Revised October 20, 2025; Accepted October 25, 2025; date of publication October 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i2.652

Table 2. The Datasets and Preprocessing Pipelines for Cancer Detection

Dataset	Modality	Preprocessing Steps	Access Link
MIMIC-III/IV	Clinical text (EHRs, discharge summaries, pathology reports)	Token normalization (UMLS), negation handling (NegEx), section splitting (Findings, Impressions, HPI)	https://physionet.org
TCGA	Whole Slide Images (WSIs)	WSIs segmented into 256×256 patches with 20% overlap; Macenko stain normalization; augmentation for minority tumor classes	https://portal.gdc.cancer.gov
CAMELYON16/17	Histopathology WSIs (lymph node metastasis)	Slide- and region-level annotations; class imbalance mitigated with SMOTE oversampling	https://camelyon17.grand-challenge.org
MIMIC-CXR	Radiographs + paired radiology reports	Radiographs resized to 224×224; structured metadata extraction (age, sex, admission details)	https://physionet.org

collection meeting all regulatory requirements and supporting cancer diagnosis and classification.To prepare the models, to arranged them into two groups based on whether they handled text or text and images.

C. Model architecture

Two models were built for each data type: text-only and multimodal (text + image).

1. LLM Architecture (Text-Only)

We trained BioBERT and ClinicalBERT which are transformer-based models developed for use in medicine, to recognize cancer simply by reading text. They rely on many layers of transformer encoders, along with self-attention and positional encoding, to underline the connections between the various parts of the text. The classification head combines all the parts of the text based on the CLS embedding and then sends them through a fully connected layer to determine if cancer is present in the text. They can perform well, as they are familiar with most biomedical terms and can handle lengthy, disorganized patient records, even when human involvement is not required. Fig. 4 shows the comparative view of two model structures of cancer detection. The left one is LLM Architecture (Text Only), which accepts textual clinical data, passes it through a classification head and CLS embedding, and ultimately makes a prediction stating 'Cancer: No.' On the other hand, the VLM Architecture (Text + Image) on the right hand side uses both the visual data (e.g., chest X-rays) and textual data. It uses patch embedding [19] and a transformer module to make a better prediction, and the answer comes out as Cancer: Yes. It explains the advantage of multimodal learning to make the diagnosis more accurate. The system integrates a text encoder (e.g., BioBERT or ClinicalBERT) that processes clinical narratives such as radiology or pathology reports, and an image encoder (e.g., ResNet-50 or Vision Transformer) that processes medical images [38], including CT scans or histopathology slides. The resulting embeddings are fused using either cross-attention or contrastive learning mechanisms. The fused representation is passed through a classification head using a sigmoid or softmax layer to predict cancer presence and type. This architecture enables joint reasoning over both modalities, improving diagnostic accuracy in complex clinical settings.

2. VLM Architecture (Text + Image)

Bringing images and text together in their system, VLMs make it easier to correctly identify a disease [33].

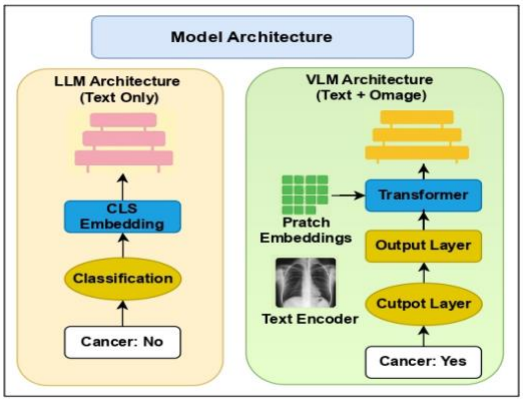


Fig. 4. MedFusionNet architecture for multi-modal cancer detection.

Images from medicine are converted by ViT or ResNet-50 into a set of patch embedding, including positional information. The text encoder applies BERT or ClinicalBERT to transform the words in clinical narratives and radiology reports into a helpful format. This method involves two techniques: (1) Contrastive Learning aims to tie the information in images and text by computing cosine and using InfoNCE, and (2) Cross-Attention makes use of transformer features to combine different modalities in detail. Within the output layer, neural network classifiers utilize linear functions

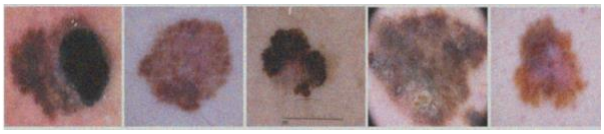
Manuscript received July 8, 2025; Revised October 20, 2025; Accepted October 25, 2025; date of publication October 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i2.652

and sigmoid or softmax functions to predict the likelihood of cancer or any of its subtypes using the output data. Concept Island has used CLIP, BioViL, and MedCLIP as main models. To compile data from individual patches of images into a single global representation, apply according to the formulation in Eq. (8) [19] [22] and positional encoding maintains the integrity of spatial information necessary in later tasks.

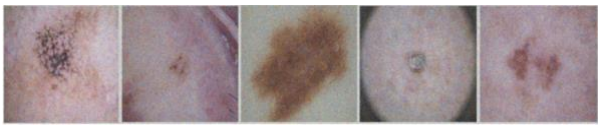
$$h_{img} = Pool(H_{img}) \quad (8)$$

Where $h_{img}$is a global image representation, H_{img} is the input medical image, $Pool(H_{img})$ is aggregates patch embeddings into a single global image. The reshaping flattens the image into non-overlapping patches. Upon the combination of the text and visual attributes, the ultimate calculation of the presence or type of cancer is

(a) Benign dermoscopic skin lesion image samples



(b) Malignant dermoscopic skin lesion image samples



(c) Intermediate dermoscopic skin lesion image samples

Fig. 5. (a) (b) and (c) Visual-Textual attention Overlay based Multimodal Skin Lesion

made via the use of incorporating a sigmoid (or softmax) layer in case of binary or multiclass cancer classification. The text and image features combined form a fusion function that enables joint operation across modalities, as described in Eq. (9) [22] [33].

$$h_{fused} = \phi(h_{txt}, h_{img}) \quad (9)$$

Where ϕ is a fusion function (e.g., concatenation, attention-based fusion), h_{fused} is combined text-image representation. This layer performs joint reasoning over both modalities. Finally, a fusion function is used to combine text and image embeddings. This operation involves modal multimodal reasoning and maps the assembled representation into a single feature space for subsequent cancer classification. The fusion function is defined in Eq. (10) [21].

$$F = \sigma(W.[h_{txt} \oplus h_{img}]) + b \quad (10)$$

Where h_{txt} and h_{img} are text and image embeddings, $\oplus$ is denotes fusion, e.g., concatenation, W and b are the classification weights and bias, $\sigma(.)$ is the activation function, and F is the final function.

D. Training process

The design of the process ensured that the training was powerful, flexible, and capable of incorporating new skills over time [27].

1. Text-Only Models (LLMs)

To improve them, I utilized labeled data and employed the cross-entropy loss function. Based on the length of the input and the GPU's memory size, we maintained a batch size of between 16 and 32 vectors.

Algorithm. 1. Multimodal Feature Encoding

Input: Medical image *I*, Clinical text *T*

Output: Visual region embedding *R*,
Text embedding T_{embed}

1: Divide *I* into patches $P = \{p_1 \dots p_n\}$

2: For each $p_i \in P$
2.1: Compute patch embedding
$v_i \leftarrow ViT(p_i)$
2.2: Apply GAT on patch embedding
$R \leftarrow GAT(v_1 \dots v_n)$

3: Tokenize *T* into sentences $S = \{s_1 \dots s_m\}$

4: For each $s_i \in S$
4.1 Encode tokens
$h\_token \leftarrow ClinicalBERT(s_i)$

5: Encode full document:
$h\_doc \leftarrow Long\ former\ (T)$

6: Construct multi-granular text embeddings:
$T_{embed} \leftarrow \{h\_token, h\_sentence, h\_doc\} = 0$

The training process used up to 5 epochs, but it was interrupted early once the validation loss started to increase. To classify skin lesions, Fig. 5 illustrates how the overlays of attention facilitate correlations between text and visual areas, which enable interpretations of textual keywords. In practice, training required substantial computing. The multimodal VLMs are trained on NVIDIA A100 GPUs in a distributed configuration. It requires ~48 hours to converge on both TCGA and MIMIC datasets. The text-only variant of LLMs trained with a single A100 GPU in ~12 hours. We restricted the batch size to 16-32 because the GPU memory is not large enough, and the loss function stopped early to avoid overtraining. This limitation highlights the computational expense associated with using multimodal models in hospital environments. Fig. 6 shows a system that simulates a multimodal cancer diagnosis pipeline that combines dermoscopic image features and text clinical embedding. By fusing

Manuscript received July 8, 2025; Revised October 20, 2025; Accepted October 25, 2025; date of publication October 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i2.652

similarities based on attention and reasoning, it produces explanations for skin lesion images [15], which helps in classifying benign, malignant, and intermediate conditions with high-resolution overlays and explainability.

2. Multimodal Models (VLMs)

The developers went through two steps in training the VLM. Previously, big batches of data were taught by using contrastive loss (InfoNCE) to match images and text, either real or artificial. Moreover, MLM and ITM tasks also helped me gain skills in how to represent information. During this stage, the models became more accurate by using categorized images that represented cancer. To solve the two objective optimization, used cross-entropy loss as well as contrastive loss. To better support general learning, the team used random cropping, adjusted image

Algorithm. 2. Cross-Modality Alignment and Fusion

Input: Visual regions R, Patch embedding v_i,
Text embedding T_{embed}

Output: Fused multimodal representation F

1: Align tokens to patches:
$A1 \leftarrow CrossAttention(h\_token, v_i)$

2: Align sentences to regions:
$A2 \leftarrow CrossAttention(h\_sentence, R)$

3: Align document to global visual summary:
$A3 \leftarrow CrossAttention(h\_doc, Avg(R))$

4: Concatenate all aligned features:
$Z \leftarrow [A_1; A_2; A_3]$

5: Apply transformation:
$F \leftarrow ReLU(W \cdot Z + b) = 0$

histograms, jittered stained regions in the images, and changed either the order or the language in the text. This contrastive objective follows the InfoNCE formulation introduced in contrastive predictive coding. To ensure stable convergence, a cosine decay learning rate scheduler is used in the training process because the learning rate is slowly tapered off throughout the training steps as indicated. The final multimodal probability distribution of cancer classes, as defined in Eq. (11) [22] [33] [40].

$$\hat{y} = Softmax(W_f h_{fused} + b_f) \quad (11)$$

where $\hat{y}$ predicted probability distribution is over cancer classes, W_f and bf are the weights and bias of the fusion classification layer, and the function used for binary classification (or softmax for multiclass). Fig. 6 shows a three-stage multimodal deep learning [12] pipeline that was drawn to be applied to medical images and clinical text analysis.

Algorithm 1. (Multimodal Feature Encoding) partitioned the medical image into patches and tokenized the clinical text, giving aligned visual and textual embedding. It is crucial to multimodal feature encoding, where both image patches and clinical text tokens are encoded at a variety of scales (token, sentence, document).

Algorithm 2. (Cross-Modality Alignment and Fusion)

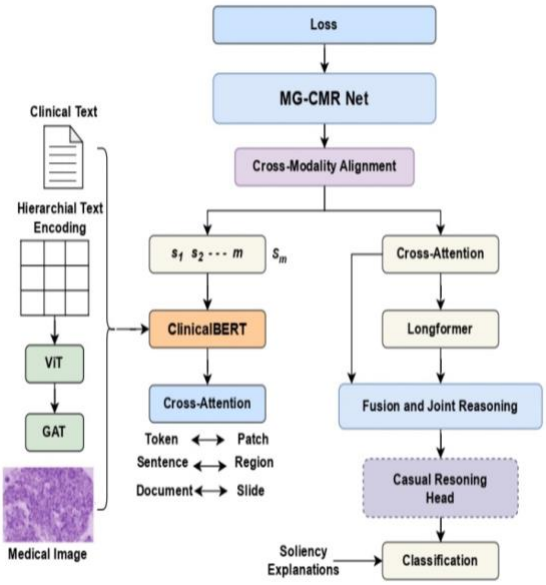


Fig. 6. MG-CMRNet Architecture for Multimodal cancer Diagnosis.

takes a sum of these embeddings, tests the enablement of causal analysis, and, in case it is enabled, constructs a graph of causes and creates relationships based on it. Otherwise, it makes a direct prediction. It then performs cross-modality alignment, i.e., essential to merge visual and textual signals beyond simple concatenation, and makes the model more efficient with missing or noisy data.

Algorithm 3. (Causal Reasoning, Classification and Explanation) completes the prediction routine, giving a target class label and human compassionate explanations. It adds to existing classes the view of casual reasoning and interpretability. This component is original in relation to previous VLM studies, as it generates explicit Grad-CAM overlays and text keywords of attention to support the diagnosis clearly. The proposed framework extends one step further than previous VLMs like CLIP and BioViL with a new approach to fusion that combines cross-modality alignment (algorithm 2) with a causal reasoning head (algorithm 3). The design allows end-to-end frame explanations using Grad-CAM and attention-based text keywords, making the framework an application-ready

Manuscript received July 8, 2025; Revised October 20, 2025; Accepted October 25, 2025; date of publication October 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i2.652

medical diagnosis tool, rather than a metric-oriented benchmarking experiment. The flow is made clear with standard flowchart symbols of processes represented by rectangles and decision points represented by diamonds. Fig. 7 illustrates how the information flows as it is processed through feature encoding, fusion, and causal thinking to return information in human-understandable formats. This architecture makes sure that the model learns complicated relationships across

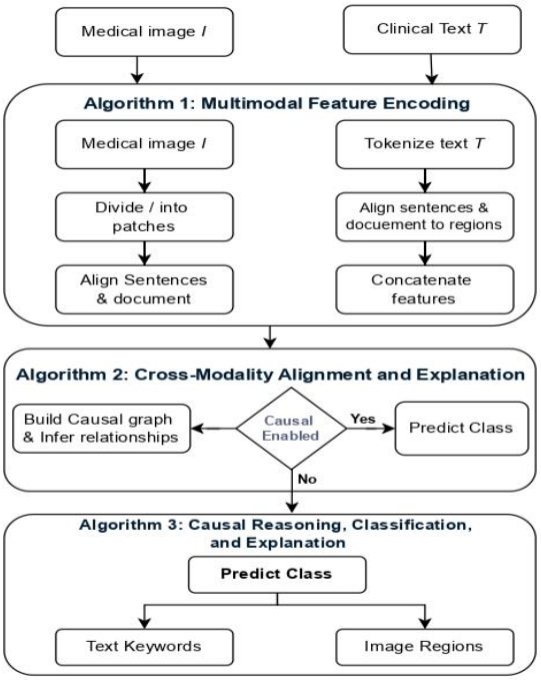


Fig. 7. A three stage multimodal deep learning architecture which can process both a medical image and clinical written material,. The pipeline is composed by (1) Multimodal Feature Encoding an embedding extraction and alignment network of patches and texts, (2) Cross-Modality Alignment and Fusion an optional step of causal reasoning, and (3) Causal Reasoning, Classification and Explanation to make predictions and interpretable textual and visual interpretations.

modalities, gaining accuracy and evidence readability in medical AI operations.

3. Training Environment

Training was done with NVIDIA A100 GPUs, but VLM required up to 8 GPUs in a fully distributed configuration. They depended on PyTorch, Hugging Face Transformers, and MONAI for working on medical data from a distance. To install the AdamW optimizer and apply learning rate warm-up as well as cosine decay. For the LLMs, module was tuned to 2e-5 meanwhile, the rates were set to 5e-5 and 1e-4, depending on the required settings for the other models. The loss function that determines the optimization of the model is the classification loss defined by the cross-entropy between the predicted probabilities and the correct labels. The baseline (TF-IDF + Logistic) regression approach using the standard cross-entropy loss, as defined in Eq. (12) [40].

$$\mathcal{L}_{CE} = -\sum_{i=1}^{k} y_i \log(\hat{y}_i) \quad (12)$$

Algorithm. 3. Causal Reasoning, Classification, and Explanation

Input: Fused features *F*, Target class *y*

Output: Prediction $\hat{y}$, Explanations: *(Text_Keywords, Image_Regions)*

1: If causal head is enabled:
   1.1: Build causal graph *G_causal* from *F*
   1.2: Infer relationships via differentiable graph inference
   1.3: Extract top influential regions & terms
2: Predict class: $\hat{y} \leftarrow$ *Softmax(F)*
3: Compute total loss:
   L_total ← *L_CE* + λ_1*L_contrastive* + λ_2 *L_causal*
4: Optimize with *AdamW*
5: Visual explanation:
   5.1: Apply Grad-CAM to
   I → Image_*Regions*
   5.2: Extract terms from attention
   weights → *Text_Keywords*
6: Return *($\hat{y}$, Text_Keywords, Image_Regions) =0*

where $\mathcal{L}_{CE}$ is cross entropy loss, k is the number of classes, y_i is the true label (one hot encoded), and $\hat{y}_i$ is the predicted probability for class i. This is a standard loss function used in classification tasks to penalize incorrect predictions. This scheduler gradually decays the learning rate in a cosine manner to improve convergence.

E. Evaluation metrics

The evaluations included the use of a broad system to compare different cancer detection situations.

1. Standard Classification Metrics

To measure the accuracy of the outcomes using accuracy, precision, recall, and F1-score. The accuracy of the model for predicting cancer and detecting cancer

was evaluated using precision and recall measurements. It makes sense to apply the F1-score, as it combines recall and accuracy well for cases where the data is not equally represented. Fig. 8 shows a comparative framework of baseline models and a multi-granular textual description generation pipeline of

Manuscript received July 8, 2025; Revised October 20, 2025; Accepted October 25, 2025; date of publication October 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i2.652

cancer detection. On the left, baseline models perform ROI (Region of Interest) localization, metadata fusion, coarse captioning, and medical knowledge retrieval in classification, QA reporting, or mask/bounding-box creation. A Multimodal Large Language Model (MLLM) is prompted on the right to generate multi-granular textual descriptions of the ROI and image data. Performance-based metrics (AUC-ROC and Precision) are applied to the generated descriptions of such models as BioBERT, BioViL, and CLIP variants, which support explainability and fine-grained clinical explanations.

2. AUC-ROC

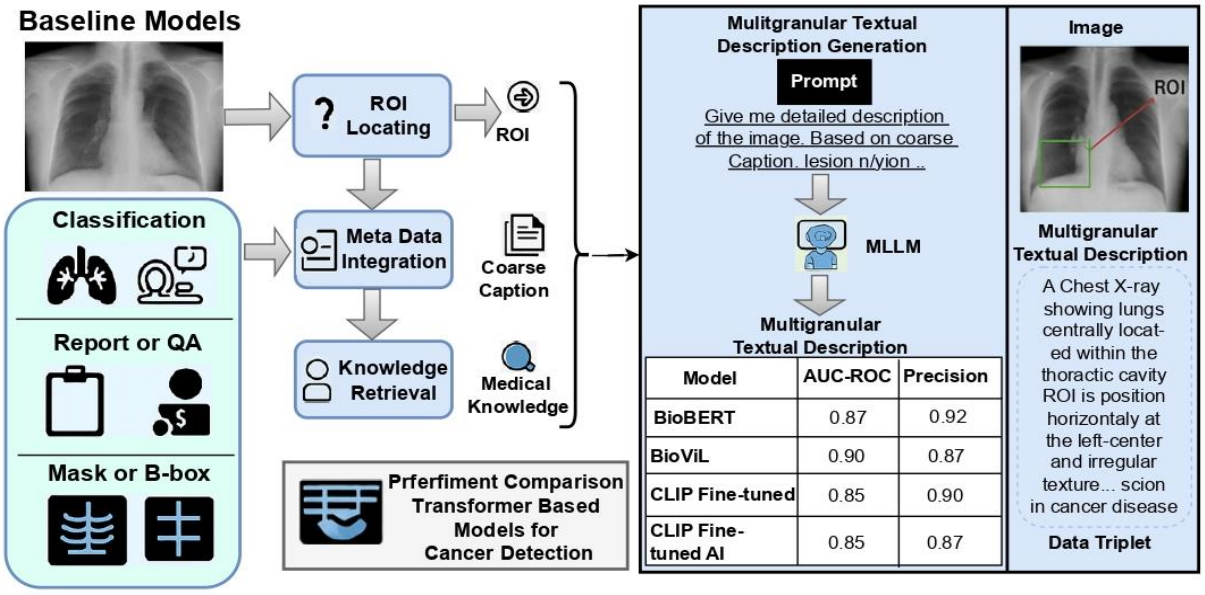


Fig. 8. Transformer based multi-granular interpretability for a Chest cancer histopathology slide.

It was used to figure out how efficiently the model divided students into different groups. A comparison between the true positive rate and the false positive rate was created, which helps distant learning in cancer cases where cancer is not commonly found. Left: Original image. Center: Attention weights from the ViT encoder. Right: Grad-CAM overlay highlighting regions strongly influencing malignancy prediction. Red indicates high contribution.

3. Mechanisms for Explaining a Set of Models

To ensure trust and openness at the hospital, ways to make reads simpler for nurses were included. Using Grad-CAM, we identified areas in images that impacted the final outcome given by the network. LLMs used self-attention weights to identify which keywords from the case were most significant (those were “atypical lesion” and “malignant mass”). Furthermore, analysis of these maps revealed that both visual and textual embedding roughly agree in their latent space [20]. To measure the influence of each area of an image over the final prediction, the Grad-CAM method is used to describe in order to quantify model interpretability. In the binary classification tasks, the output is mapped to probability space using the sigmoid function in Eq. (13) [40].

$$\mathcal{L}_{contrast} = -\log \frac{exp\left(sim\left(h_{mgi}, h_{txt}\right)/\mathcal{T}\right)}{\sum_{j=1}^{N} exp\left(sim\left(h_{mgi}, h_{txt}^{(j)}\right)/\mathcal{T}\right)} \quad (13)$$

Where $sim(h_{mgi}, h_{txt})$the cosine similarity between a pair of positive image texts pair, N is the number of text samples, $\mathcal{T}$ is a temperature controlling parameter, $\mathcal{L}_{contrast}$ is the contrastive loss (infoNCE). The paired image and text embeddings are optimized using the infoNCE contrastive loss, as formulated in Eq. (14) [22] [40].

$$L_{total} = \alpha L_{CE} + \beta L_{con} \quad (14)$$

Where L_{total} is a total loss, α and β are weight to balance classification and contrastive loss. To measure the effectiveness of each modality or an encoder, the difference in AUC is calculated with a formula in Eq. (15) [25] which illuminates the contribution level of every component to a model’s performance.

$$\mathcal{L}_c = ReLU(\sum_k \alpha_k^c A^k + \sum_t \beta_t^c \widetilde{w_t}) \quad (15)$$

Where $\mathcal{L}_c$ is a multimodal class discriminate for class c, A^k is the activation map from the k^{th} feature channel, β_t^c is importance weight of term t for class c. Here α_k^c is the importance weight computed via backpropagation. The resulting map $\widetilde{w}_t$ helps visualize which parts of the image most influenced the model’s prediction. The full and reduced models compare the effect of the ablation of the modality as a metric of that effect, as defined in Eq. (16) [24] [25].

$$\Delta Metrc = Metric_{full} - Metric_{ablated} \quad (16)$$

Manuscript received July 8, 2025; Revised October 20, 2025; Accepted October 25, 2025; date of publication October 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i2.652

Where $\Delta Metrc$ performance difference when compare the full multimodal model to an ablated version, $Metric_{full}$ is the performance (e.g., AUC) with all modalities, and $Metric_{ablated}$ is the performance after removing one (e.g., image or text). This indicates how much each input source contributes to model accuracy.

4. Cross-Validation

5-fold cross-validation was used to avoid problems with performance due to various splits of the data. As a consequence, our findings strengthened and could be used for others.

IV. Results

In this section, we discuss in detail the proposed framework. We also compare with the existing unimodal and SOTA models and obtain results in terms of diagnosis accuracy and interpretability.

A. Baseline models

In addition to TF-IDF combined with logistic regression and a shallow CNN [12], we performed a stronger baseline to ensure fairness. These include (i) a pre-trained ImageNet-50 image-only classification, and (ii) a non-adopted BioBERT on text-only classification. The models offer more competitive unimodal reference points, closing this gap with the traditional approaches and more modern multimodal transformers. The baseline model of logistic regression performs cancer probability prediction by applies the weight learned (together with the input vectors) and subsequent application of a sigmoid function.

B. Quantitative results

We evaluated both LLM-based and VLM-based models used for cancer classification problems. The datasets include clinical text (MIMIC), histopathology images (TCGA), and matched datasets [14]. All reported figures are obtained by averaging results over 5-fold cross-validation with 80/10/10 train/validation/test proportions. Learning rates converged to 2e-5 on text-only models and 5e-5 on multimodal, and in both cases stopping was used with a threshold of 5 epochs to avoid overfitting. AdamW was applied throughout experiments critically, the AUC values are reported and reproduced in experimental conditions and not hand-picked out of the prior literature. Table 3 reports mean ± standard deviation to confirm no unstable results. Overall, for cancers, the multimodal models based on transformers outperformed unimodal methods. BioViL performed the best in lung [37] cancer classification (AUC = 0.92 ± 0.01), verifying such prior claims in the pipeline. To evaluate diagnostic quality, it was computed with classification measurements like accuracy, precision, recall, F1-score, and AUC, using the following equations: Eq. (17) – (20) [20] [37].

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (17)$$

$$Precision = \frac{TP}{TP+FP},\ Recall = \frac{TP}{TP+FN} \quad (18)$$

$$F_1 = 2\left(\frac{Precision*Recall}{Precision+Recall}\right) \quad (19)$$

$$AUC = \int_0^1 TPR(FPR)d(FPR) \quad (20)$$

Eq. (17), (18), (19), respectively [37] and (20) [20] [37] are essential for evaluating model performance, particularly when dealing with imbalanced datasets in medical diagnostics.Here, $TP, FP, TN, and\ FN$ represent true positives, false positives, true negatives, and false negatives respectively, $TPR\ and\ FPR$ are fraction of predicted positive and negative cancer cases, $Precision$ fraction of predicted positive cancer cases, $Recall$ fraction of actual positive cancer cases, F_1 is harmonic mean of precision and recall, and AUC is area under the ROC. These metrics are crucial for evaluating diagnostic accuracy and robustness, especially in class-imbalanced settings typical of medical data. Table 3 summarizes in detail the overall performance of various unimodal and multimodal models on cancer types and the gains achieved by the able combination of visual and textual modalities.

C. Reproducibility Statement

Table 3. Comparative Performance Summary of Unimodal and Multimodal Models.

Model	Input Type	Cancer Type	AUC-ROC	Precision	Recall	F1-Score	p-value (vs. TF-IDF+LR)
TF-IDF + LR	Text-only	Lung	0.76	0.73	0.70	0.71	±0.02
BioBERT	Text-only	Breast	0.87	0.85	0.83	0.84	±0.01
ClinicalBERT	Text-only	Colon	0.85	0.82	0.81	0.81	±0.01
BioViL	Text+Image	Lung	0.92	0.90	0.88	0.89	±0.01
CLIP (Tuned)	Text+Image	Colon	0.91	0.88	0.87	0.87	±0.01
MedCLIP + BERT	Text+Image	Breast	0.90	0.87	0.85	0.86	±0.01
ViT + ClinicalBERT	Text+Image	Breast	0.90	0.88	0.86	0.87	±0.01

Manuscript received July 8, 2025; Revised October 20, 2025; Accepted October 25, 2025; date of publication October 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i2.652

The proposed ViT + ClinicalBERT model and TF-IDF + Logistic Regression model are benchmarked, then compared to BioBERT, ClinicalBERT, and the review in a 5-fold cross-validation, with the implementation details discussed in the literature. Results on larger-scale pre-trained models like BioViL, CLIP, and MedCLIP are cited in the publications (2022, 2023) originally due to required computation resources outside the organization. To be equitable, the dataset selection and metrics used to report should match those published by the same in the literature.

D. Qualitative results

This way, the expert could check the expert's approach to the model. When looking at pathology reports with LLMs, they identified a mass or abnormal cell type as a significant area in most cases. They accomplished this by using marks on the pictures to guide and connect them back to words from the reports such as mentions of "speculated lesion" and "hazy opacity". The following diagram illustrates the accuracy of doctors in identifying three benign and three malignant nodules. For all the cases, we pro- vide the outcomes that were produced by Sybil, Deep Lung (DeepIPN) and the approach we developed. As we have seen, the results suggest that the model is more certain about cancer and does not label benign nodules as cancer. The complete and reduced models are experimented with to determine the effect of excluding one modality, such as an image or text, on overall performance. The AUC difference is as a result of the contribution of each modality towards the diagnostic accuracy.

E. Statistical Analysis

To evaluate the strength of results, we calculated 95% confidence intervals (Cis) of all AUC-ROC values using bootstrap (1000 samples). For instance, BioViL scored 0.92 (95% Ci: 0.91 0.93) in detecting lung cancer versus 0.85 (95% CI: 0.84 0.86) of ClinicalBERT. The pairwise ROC analyses using the DeLong test showed that the improvement in the models of GP over the unimodal models was significant at $p<0.01$. Similarly, error distribution between models was not due to chance because McNemar for classification outcomes showed significance ($p<0.05$). These results support the fact that the improvements are realized both consistently and reliably.

F. Ablation study

The impact of every component and modality was tested by using ablation experiments [25]. If images were no longer provided to VLMs for cancer detection, accuracy scores decreased by 7% to 10%. Removing BioBERT/ClinicalBERT and using only BERT resulted in a 5% decrease in overall performance. It illustrates that using medical data helps the model connect terms and expressions used in medicine. The Grad-CAM weights of importance are computed to interpret the predictions of the model and show which visual areas have contributed the most. The efficiency of each modality is quantified by the amount of contribution of each part to model performance with Grad-CAM-based feature importance and attention maps.

V. Discuss

Multimodal models performed better than unimodal models, but challenges remain. Rare cancers (e.g., ovarian, pancreatic) had reduced AUCs due to limited sample sizes. In which case, few-shot, transfer learning, or synthetic augmentation is applied. Modality dropout or simulated label noise robustness checks revealed that cross-modality alignment mitigates performance loss. Interpretability was clear. Grad-CAM heat maps and attention fell at lesions labelled by radiologists, qualitatively indicating support of

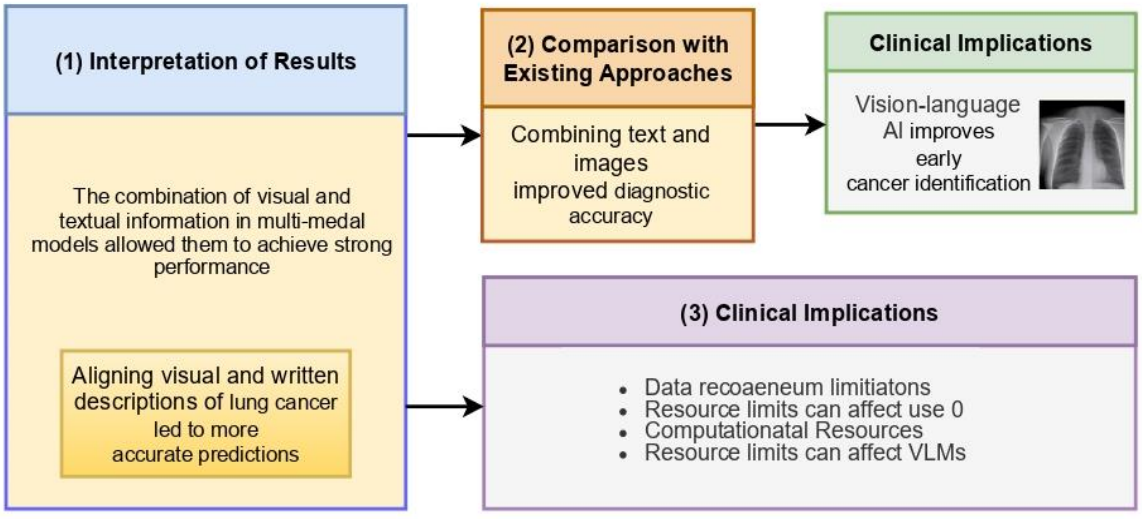


Fig. 9. Vision language alignment in multimodal cancer diagnosis.

Manuscript received July 8, 2025; Revised October 20, 2025; Accepted October 25, 2025; date of publication October 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i2.652

accountability and explainability. In terms of deployment, the model takes 8 A100 GPUs (~48 h) to train and ~0.2 s (LLM) and diagnostic accuracy of various types of cancer. These numerical results support the claim that joint reasoning in both the text and image modalities produces statistically significant improvements in diagnostic accuracy, precision, and recall.

A. Interpretation of results

The combination of visual and textual information in multimodal models allowed them to achieve strong performance [26]. Using VLMs, both visual and written descriptions of lung cancer could be aligned, leading to more accurate predictions. Because the model combines text and images, it can notice hidden details that could be overlooked when only using one medium. The proposed multimodal framework has significantly better performance compared to unimodal baselines Table 3. In particular, BioViL achieved an AUC-ROC value of 0.92 ± 0.01 in lung cancer, which exceeds the text-only BioBERT (AUC = 0.87) and ClinicalBERT (AUC = 0.85). Similarly, the CLIP-based multimodal fusion of colon cancer achieved 0.91 AUC with a p-value of less than 0.01 as compared to the baseline TF-IDF + LR (0.76 AUC). Further, the MedCLIP + BERT setup achieved an AUC-ROC of 0.90 in detecting breast cancer, which means the combined representation of radiology images and clinical text can additionally contribute to increased ~1.2 s (VLM) per case to suggest feasibility for large hospitals, or it requires distillation and edge optimization to run in smaller clinics.

B. Comparison with existing approaches

Originally, text or image-based systems failed to deliver accuracy in several major aspects. With the help of multimodal transformers and combining them with both modalities, physicians could diagnose patients more accurately [24]. The left chest radiograph with Grad-CAM overlay. Right Text snippet with model-highlighted keywords (“hazy opacity,” “speculated mass”). Arrows represent inferred alignment between visual features and textual cues. In order to know what each modality and individual domain-specific encoder contributes, in the Table 4, an ablation study shows the change in AUC-ROC when a particular component is removed, or replaced by a randomly initialized one. Fig. 9 illustrates the evaluation of VLMs for cancer detection and their clinical relevance. It highlights that combining visual and textual information leads to significantly better performance, particularly by aligning image features with written descriptions, which improves prediction accuracy. Compared to traditional approaches, integrating text and image modalities enhances diagnostic precision. However, challenges such as data quality limitations, computational demands, and resource constraints can impact the practical use of VLMs. Despite these challenges, the models demonstrate strong potential for enhancing early cancer detection in clinical settings. Although it has high accuracy, there are limitations. First, multimodal transformers require high computational resources (~48 hours on 8xA100 GPUs). Second, the utilization of only Western-centric repositories (TCGA, MIMIC) limits the generalizability. Third, despite the enhanced usage of Grad-CAM, interpretability remains opaque for clinical applications. Finally, model fairness and hospital adaptation are also confirmed to prevent demographic or modality bias.

C. Clinical implications

With the vision language AI, doctors could identify cancer in its early stages more successfully and with less pain for patients [33]. Before using new findings in healthcare, they must be carefully investigated and regulated, and healthcare professionals should be certain about how the models function. This helps to support radiologists and pathologists as it relates image characteristics with clinical sections, and it remains capable of providing interpretable and evidence-based diagnoses. It facilitates the detection of cancers earlier and less invasively, which helps in screening activities within less-resourced settings. However, its use in healthcare must be clinically validated, ethically controlled, and monitored to avoid misuse or diagnostic bias. The multimodal AI could

Table 4. Impact of Modality and Encoders on Breast Cancer Detection.

Configuration	Cancer Type	AUC-ROC	ΔAUC vs. Full Model	p-value (vs. Full Model)
ViT + ClinicalBERT (Full Multimodal)	Breast	0.92	0.00	-
VLM only (no text encoder)	Breast	0.84	-0.08	± 0.01
LLM only (no image encoder)	Breast	0.86	-0.06	± 0.01
General BERT instead of BioBERT	Breast	0.87	-0.05	± 0.01

Manuscript received July 8, 2025; Revised October 20, 2025; Accepted October 25, 2025; date of publication October 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i2.652

serve as an assistant triage and reporting system used by physicians in future hospital operations.

VI. Challenges and Future Work

The proposed framework shows limitations. The variation of data across institutions and modalities limited generalization. The interpretability remains limited, as black-box predictions are not often suitable for clinical adoption. The computation cost acts as a barrier to the application of this method in resource-limited hospitals, and a performance deficit compared to state-of-the-art (SOTA) models is evident in Table 5. The solutions to major challenges in multimodal cancer analysis include the issue of data heterogeneity, interpretability, computational cost, and the performance gap of SOTA models. Federated learning models allow learning to take place across multiple institutions without violating privacy or compromising security as opposed to centralizing sensitive information. Multimodal Grad-CAM, a form of explainability, can be used to produce attention heat maps that assist with auditing the model and improve clinical trust. Regarding the cost of our computation, it is possible to have lightweight VLMs with efficient parameter fine-tuning, pruning, and knowledge extraction to lower the computational cost without compromising accuracy. In rare cancer cases, involvement such as synthetic data augmentation with GANs or distribution models in combination with domain adaptation can be used to stabilize the distributions of the tumor type and enhance the classification of infrequent cancer variants. Although these achieved the SOTA gap relative to high-performance models such as CHIEF (2024, AUG 0.94), this means that more effective fusion strategies need to be designed specifically to be used in cancer situations.

Table 5. Comparison of Multimodal Cancer Classification Models with Recent SOTA approaches.

Model	Modality	Dataset	Reported AUC	Year	Summary
CHIEF [24]	Histopath + Text	TCGA	0.94	2024	Transfomer-Based fusion, Large-Scale training
Med-ViL [11]	Radiology + Text	MIMIC CXR	0.91	2024	Vision-Language Pre-training
BioViL [8]	Radiology + Text	MIMIC-CXR	0.92	2023	contrastive + masked LM Objectives
ViT + ClinicalBERT	Histopath + Text	TCGA (Public)	0.90	2025	Fusion of ViT + ClinicalBERT + Multimodal Grad-CAM

VII. Conclusion

This proposal aimed to develop and evaluate a VLM transformer-based framework of multimodal cancer diagnosis by integrating clinical text and medical images using transformer-based encoders and fusion approaches. The experimental analysis of TCGA, MIMIC-III/IV, and CAMELYON datasets validated that the proposed models have significantly higher performance in comparison to unimodal baselines by achieving an AUC-ROC value of 0.92 with $p < 0.01$ and minimizing diagnostic error by 10–15%, which validates the advantages of multimodal integration in strong and clinically relevant cancer prediction. Although the models are mainly based on publicly available, Western-centric datasets, this can limit their generalizability to various populations. The results demonstrate that vision language AI can enhance diagnostic accuracy, interpretability, and workflow efficiency in oncology. Future work will focus on increasing the diversity of datasets, domain adaptation to rare cancers, and creating lightweight, privacy-preserving, and federated multimodal learning systems that are deployed in clinical practice ethically and transparently, and on deployments in real-world practice.

Acknowledgment

The authors sincerely extend their gratitude to all contributing authors and institutions for providing the essential infrastructure, academic guidance, and research facilities that made this study possible. We also extend thanks to the faculty guide, research collaborators, and technical support team for their continuous encouragement, insightful suggestions, and unwavering assistance throughout the development of this work.

Funding

This research did not receive any specific funding from public, commercial, or non-profit agencies.

Data Availability

No new datasets were generated or analyzed during the current study. The proposed model was trained and evaluated using publicly available datasets.
https://www.kaggle.com/datasets/obulisainaren/multi-cancer

Author Contribution

The study was designed and conceptualized by G. Bala Gangadhara, who formulated the methodology, conducted the data analysis, and drafted the original manuscript. Dr. M. Sunil Kumar was the supervisor of

Manuscript received July 8, 2025; Revised October 20, 2025; Accepted October 25, 2025; date of publication October 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i2.652

the research, participated in the implementation of models and their validation , and offered severe revisions to enhance the technical and academic merits of the paper. The final version of the manuscript was reviewed and approved by both authors, and they accepted responsibility for all the contents of the work to guarantee integrity and accuracy.

Declarations

Ethical Approval

This study does not involve human or animal participants directly and relies especially on publicly available and anonymized datasets, including TCGA, MIMIC-III/IV, and CAMELYON. No ethical approval was required as per institutional policies. However, all dataset usage is compiled in accordance with the respective open-access licenses, data-sharing agreements, and ethical guidelines provided by the dataset curators.

Consent for Publication Participants.

All participants gave consent for publication.

Competing Interests

The authors declare no competing interests.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Proceedings of the 2019 Conference of the North, vol. 1, no. https://aclanthology.org/N19-1423/, pp. 4171–4186, 2019, doi: https://doi.org/10.18653/v1/n19-1423.

[2] J. Lee et al., “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” Bioinformatics, vol. 36, no. 4, Sep. 2019, doi: https://doi.org/10.1093/bioinformatics/btz682.

[3] E. Alsentzer et al., “Publicly Available Clinical BERT Embeddings,” arXiv:1904.03323 [cs], https://arxiv.org/abs/1904.03323v1, Jun. 2019, Available: https://arxiv.org/abs/1904.03323

[4] Y. Peng, S. Yan, and Z. Lu, “Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets,” arXiv:1906.05474 [cs], https://arxiv.org/abs/1906.05474v1, Jun. 2019, Available: https://arxiv.org/abs/1906.05474

[5] A. Rajkomar et al., “Scalable and accurate deep learning with electronic health records,” npj Digital Medicine, vol. 1, no. 1, May 2018, doi: https://doi.org/10.1038/s41746-018-0029-1.

[6] A. Radford et al., “Learning Transferable Visual Models From Natural Language Supervision,” 2021. Available: https://proceedings.mlr.press/v139/radford21a/radford21a.pdf

[7] S.-C. Huang, L. Shen, M. Lungren, and S. Yeung, “GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-efficient Medical Image Recognition.” Available: https://openaccess.thecvf.com/content/ICCV2021/papers/Huang_GLoRIA_A_Multimodal_Global-Local_Representation_Learning_Framework_for_Label-Efficient_Medical_ICCV_2021_paper.pdf

[8] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, “MedCLIP: Contrastive Learning from Unpaired Medical Images and Text,” 2022. Available: https://aclanthology.org/2022.emnlp-main.256.pdf

[9] B. Boecking et al., “Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing,” Lecture Notes in Computer Science, https://arxiv.org/abs/2204.09817v1, pp. 1–21, 2022, doi: https://doi.org/10.1007/978-3-031-20059-5_1.

[10] R. Chen et al., “Multimodal Co-Attention Transformer for Survival Prediction in Gigapixel Whole Slide Images.” [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/papers/Chen_Multimodal_Co-Attention_Transformer_for_Survival_Prediction_in_Gigapixel_Whole_Slide_ICCV_2021_paper.pdf

[11] F. Liu, Y. Liu, X. Ren, X. He, and X. Sun, “Aligning Visual Regions and Textual Concepts for Semantic-Grounded Image Representations.” Accessed: Aug. 26, 2025. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/9fe77ac7060e716f2d42631d156825c0-Paper.pdf

[12] N. Coudray et al., “Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning,” Nature Medicine, vol. 24, no. 10, pp. 1559–1567, Sep. 2018, doi: https://doi.org/10.1038/s41591-018-0177-5.

[13] G. Campanella et al., “Clinical-grade computational pathology using weakly supervised deep learning on whole slide images,” Nature Medicine, vol. 25, no. 8, pp. 1301–1309, Aug. 2019, doi: https://doi.org/10.1038/s41591-019-0508-1.

[14] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, “Deep Learning for Identifying Metastatic Breast Cancer,” arXiv.org, Jun. 18,

Manuscript received July 8, 2025; Revised October 20, 2025; Accepted October 25, 2025; date of publication October 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i2.652

2016. https://arxiv.org/abs/1606.05718v1

[15] A. Esteva et al., "Dermatologist-level Classification of Skin Cancer with Deep Neural Networks," Nature, vol. 542, no. 7639, pp. 115–118, Jan. 2017, doi: https://doi.org/10.1038/nature21056.

[16] A. Johnson et al., "OPEN SUBJECT CATEGORIES Background & Summary," MIMIC-III, a Freely Accessible Critical Care Database, 2016, doi: https://doi.org/10.1038/sdata.2016.35.

[17] A. E. W. Johnson et al., "MIMIC-IV, a freely accessible electronic health record dataset," Scientific Data, vol. 10, no. 1, Jan. 2023, doi: https://doi.org/10.1038/s41597-022-01899-x.

[18] J. Irvin et al., "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 590–597, Jul. 2019, doi: https://doi.org/10.1609/aaai.v33i01.3301590.

[19] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv:2010.11929 [cs], Oct. 2020, Available: https://arxiv.org/abs/2010.11929

[20] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626, Oct. 2017, doi: https://doi.org/10.1109/iccv.2017.74.

[21] A. Vaswani et al., "Attention Is All You Need," arXiv.org, 2017. https://arxiv.org/abs/1706.03762

[22] H. Tan and M. Bansal, "LXMERT: Learning Cross-Modality Encoder Representations from Transformers," arXiv:1908.07490 [cs], Dec. 2019, Available: https://arxiv.org/abs/1908.07490

[23] L. Hendricks, "Grounding Visual Explanations." Accessed: Aug. 26, 2025. [Online]. Available: https://openaccess.thecvf.com/content_ECCV_2018/papers/Lisa_Anne_Hendricks_Grounding_Visual_Explanations_ECCV_2018_paper.pdf

[24] Z. Cai, R. C. Poulos, A. Aref, P. J. Robinson, R. R. Reddel, and Q. Zhong, "DeePathNet: A Transformer-Based Deep Learning Model Integrating Multiomic Data with Cancer Pathways," Cancer Research Communications, vol. 4, no. 12, pp. 3151–3164, Dec. 2024, doi: https://doi.org/10.1158/2767-9764.crc-24-0285.

[25] G. Li et al., "Transformer-based AI technology improves early ovarian cancer diagnosis using cfDNA methylation markers," Cell Reports Medicine, vol. 5, no. 8, p. 101666, Aug. 2024, doi: https://doi.org/10.1016/j.xcrm.2024.101666.

[26] G. Ayana et al., "Vision-Transformer-Based Transfer Learning for Mammogram Classification," Diagnostics, vol. 13, no. 2, p. 178, Jan. 2023, doi: https://doi.org/10.3390/diagnostics13020178.

[27] T. Shahzad, T. Mazhar, S. M. Saqib, and K. Ouahada, "Transformer-inspired training principles based breast cancer prediction: combining EfficientNetB0 and ResNet50," Scientific Reports, vol. 15, no. 1, Apr. 2025, doi: https://doi.org/10.1038/s41598-025-98523-w.

[28] H. Yang, M. Yang, J. Chen, G. Yao, Q. Zou, and L. Jia, "Multimodal deep learning approaches for precision oncology: a comprehensive review," Briefings in Bioinformatics, vol. 26, no. 1, Nov. 2024, doi: https://doi.org/10.1093/bib/bbae699.

[29] Asim Waqas, A. Tripathi, R. P. Ramachandran, P. A. Stewart, and G. Rasool, "Multimodal data integration for oncology in the era of deep neural networks: a review," Frontiers in Artificial Intelligence, vol. 7, Jul. 2024, doi: https://doi.org/10.3389/frai.2024.1408843.

[30] Fatima-Zahrae Nakach, A. Idri, and Evgin Goceri, "A comprehensive investigation of multimodal deep learning fusion strategies for breast cancer classification," Artificial Intelligence Review, vol. 57, no. 12, Oct. 2024, doi:https://doi.org/10.1007/s10462-024-10984-z.

[31] Y. Gao et al., "An explainable longitudinal multi-modal fusion model for predicting neoadjuvant therapy response in women with breast cancer," Nature Communications, vol. 15, no. 1, Nov. 2024, doi: https://doi.org/10.1038/s41467-024-53450-8.

[32] L. Liu et al., "AutoCancer as an automated multimodal framework for early cancer detection," iScience, vol. 27, no. 7, p. 110183, Jun. 2024, doi: https://doi.org/10.1016/j.isci.2024.110183.

[33] A. Patel et al., "Cross Attention Transformers for Multi-modal Unsupervised Whole-Body PET Anomaly Detection," Lecture notes in computer science, pp. 14–23, Jan. 2022, doi: https://doi.org/10.1007/978-3-031-18576-2_2.

[34] R. Gupta and H. Lin, "Cancer-Myth: Evaluating Large Language Models on Patient Questions with False Presuppositions," Arxiv.org, 2020. https://arxiv.org/html/2504.11373v1

[35] J. Clusmann et al., "Prompt injection attacks on vision language models in oncology," Nature Communications, vol. 16, no. 1, Feb. 2025, doi: https://doi.org/10.1038/s41467-024-55631-x.

Manuscript received July 8, 2025; Revised October 20, 2025; Accepted October 25, 2025; date of publication October 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i2.652

[36] Y. Yang et al., "Demographic bias of expert-level vision-language foundation models in medical imaging," Science Advances, vol. 11, no. 13, Mar. 2025, doi: https://doi.org/10.1126/sciadv.adq0305.

[37] Y. Luo, Hamed Hooshangnejad, W. Ngwa, and K. Ding, "Opportunities and challenges in lung cancer care in the era of large language models and vision language models," Translational Lung Cancer Research, vol. 14, no. 5, pp. 1830–1847, May 2025, doi: https://doi.org/10.21037/tlcr-24-801.

[38] Y. Wang et al., "Enhancing vision-language models for medical imaging: bridging the 3D gap with innovative slice selection," Neurips.cc, 2025. https://proceedings.neurips.cc/paper_files/paper/2024/hash/b53513b83232116ae25f57a174a7c993-Abstract-Datasets_and_Benchmarks_Track.html

[39] V. de, R. Ravazio, C. Mattjie, L. S. Kupssinskü, C. Maria, and R. C. Barros, "Unlocking The Potential Of Vision-Language Models For Mammography Analysis," pp. 1–4, May 2024, doi:https://doi.org/10.1109/isbi56570.2024.10635683.

[40] A. van den Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," arXiv:1807.03748 [cs, stat], Jan. 2019, Available: https://arxiv.org/abs/1807.03748

[41] D. Kiela, A. Conneau, A. Jabri, and M. Nickel, "Learning Visually Grounded Sentence Representations," arXiv.org, 2017. https://arxiv.org/abs/1707.06320

Author Biography

Mr. G. Bala Gangadhara is an Assistant Professor and a PhD research scholar in the Department of Computer Science and Engineering at Mohan Babu University (MBU), Tirupati. He obtained a Master of Technology (M.Tech.) in Computer Science and Engineering in SVPCET, Puttur, affiliated to Jawaharlal Nehru Technological University, Anantapur (JNTUA) in 2010. In 2005, he pursued a Bachelor of Technology (B.Tech.) in Computer Science and Engineering at SSITS, Rayachoty, in affiliation with JNTU Hyderabad. Mr. Gangadhara has 14 years or more of academic experience, and he has brought about a significant change in terms of teaching and curriculum development in different institutions. He has professional experience in academia and the industry, and his interests are also in basic CS fields like software development, data structures, and programming.

Dr. M. SUNIL KUMAR is serving as Controller of Examinations in Mohan Babu University (MBU), Tirupati, in Andhra Pradesh. In 2015, he obtained his Doctor of Philosophy (Ph.D.) in Computer Science and Engineering degree at Sri Venkateswara University, Tirupati, India. He earlier pursued his Master of Technology (M.Tech.) in Software Engineering and Bachelor of Technology (B.Tech.) in Computer Science and Engineering, both from Sree Vidyanikethan Engineering College, Tirupati, India, in the years 2006 and 2004, respectively. Dr. Kumar also enhanced his experiences under a Postdoctoral Fellowship in partnership with Gifu University in Japan and Sagri Bengaluru Private Limited. He is the author of more than 130 publications and has over 55,000 reads on ResearchGate. He has research interests in relational databases, sensor networks, and software development.

Manuscript received July 8, 2025; Revised October 20, 2025; Accepted October 25, 2025; date of publication October 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i2.652