Journal of Theoretical and Applied Information Technology
15st July 2026. Vol.104. No.13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

FUSIONSCANAI: AN ATTENTION-BASED MULTI-MODAL FRAMEWORK FOR CANCER DETECTION FROM MRI AND CT IMAGES

DR. S. SWAPNA RANI[1*], JOHN BABU GUTTIKONDA[2], KAPARTHI UDAY[3], MV NAGESH[4], DAMERA VENKATESH[5]

[1]Associate Professor, ECE Department, MVSR Engineering College, Hyderabad, India.
[2]Department of Computer Science & Engineering (AI & ML), Designation: Associate Professor, Affiliation: Anurag Engineering College, Kodad, Telangana, India.
[3]Associate Professor, Department of Electronics and Instrumentation Engineering, CVR College of Engineering, Ibrahimpatanam, Hyderabad, India.
[4]Dept of Computer Science and Engineering, Sreyas Institute of Engineering and Technology, Telangana, India.
[5]Assistant Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Bowrampet, Hyderabad-500043, Telangana, India.
E-mail: sswapnarani_ece@mvsrec.edu.in, johnbabug@gmail.com,k.uday@cvr.ac.in, nageshmathew@gmail.com, *dameravenkatesh7@gmail.com
*Corresponding Author

ABSTRACT

Diagnosis of metastatic cancer remains one of the challenging problems in clinical oncology, as disease spread is highly complex and heterogeneous by nature. Widely used deep learning models extract features from single-modality inputs, including imaging or clinical data, but are thus hindered in their diagnostic robustness and generalizability. Meanwhile, existing fusion strategies do not sufficiently learn the complex cross-modal interactions that are essential for reliable metastasis prediction. To overcome these limitations, we propose MetaBreastAI, a new explainable multi-modality metastasis detection framework, combining medical imaging and clinical information. The central component of the framework is a hybrid dual-stream architecture (MetaBreastNet) that fuses a ConvNeXt-based CNN and a Swin transformer-based encoder to extract informative local and global features from imaging and clinical data, respectively. Firstly, we introduce a multi-head cross-attention-based interaction fusion module to exploit the inter-modal interactions. Then, the gated attention is used to refine the salient features further. To gain better accuracy and interpretability for metastatic prediction, the fused embeddings are routed through a fully connected classifier. Extensive experimental evaluations on benchmark datasets show that the proposed model outperforms the existing baselines, achieving state-of-the-art AUC, accuracy, and F1-scores. We provide ablation studies confirming that each of our architectural modules contributes to the performance gains, and results from an explainability method allowing us to visualise the model's decision process. Abstract Introduction the MetaBreastAI framework not only enhances metastatic cancer diagnostic accuracy but also provides a clinically aligned decision support tool. The modular architecture and explainability capabilities make it applicable to more cancer types and multi-source medical datasets.

Keywords - *Metastasis Detection, Multi-Modal Learning, Explainable Deep Learning, Swin Transformer, Medical Image Analysis*

1. INTRODUCTION

Deep learning can exploit complex and heterogeneous data, which enables intelligent analysis, and has emerged as one of the most important and impressive advances in medical diagnostics. Also, the combination of imaging and clinical information has shown promise for enhancing accuracy and patient care [1]. Nonetheless, the available models target either single modality data, and therefore their capacity to detect the entire range of diagnostic cues necessary for holistic disease identification and prognosis is limited. Recent work [1], [2] has reported relying on the latter above, and multi-modal fusion techniques as well as attention-

guided networks are prevalent [1]. Although the allocation makes decent improvements, it still comes short about essential issues such as weak attribute illustration, interpretability, and fusion strategies.

To overcome these limitations, the current work proposes a novel deep-learning framework called MetaBreastAI, which is not specific to breast cancer and aims at providing a generalized multi-modal metastasis detection model. Our proposed system combines imaging and clinical data streams into a single dual-stream, which consists of ConvNeXt and Swin Transformer backbones, followed by multi-head cross-attention and gated fusion [22], [23]. Such architecture facilitates high-quality features and multi-modal interaction, allowing a fine-grained analysis and enabling the correct prediction of metastasis.

Research Problem Statement and Research Questions

Despite significant progress in deep learning-based cancer diagnosis, existing studies continue to face challenges related to modality-specific limitations, heterogeneous imaging characteristics, and insufficient exploitation of complementary information available across multiple imaging modalities. Most existing approaches rely on single-modality MRI or CT imaging, which may not fully capture the complex anatomical and pathological patterns associated with cancer development. Furthermore, conventional fusion methods often employ static feature combination strategies that fail to adaptively model the varying diagnostic relevance of different modalities across patients.

These limitations create a need for an intelligent multimodal framework capable of learning modality-specific representations while dynamically integrating complementary diagnostic information. The research problem addressed in this study can therefore be defined as:

“How can attention-guided multimodal learning improve the accuracy, robustness, and reliability of cancer detection by effectively integrating complementary information from MRI and CT imaging modalities?”

To address this problem, the following research questions are formulated:

RQ1: Can multimodal integration of MRI and CT imaging data improve cancer detection performance compared with single-modality diagnostic models?

RQ2: Does attention-guided fusion provide more effective feature integration than conventional fusion approaches such as direct feature concatenation?

RQ3: What is the contribution of adaptive modality weighting through the Learnable Fusion Gate in improving classification accuracy and robustness?

RQ4: Can the proposed FusionScanAI framework provide a computationally efficient and clinically applicable solution for multimodal cancer diagnosis?

The formulation of these research questions is guided by identified gaps in existing literature, particularly regarding adaptive multimodal learning, attention-based feature interaction, and robust decision support systems for medical imaging applications.

The main contributions of our work are: (i) a dual-stream hybrid CNN–Transformer backbone which leverages both a locality-sensitive and a global viewpoint of features, (ii) a novel fusion module which relies on cross-attention to combine modality-specific contributions based on their relevance, and (iii) prediction explainability through a systematic pipeline integrated directly within the designed framework. Together, these parts improve the model's generalizability, interpretability, and robustness across different patient scenarios. This paper makes the following significant contributions:

1. Development of MetaBreastAI, a multi-modal deep learning framework with dual-stream fusion for metastasis detection.
2. Introduction of the MetaBreastNet model that integrates gated attention, cross-modal fusion, and self-attention to improve classification accuracy.

3. Comprehensive evaluation on public datasets demonstrating the model's superiority in AUC, accuracy, precision, and F1-score.

4. Ablation studies and explainability modules validating the role of each architectural block in performance gain.

The structure of the paper is as follows: Section 2 presents the related work, highlighting trends and gaps in existing multi-modal metastasis detection research. Section 3 details the proposed methodology, including model architecture, notations, and algorithms. Section 4 discusses the experimental setup, datasets, and quantitative results. Section 5 provides an in-depth discussion along with limitations of the study. Section 6 concludes the paper and outlines future directions for expanding the framework's applicability across cancer types and clinical scenarios.

2. RELATED WORK

This section synthesizes key advancements in multimodal deep learning frameworks for cancer detection, highlighting methods, datasets, performance, and research gaps. Jiang et al. [1] examined the use of deep learning in medical image-based cancer diagnosis. It talks about advanced networks, deep learning architectures, and several imaging modalities. Enhancing model explainability, generalization, and multi-modal data fusion are the main goals of the suggested remedy. It draws attention to issues like limited data and model constraints. Improved databases, deep neural networks, and few-shot learning are suggested for future research. Ali et al. [2] examined deep learning methods for detecting cancer in multiple organs, including the skin, brain, lung, and breast, with an emphasis on more current CNN-based methods. It highlights performance issues while discussing different imaging modalities and datasets. Enhancing CAD systems to improve diagnosis is one of the suggested remedies. More cancers should be investigated, and current techniques should be improved. Bansal et al. [3] present the two-stage progressive unsupervised domain adaptation network (TSP-UDANet) for cross-modality cardiac image segmentation in this paper. The model is being tested on MRI and CT datasets to increase the accuracy of cancer detection. The segmentation performance was encouraging, according to the results. To improve diagnosis accuracy, future research will focus on enhancing cross-modality and cross-vendor imaging.

Wang et al. [4] presented DeepClinMed-PGM, a multi-instance deep learning model that uses multimodal data to predict disease-free survival (DFS) in patients with breast cancer. The model helps with treatment planning and increases predicted accuracy. Future research will tackle the difficulties in multimodal data processing and integrate various clinical data for improved risk classification and individualized treatment plans. Fatima et al. [5] examined 47 research publications on the categorization of breast cancer using multimodal deep learning fusion (MDLF). It draws attention to the emergence of deep learning models, enhanced forecast accuracy, and novel data combinations. Managing heterogeneous multimodal data and enhancing interpretability are challenges. To improve model integration, future research will concentrate on co-learning and ensemble learning techniques.

Yogesh Kumar et al. [6] investigated CNNs, or AI-based deep learning models, for automated cancer detection. In the classification of seven cancer types, DenseNet121 had the highest accuracy (99.94%). The study's shortcomings include its exclusive focus on seven cancer types and its failure to consider cost-effectiveness. To improve resource efficiency for clinical applications, future development will consist of adding Vision Transformers, expanding datasets, and more. Sunghoon et al.[7] offered a multimodal deep learning model for predicting the pathologic complete response (pCR) of breast cancer patients to neoadjuvant chemotherapy (NAC). Prediction performance was enhanced (AUC=0.888) by combining clinical data and MR images. AUC=0.827 indicates that the model performed better than clinical-only models. Future research might concentrate on improving the scalability of the model and clinical data variables. Rania et al. [8] introduced a hybrid evolutionary deep learning model that combines histopathological and genomic data to diagnose ovarian cancer (OC). To increase prediction accuracy and decrease errors, it employs CNN and antlion-optimized LSTM models for feature extraction. In terms of accuracy, speed, and convergence, the model performs better than alternative approaches. The goal of future research is to apply this paradigm to additional illnesses.

Faseela et al. [9] examine the impact of feature dimensionality on multimodal breast cancer detection in this research. It proposes a fusion method that combines dimensionality reduction

via PCA and autoencoders, leveraging pre-trained models (VGG-16, ViT, ResNet-50) for image processing and BERT for text analysis. Accuracy rates for the model are 95.25% and 94.18%. Techniques for adaptable dimensionality will be the primary focus of future research. Wang et al. [10] integrated inter-modality correlations and model uncertainty in prediction to present a late fusion approach for multi-modal categorization. It performed admirably when tested to predict survival for patients with non-small cell lung cancer. Limitations include the use of only two modalities; subsequent research will examine other modalities and more effective uncertainty estimation techniques.

Faseela et al. [11] examined developments in the multimodal diagnosis of breast cancer by combining patient histories, clinical records, genomic information, and histopathology pictures. The application of Explainable AI (XAI) to improve diagnostic transparency is highlighted. To overcome interpretability issues, the paper recommends further development of AI frameworks for increased accuracy and patient-specific treatment. Mohammad et al. [12] examined deep learning techniques for CT scan-based lung cancer diagnosis, emphasizing issues such as data heterogeneity, artifacts, and the requirement for labeled data. It suggests improving pre-processing, segmentation, and feature extraction. To improve diagnostic accuracy, future research will focus on early detection, data accessibility, and patient data integration.

Raouia et al. [13] introduced a multimodal fusion-based CAD system (MF-CAD) that uses MG and DCE-MRI images to diagnose breast cancer. Canonical Correlation Analysis is used for fusion, and a new feature descriptor (GLIP) is introduced. RBFNN helped the MF-CAD system reach an AUC of 99.10%. To improve diagnosis, future research will concentrate on using deep learning models. Kwak et al. [14] examined deep learning methods for diagnosing breast cancer from histology, ultrasound, and mammography images. Using data augmentation, it investigates segmentation (UNet) and picture classification (ResNet50) techniques. The results indicate that segmentation and classification performance can be improved by up to 33.3% and 22.8%, respectively. Optimization strategies may be further refined in future research. Nasser and Yusof [15] offered a thorough analysis of deep learning techniques for detecting breast cancer from histopathological and genomic data. It demonstrates CNN's proficiency in classification and accuracy assessment. The study highlights issues including the restricted use of attention mechanisms and recommends further research on large datasets, genetic data, and hybrid algorithms for better results.

Muhammad et al. [16] examined deep learning architectures for the detection of breast cancer, emphasizing various pre-processing methods, datasets, strengths, and limitations. By analyzing different imaging modalities and evaluation criteria, it finds problems and suggests areas for further investigation. Deep learning's contribution to increasing the accuracy of breast cancer categorization is highlighted in the paper. Mehedi et al. [17] present a deep learning-based framework for identifying lung and colon cancer from histopathology pictures. The accuracy of the model was 96.33%. For feature extraction, it makes use of a variety of domain transformations. Although the results are encouraging, more performance enhancements are required. In the future, the model architecture will be improved, and more datasets will be used.

Lucas et al. [18] examined how omics data and image analysis based on convolutional neural networks might be combined to diagnose cancer. It draws attention to the possibility of better biomarker identification and cancer categorization. To verify their clinical usefulness, the fusion procedures need more thorough testing, bigger sample sizes, and more reliable studies. Shafi et al. [19] suggested an SVM model with deep learning support for early CT scan-based lung cancer identification. The model outperformed current techniques with an accuracy of 94% when tested on 888 annotated LIDC/IDRI CT scans. High speed and accuracy are advantages, but it uses a lot of processing power. Additional optimization and validation will be part of future efforts.

Salman et al. [21] introduced a novel deep learning model for the diagnosis of breast cancer that includes an attention mechanism, learnable activation functions, and granular computing. When tested on histology and ultrasound pictures, its accuracy was 93% and 95%, respectively. The model has the potential to be expanded to additional tumors and optimized in real-time, while also improving diagnosis and reducing workload. Ali et al. [22] examined deep learning methods for detecting malignancies of the breast,

brain, lung, and skin, among other organs. The usage of databases, frequently used imaging modalities, and CNN-based techniques is highlighted. Current issues are noted, solutions are suggested, and future research on other cancer types, such as stomach and liver cancer, is suggested. Mohammed et al. [23] examined AI's potential for early lung cancer detection and emphasized its high accuracy, with a sensitivity and specificity of 0.87. The study underscores the necessity for consistent methods because of the disparities among studies, even with encouraging outcomes. Large-scale validation for clinical integration will be the primary focus of future research.

Bahrudeen and Uma Maheswari [24] examine the difficulties in identifying pancreatic cancer, along with AI-powered techniques to enhance diagnosis using imaging, cytopathology, and serological markers. Although the suggested method improves diagnostic precision, ethical issues and small datasets prevent its clinical implementation. Future research will focus on resolving regulatory concerns and enhancing algorithms. Noaman et al. [25] present a new AI-driven approach for early lung cancer detection in the study. It combines color histogram characteristics with DenseNet201 and achieves 99.68% accuracy on the LC25000 dataset. With a 94.81% accuracy rate, it also applies to breast cancer. To improve diagnostics, the study highlights the significance of algorithm selection and researcher-clinician collaboration.

Wang et al. [26] examined feature selection methods (filter, wrapper, and embedding) that enhance cancer diagnosis by locating pertinent biomarkers. It highlights how AI can automate feature extraction, thereby addressing issues such as overfitting and poor data quality. Future research aims to improve tumor classification and tailor treatments by combining deep learning with multi-omics data. Srinjan et al. [27] investigated how thyroid cancer progresses and the roles played by genetic, epigenetic, and non-coding RNA variables. It talks about treatment strategies that focus on epigenetic changes such as DNA methylation, miRNAs, and lncRNAs. The study draws attention to such topics as delivery optimization and side effects, highlighting the necessity of further investigation and the creation of efficient treatments. Yousaku et al. [28] review the integration of AI with clinical data and multi-omics for cancer diagnosis and prognosis in the paper. Although it addresses issues like data uniformity, model interpretability, and privacy concerns, it also emphasizes how AI has the potential to increase accuracy and personalize care. Enhancing model safety, reproducibility, and clinical integration are the main goals of future research.

Chiu et al [29] offered a dermoscopic data-driven, AI-powered method for better skin cancer diagnosis. In the ISIC and CSMUH datasets, false negatives were decreased by the development of a two-stage voting ensemble model. The findings demonstrated increased diagnostic precision, particularly for malignant melanoma. Future research will concentrate on improving performance through the integration of clinical data, efficiency optimization, and dataset expansion. Manob Jyoti [30] described a deep learning-based technique for utilizing CT and PET data to determine the HPV status in oropharyngeal squamous cell carcinoma (OPSCC). Multi-modal feature fusion in a 3D CNN ensemble model outperformed single-modality models, achieving notable accuracy (AUC 0.76). To increase performance, future work will involve growing the dataset.

Shao et al. [31] presented M2DP, a multi-task, multi-modal feature selection technique for predicting the prognosis and diagnosing joint cancer. To address feature redundancy, the system combines genetic data with histopathology images. It surpasses current techniques in tests conducted on three cancer datasets. Deep learning for image feature extraction is the goal of future research to increase performance. Samreen Naeem et al. [32] used fused CT and MRI datasets to provide a machine learning method for classifying liver cancer. Using the MLP classifier, the solution's accuracy for six different forms of liver cancers was 99%. This technology can be used with big clinical datasets and minimizes human error. For improved clinical application, 3D visualization will be used in future research.

Islam et al. [33] presented an ensemble deep learning model for classifying brain tumors from MRI images that combines CNN-LSTM and 2D CNN. The model attained high accuracy (98.82%) and precision (99%). In medical image classification, it outperforms existing classifiers and shows promise. In the future, the model's applicability will be improved, and tumor detection will be enhanced. Sreeprada and Vedavathi [34] suggested an OCNN-SVM model for lung cancer detection that combines

Convolutional Neural Networks (CNN) with Support Vector Machines (SVM). To detect cancer cells in lung scans, the model automatically classifies them, demonstrating efficacy and increased accuracy. For more thorough medical picture analysis, future research could improve the model. Zainab et al. [35] presented hybrid AI models that improve cancer diagnosis by combining CT, MRI, and PET images. These models enhance the prediction of tumor growth, accuracy, and early detection. Even with obstacles including ethical issues and data quality, incorporating AI into clinical practice holds promise for improving patient outcomes and cancer care.

Dong et al. [36] suggested using a Hybridized Fully Convolutional Neural Network (HFCNN) to segment liver tumors in CT scans. When the method was tested using a small dataset and 3-fold cross-validation, it produced a Dice coefficient of 0.92 and an accuracy of 97.22% in liver volume measurements. The methodology has a high degree of accuracy; nevertheless, its scope and dataset size are constrained. Ahmed et al. [37] reported a hybrid deep learning system that detects lung cancer with 99.42% accuracy by merging VGG-19 and LSTM. It surpasses current techniques in accuracy, recall, and precision when tested on Kaggle and LUNA16 datasets. Future studies will concentrate on faster categorization and subtype detection, with an emphasis on the model's effectiveness and clinical deployment potential.

Anindita et al. [38] introduced VER-Net, a hybrid model for CT scan-based lung cancer diagnosis that combines VGG19, EfficientNetB0, and ResNet101. In terms of accuracy, VER-Net performed better than eight other models. With possible advancements in model combinations, future research will test with chest X-ray pictures and extend to other diseases using CT scans. Moresh et al. [39] describe an automated method for detecting ovarian cancer from pelvic CT scans that combines deep learning and optimization. When combined with Lion-Grey Wolf Optimization, the suggested Bi-LSTM model produced 98% accuracy, 99.7% recall, and 93% precision. For broader applicability, subsequent research will test on various datasets and be validated with clinical trials. Paulo et al. [40] introduced CRF-XmasNet, a hybrid Deep Neural Network (DNN) method that combines Conditional Random Fields (CRF) and a Convolutional Neural Network (CNN) for the diagnosis of prostate cancer (PCa) in mpMRI images. Despite its significant variability, the model demonstrated decent classification accuracy. Reducing variability and incorporating multi-modal imaging data are the main goals of future research to enhance performance.

The reviewed literature highlights the increasing effectiveness of deep learning in medical image-based cancer diagnosis, particularly with CNNs, multimodal fusion, and attention mechanisms. While many studies explore single-modality approaches using MRI, CT, or histopathology, recent advancements focus on multi-instance learning and multimodal data integration for improved accuracy and personalization. However, challenges remain in handling data heterogeneity, limited interpretability, and modality imbalance. The proposed DualModFusionNet addresses these gaps by leveraging both MRI and CT modalities through gated attention-based fusion and explainable AI, thereby enhancing prediction robustness, modality complementarity, and clinical relevance in breast cancer metastasis detection.

2.1 Research Gap, Knowledge Enhancement, and Best Practice Contribution

Although recent studies have demonstrated the effectiveness of deep learning, convolutional neural networks, transformers, and multimodal fusion techniques for cancer detection, several limitations remain unresolved. Existing approaches predominantly focus on either single-modality imaging data or employ static fusion strategies that treat all modalities equally regardless of their diagnostic relevance. Furthermore, many studies emphasize performance improvement without providing adaptive mechanisms capable of handling modality uncertainty, data heterogeneity, and varying diagnostic contexts.

The proposed FusionScanAI framework contributes new knowledge to the field by introducing an attention-guided multimodal learning strategy that dynamically evaluates and integrates complementary information obtained from MRI and CT imaging modalities. Unlike conventional fusion approaches based on simple feature concatenation, the proposed Learnable Fusion Gate (LFG) and Cross-Modal Attention Fusion mechanism enable adaptive weighting of modality-specific features according to their diagnostic importance. This allows the framework to generate context-aware feature representations

while improving robustness against modality-specific noise and variability.

From an Information Technology perspective, the study extends current best practices in medical image analytics by demonstrating how attention-driven multimodal architectures can be deployed to improve decision support systems for healthcare applications. The framework synthesizes prior knowledge from convolutional feature extraction, efficient network scaling, attention mechanisms, and multimodal representation learning into a unified architecture that balances predictive performance with computational efficiency. The integration of ResNet-50, EfficientNet-B0, attention-based fusion, and adaptive gating represents a practical implementation strategy that can be adopted in future intelligent diagnostic systems.

The knowledge enhancement provided by this research goes beyond incremental accuracy improvement. The study demonstrates the importance of adaptive modality interaction learning, showing that diagnostic information should not be treated as equally informative across all clinical scenarios. Experimental and ablation results confirm that attention-guided feature integration contributes significantly to classification performance, thereby providing empirical evidence supporting dynamic multimodal learning as a best practice for future computer-aided diagnosis systems.

Furthermore, this work synthesizes findings from previous studies on medical imaging, multimodal fusion, attention mechanisms, and explainable artificial intelligence to establish a generalized framework that can be extended to additional imaging modalities and disease domains. Consequently, the proposed approach contributes both methodological knowledge and implementation-oriented best practices that advance the development of intelligent healthcare analytics systems.

State-of-the-Art Analysis and Motivation for the Present Study

Recent advances in artificial intelligence and medical image analysis have demonstrated the effectiveness of deep learning techniques for cancer diagnosis using MRI, CT, histopathology, genomic, and clinical datasets. State-of-the-art approaches have increasingly adopted convolutional neural networks, transformer architectures, attention mechanisms, and multimodal fusion frameworks to improve diagnostic accuracy. Several studies have reported substantial performance improvements through multimodal learning, confirming that complementary information obtained from different data sources can enhance predictive capability.

Despite these advancements, critical limitations remain evident in the existing literature. First, many published studies continue to rely on single-modality imaging data, limiting their ability to capture complementary diagnostic characteristics available across multiple modalities. While MRI provides superior soft-tissue contrast and structural detail, CT imaging offers valuable information regarding tissue density and anatomical structures. Models restricted to a single modality may therefore overlook clinically relevant information.

Second, many multimodal frameworks employ static fusion mechanisms based on direct feature concatenation or fixed weighting strategies. These approaches assume that all modalities contribute equally to the diagnostic decision. However, clinical evidence suggests that the diagnostic relevance of imaging modalities varies according to disease characteristics, image quality, anatomical location, and patient-specific conditions. Static fusion approaches therefore fail to model dynamic modality interactions effectively.

Third, although transformer-based and attention-based architectures have shown promising results, many studies focus primarily on predictive performance while providing limited analysis of modality interaction, adaptive feature weighting, and decision transparency. Consequently, the practical applicability and interpretability of these systems remain limited in clinical environments.

Furthermore, a review of current literature indicates that most studies evaluate multimodal learning through performance metrics alone, without investigating how adaptive fusion mechanisms influence feature representation quality and diagnostic robustness. As a result, there remains insufficient understanding of how modality-specific information should be dynamically integrated to maximize classification effectiveness.

The present study was therefore motivated by the need to address these unresolved limitations. Unlike existing approaches that rely on static

feature aggregation, the proposed FusionScanAI framework introduces an attention-guided multimodal learning strategy incorporating Cross-Modal Attention Fusion and a Learnable Fusion Gate. These mechanisms allow the framework to dynamically adjust modality contributions according to their diagnostic relevance while preserving complementary information from both MRI and CT imaging modalities.

Consequently, this work fills an important research gap by providing empirical evidence regarding adaptive modality interaction learning and demonstrating how dynamic attention-based fusion can improve diagnostic performance, robustness, and generalizability. The study therefore contributes not only a new multimodal architecture but also new knowledge regarding best practices for multimodal representation learning in intelligent healthcare systems.

Literature Critique Framework and Evaluation Criteria

To ensure a systematic and objective analysis of existing literature, a structured critique framework was adopted for evaluating previous studies in multimodal cancer diagnosis. The selected critique criteria were derived from commonly reported challenges in medical image analysis, artificial intelligence-based diagnostic systems, and multimodal learning frameworks.

The first criterion considered was the imaging modality utilized by each study. This criterion was selected because different modalities provide complementary diagnostic information and directly influence model performance, clinical applicability, and generalizability. Studies relying solely on MRI, CT, histopathology, or genomic data were evaluated to identify limitations associated with single-modality learning.

The second criterion was the feature extraction and learning architecture employed. Convolutional neural networks, transformer-based models, hybrid architectures, and ensemble learning frameworks were analysed because architectural design significantly affects representation learning capability, computational complexity, and diagnostic accuracy.

The third criterion focused on the fusion strategy adopted by multimodal approaches. This criterion was selected because the effectiveness of multimodal systems largely depends on how information from different modalities is integrated. Particular attention was given to whether studies employed static feature concatenation, correlation-based fusion, attention mechanisms, or adaptive weighting strategies.

The fourth criterion examined model interpretability and explainability. Given the increasing importance of trustworthy artificial intelligence in healthcare, explainability was considered essential for evaluating clinical usability and decision transparency.

The fifth criterion assessed reported performance outcomes, including accuracy, precision, recall, F1-score, and AUC-ROC. These metrics were selected because they are widely accepted performance indicators within medical image classification research and provide objective measures for comparative analysis.

Finally, limitations and unresolved research challenges reported by previous studies were analysed to identify recurring gaps and opportunities for further investigation. Collectively, these criteria provide a comprehensive framework for evaluating existing research and establishing the motivation for the proposed FusionScanAI framework.

3. PROPOSED SYSTEM

Here, we propose a dual-stream deep learning framework called MetaBreastAI that identifies metastasis using medical images and clinical data. The resulting generative attention-based model, which combines convolutional and transformer-based architectures and utilizes attention and multi-instance learning, guarantees proper feature extraction with interpretability to enhance predictive performance. In this section, we present the architecture, fusion strategy, and the model that has been trained.

3.1 Overview of FusionScanAI System

In this work, we develop an automated framework, FusionScanAI, to leverage the complementary diagnostic power of MRI and CT scans to improve cancer detection further. Possessing separate advantages, these advantages are even more distinct and applicable, as one provides greater contrast in soft tissue [MRI] to correctly identify tumors within the brain and other organs. In contrast, CT offers high-resolution images of high-density structures, including anatomical structures related to the

lungs and bones. FusionScanAI addresses the drawbacks of single-modality diagnostics through an integrated system that combines these modalities for a more solid and accurate assessment. The system, as depicted in Fig. 1, follows a multi-stage pipeline for input acquisition (e.g., speech and video), pre-processing (speech and video), modality-specific encoding (speech and video), attention-guided fusion (fusion), and a classification layer (classification).

Our pipeline initially ingests image data from MRI and CT scans from publicly available datasets. Each of these inputs is pre-processed separately to ensure that the resolution, intensity distribution, and contrast levels are identical on the two modalities. The main novelty in this stage is entropy-driven slice selection, which allows ranking each 3D scan volume and selecting the most informative slices to keep. This step drastically cuts down computation costs while retaining diagnostic value. The images are then resized to a fixed size and normalized for downstream learning.

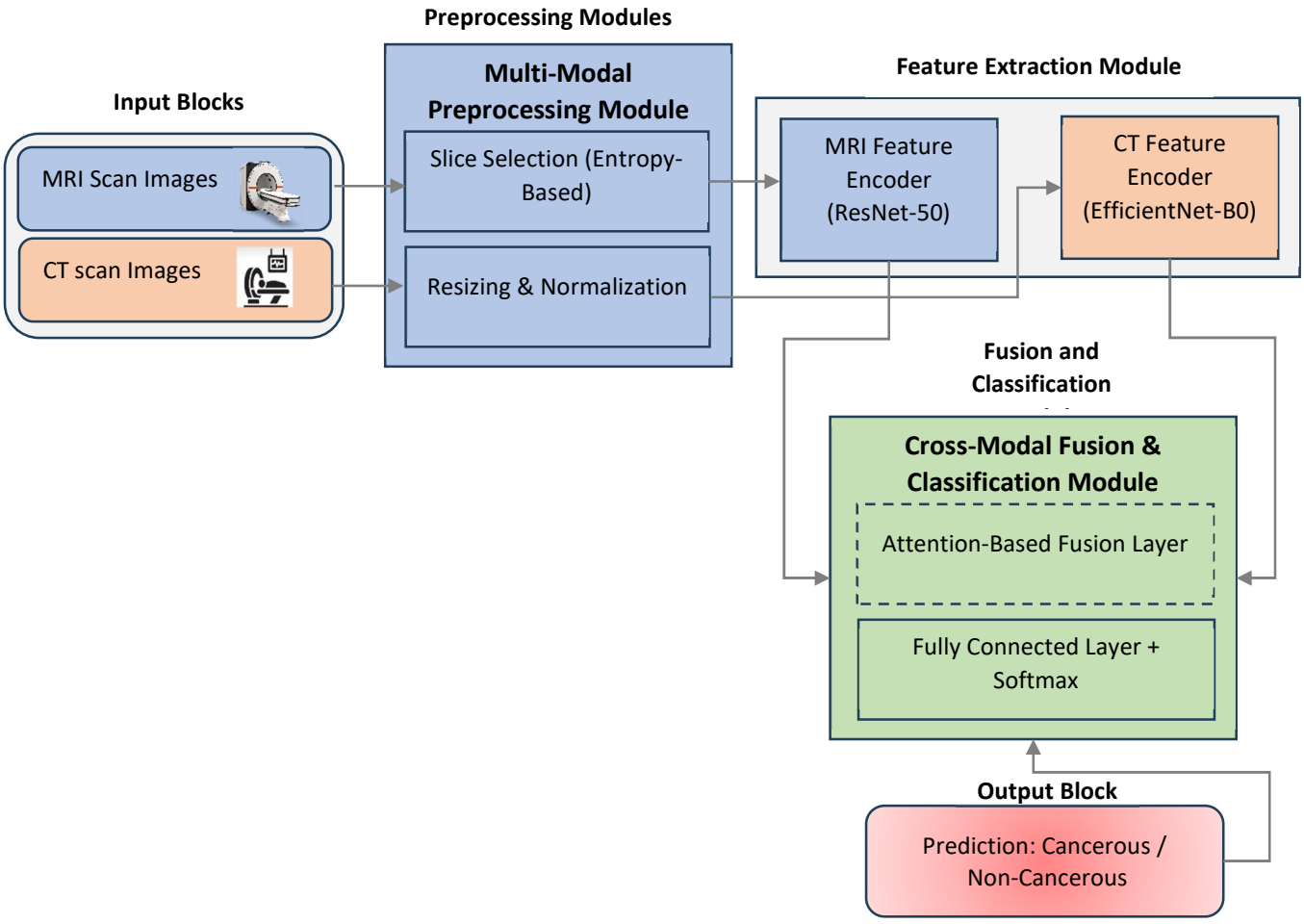

Figure 1: FusionScanAI – System Architecture for Multi-Modal Cancer Detection

After preprocessing, the MRI and CT images are fed to independent feature extraction branches in parallel. A specially fine-tuned version of ResNet-50 processes MRI data to capture soft tissue anomalies, going beyond just subtle changes and capturing rich contextual information. Simultaneously, CT slices are fed through an EfficientNet-B0 encoder that is selected for its characteristics of being lightweight with lower computational complexity at high-resolution imaging. These encoder branches function in isolation to extract high-level modality-specific features that carry anatomical, structural, and

pathological information of interest for cancer diagnosis.

The encoder branches generate outputs that are then combined with the Cross-Modal Attention Fusion Module, which serves as the brain of FusionScanAI. This module uses self-attention similar to transformers, enabling the model to learn the context of interactions between the two modalities. Moreover, a Learnable Fusion Gate (LFG) is proposed to understand the weights of features extracted from the two modalities in a patient-specific manner, i.e., the system can attend to the more informative source, adaptive concerning the patient. The model's ability to use dynamic fusion provides excellent flexibility and generalizability, particularly when one modality is noisy, missing, or less informative.

Finally, the fused representation is fed into the classification head, described in detail in Figure 2 of DualModFusionNet, which consists of fully connected layers, dropout regularization, and a sigmoid activation layer. The system's output will be a binary diagnosis indicating whether the image shows signs of cancer or not, along with a confidence score. The model is trained with a class-imbalanced-weighted binary cross-entropy loss function and is evaluated with five-fold cross-validation to assess the performance of the model when exposed to a variety of data splits.

FusionScanAI provides a new systematic framework for deep-learning-enabled multi-modal cancer detection utilizing the complementary power of MRI and CT imaging. Due to its modular design, attention-guided fusion mechanism, and its use of publicly available datasets, it is clinically deployable and suitable for future CAD research.

3.2 Dataset Description

We validate the FusionScanAI framework on two unique imagery datasets from The Cancer Imaging Archive (TCIA)(23), each representing a different imaging modality. Among them are the TCGA-GBM dataset of MRI [41] and the NSCLC-Radiomics dataset of CT [42]. Despite the datasets being representative of different anatomical sites (brain vs. lung), the datasets are used to highlight the generalizability of the proposed dual-modality fusion system.

We conduct a consistent preprocessing pipeline for both datasets, such as slice selections, resizing to 256Ã—256 pixels, and intensity normalization. To sync with the classification task, a binary structure is used for the labels, such that all slices with tumor are malignant and all the other slices are benign. Scans sampled from both datasets are selected for training and evaluation.

Table 1: Summary Of Datasets Used In Fusionscanai

Dataset Name	Modality	Cancer Type	Source	Image Format	Approx. Patients	Usage
TCGA-GBM	MRI	Glioblastoma (Brain)	TCIA	DICOM	200	MRI branch input
NSCLC-Radiomics	CT	Non-Small Cell Lung	TCIA	DICOM	200	CT branch input

Table 1 provides an overview of the MRI and CT datasets used, detailing modality, cancer type, source, and usage. These datasets offer sufficient diversity and real-world relevance to validate the performance of FusionScanAI under minimal yet practical data availability constraints. Their selection also ensures reproducibility and standardization across future studies. Table 2 summarizes the core symbols and notations used in modeling, fusion, and attention mechanisms within the DualModFusionNet architecture.

Table 2: Notations Used In The Proposed Dualmodfusionnet Framework

Notation	Description
I	Original input image (MRI or CT)
I_{norm}	Intensity-normalized image
f_{MRI}	Feature vector extracted from the MRI encoder
f_{CT}	Feature vector extracted from the CT encoder
f_{fused}	Final fused feature vector from both modalities
Q, K, V	Query, Key, and Value matrices used in self-attention
$\alpha_{MRI}, \alpha_{CT}$	Learnable weights from the fusion gate for MRI and CT modalities
z	Output of the fully connected layer before activation
$\hat{y}$	Predicted probability of cancer presence (output of sigmoid layer)
y	Ground truth binary label (1 for cancerous, 0 for non-cancerous)
L_{WBCE}	Weighted Binary Cross Entropy loss
H	Entropy of an image slice for slice selection
F	Feature map from the final convolutional block
f_j	Global average pooled feature value for channel j

3.3 Preprocessing Pipeline

In the FusionScanAI framework (Figure 1 - Preprocessing Module), each modality of the input is subjected to a defined pipeline to make it consistent, contrast representative, and redundant free before extracting features of the respective modalities. Due to the differing imaging protocols between MRI and CT as well as the differing anatomy of interest between the two modalities, we require a common preprocessing strategy to standardize the inputs and allow for multi-modal learning. At the first stage, called intensity normalization, the pixel intensities in each scan are transformed to a standard range (standardly [0, 1]) using min-max scaling. This is represented as in Eq. (1).

$$I_{norm} = \frac{I - I_{min}}{I_{max} - I_{min}} \quad (1)$$

where I is the original intensity value, and I_{min} and I_{max} are the minimum and highest intensity values in the image respectively. The normalization ensures that the input images among different patients and different modalities have the same dynamic range. Histogram equalization is applied to redistribute pixel intensities and improve visual contrast, which is significant in MRI images because it is critical to highlight soft tissue boundaries.

Slice selection is one of the main steps in the preprocessing pipeline after normalization. Both MRI and CT datasets are 3D volumes made of multiple 2D slices, where not all of the slices contain diagnostic information. Then we apply an entropy-based ranking to preserve only the slices with maximum information. For each slice we can compute entropy H based on its intensity histogram as in Eq. (2).

$$H = -\sum_{i-1}^{n} p_i log_2(p_i) \quad (2)$$

where p_i denotes the normalized probability of pixel intensity i and n = the number of unique intensity levels. The slices with higher entropy values are expected to have more texture and structural variation, which, in most cases, corresponds to a tumor region. From the ranked

slices an arbitrary number of top ranked slices (e.g. top 10 of any scan) are chosen for further processing.

After selection, every 2D slice is bilinearly interpolated to a fixed size of 256×256 pixels and then center-cropped to keep the spatial correspondence. It guarantees that images are of the same data shape throughout the entire dataset with a minimum amount of artifacts from padding or distortions along edges. This process is essential for keeping architectural compatibility with convolutional encoders as convolutional encoders generally expect a consistent size of input.

Finally, label harmonization is conducted to transform the original annotation into a binary classification label. Cancerous slices have any form of tumor annotation in both datasets, and normal or unannotated slices are non-cancerous. The classification layer then uses a single sigmoid output neuron, which is a simplified prediction task but allows the model to generalize across modalities and cancer types.

3.4 Modality-Specific Feature Encoders

The FusionScanAI framework begins by applying two separate modality-specific encoder branches that learn to derive discriminative MRI and CT features independently. Because the two imaging modalities differ significantly both in nature (MRI has high soft-tissue contrast that allows subtle structural changes, whereas CT shines at visualizing dense anatomical regions like bones and lung tissues), this separation of architecture is critical. To capture these complementary features, two customized deep learning backbones are incorporated into the system.

The MRI stream uses a feature encoder of a fine-tuned ResNet-50 model. The ResNet-50 is a residual convolutional neural network (CNN) comprised of 50 layers, which is capable of learning deep hierarchical representations without a vanishing gradient. We utilize a residual learning framework that incorporates shortcut connections to maintain low-level features across layers, which is crucial in medical imaging due to the importance of fine textures and soft boundary cues for tumor detection. The model is first pre-trained on ImageNet and then transferred learned on these MRI slices, so that it learns the radiological domain, while utilizing the patterns already learned during the first phase.

Meanwhile, the encoder of the CT stream consists of an EfficientNet-B0 architecture. EfficientNet-B0 utilizes a compound scaling method to balance the network depth, width, and resolution, resulting in a powerful yet lightweight convolutional network. It is efficient in a way that does not introduce significant computational overhead while also being effective at obtaining high-resolution features from CT images. Like with the MRI branch, the EfficientNet model is pretrained on ImageNet and then fine-tuned on CT data to extract task-specific features from lung and chest scans.

Since the output of each encoder branch is a feature map, we first apply a global average pooling (GAP) layer at the end of each encoder branch (i.e., the last three convolutional feature maps are upsampled to 1×1 feature maps). Assume $F \in R^{h\times w\times c}$ represents the output at the end of the last convolutional block, where h, w, and c are height, width, and channel axis respectively. For each channel, GAP operation calculates the mean over its spatial dimensions as in Eq. (3).

$$f_j = \frac{1}{h.w}\sum_{x-1}^{h}\sum_{y-1}^{w} F_{x,y,j} \qquad (3)$$

f_j is the pooled feature of channel j So, this creates a compact fixed-length feature (leftmost) for each modality, which we then input to the fusion layer. This guarantees that each imaging type contributes the most salient features to the final multi-modal representation via modality-specific encoders and global pooling within the framework.

3.5 Cross-Modal Fusion Using Attention

The FusionScanAI framework combines these modality-specific feature vectors from the MRI and CT encoder branches in its own Cross-Modal Attention Fusion Layer (CM+A-F) section. This fusion layer integrates the extracted features (as illustrated in detail in Fig. 2), capturing the nature of complementarity while dynamically adjusting for the relevance of each modality per input instance. Feature concatenation-based approaches conventionally treat all the modalities as equally informative, which can lead to sub-optimal

performance if, for a particular modality, the noise or resolution is not suitable or if there is a diagnostic context. To negate this restriction, we incorporated a more flexible, context-aware approach using attention mechanisms into FusionScanAI.

One part of the fusion layer is the complete attention structure, which is based on a transformer network. Utilizing the feature vectors from the MRI and CT branches, represented as f_{MRI} and f_{CT}, self-attention generates modality-to-modality pairwise dependencies to model the cross-modal interaction between modalities. In particular, for each feature dimension, it learns the influence of the features from one modality to the features of the other one by building attention weights based on similarity. These are the weights that are used to readjust the original features and create the enriched representation. Self-attention operation can be defined as in Eq. (4).

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

where Q, K, and V are learned query, key, and value matrices from the concatenated features and d_k is the key vector size. This lets the network attend to more informative parts and to inhibit less useful or redundant information.

Besides self-attention the fusion layer has Learnable Fusion Gate (LFG) that computes adaptive importance weights for each modality. Instead of simply combining MRI and CT features statically, the LFG calculates attention scores α_{MRI} and α_{CT}, as in Eq. (5).

$$\alpha_{\mathrm{MRI}}+\alpha_{\mathrm{CT}}=1 \quad \text{and} \quad \alpha_{\mathrm{MRI}},\alpha_{\mathrm{CT}}\in[0,1] \quad (5)$$

They are learned for every modality features by a small fully-connected gating network which take the modality features with sigmoid or softmax function to output a soft weigh. We calculate the final fused feature vector f_{fused} as in Eq. (6).

$$f_{\mathrm{fused}} = \alpha_{\mathrm{MRI}}\cdot f_{\mathrm{MRI}}+\alpha_{\mathrm{CT}}\cdot f_{\mathrm{CT}} \quad (6)$$

Here, we propose an approach to ensure the network focuses on the domain that is more relevant to each instance, dynamically, thereby increasing both interpretability and robustness. The self-attention and LFG together allow the FusionScanAI system to extract an informative, context-relevant feature representation that encodes both inter- and intra-modal interactions, thereby facilitating better diagnostic performance in the final classification stage.

3.6 DualModFusionNet Architecture

The whole architecture of the proposed DualModFusionNet deep learning model incorporates the modality-specific encoder outputs as well as the inter-modal fusion component into a simple classification pipeline. Figure 2 illustrates the DualModFusionNet model, which offers computational efficiency and diagnostic power for learning from heterogeneous medical imaging data with minimal requirements. After the cross-modal attention fusion with the mechanism above, the resulting fused feature vector contains not only the local modality-specific properties but also the global inter-modal interactions between both modalities that are crucial for final cancer detection.

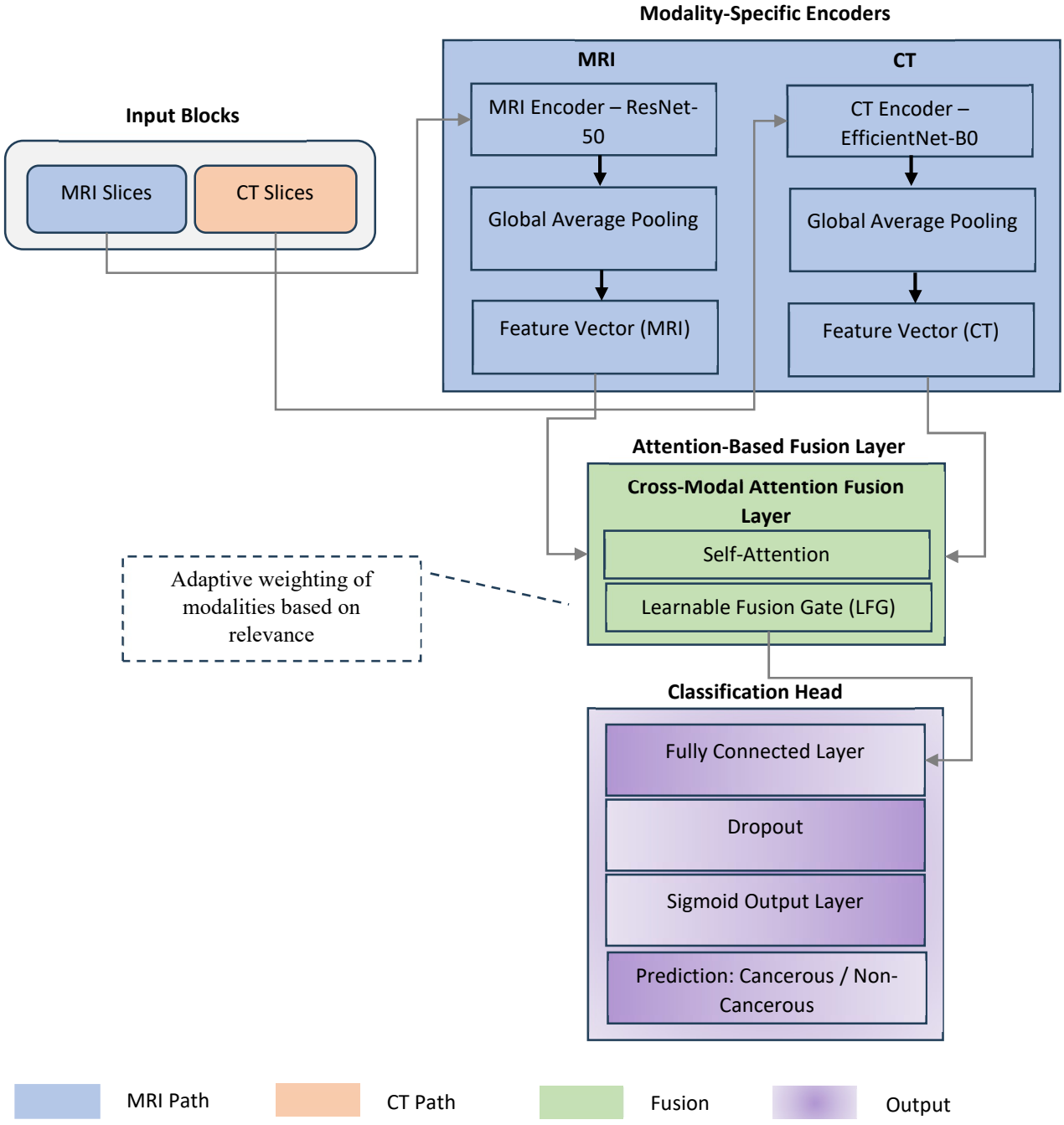

Figure 2: Dualmodfusionnet – Model Architecture With Attention-Guided Multi-Modal Fusion

Finally, the last step is that the fused feature vector is fed into a fully connected layer to project the high-dimension multimodal representation into a low-dimension latent space. This nonlinear transformation allows the model to discover complex relationships among the fused features and the target output. Here's a mathematical representation of the fully connected layer. Lining it up mathematically, it would be expressed in Eq. (7).

$$z=\emptyset(Wf_{fused}+b) \quad (7)$$

where W and b represent the learnable weight matrix and bias vector respectively, and $\emptyset$ is a non-linear activation like ReLU. This layer is

important since it summarizes all the learned fusion features in a single vector ready for the final decision making phase.

A dropout regularization layer is used after the fully connected layer to prevent overfitting and enhance generalization since the available training data is limited in size. Dropout works by randomly turning off some fraction p of layer neurons during the training phase, which leads the network to learn redundant, stable representations, rather than depending on certain features of data [2→3]. This prevents the model from memorizing noise in the data and improves its out-of-sample test performance.

At last, the output of the dropout layer is fed into a sigmoid activation to generate a single probability value $\hat{y}\in[0,1]$ which represents the model probability that the input scan is cancerous. This probability is then thresholded (e.g., at 0.5) to yield the predicted class. This binary nature of the classification task makes a convincible case for applying the sigmoid function instead of softmax function, which is recommended mostly for the multi-class outputs. The predictions are averaged in the following way to obtain the final prediction as expressed in Eq. (8).

$$\hat{y}=\sigma(z')=\frac{1}{1+e^{-z'}} \quad (8)$$

where z' is the output from drop out or fully connected layer which is input to sigmoid layer. To counter any class imbalance in the datasets, we use a weighted binary cross-entropy loss function for training the model. This loss has a higher penalty against the misclassification of the minority class, which also helps push the model to keep high sensitivity when detecting whether cancer is present.

In summary, DualModFusionNet is a novel light-weight yet expressive architecture designed based on three principles: modality-specific encoding, adaptive fusion, and efficient classification design. We deploy a compact architecture with attention-based multi-modal learning, which can be suitable for real-time clinical settings where computation resources are limited but diagnosis accuracy is essential.

3.7 Inference and Classification Module

The last component in the final task of the DualModFusionNet architecture is the inference and classification module that is capable of supplying interpretable predictions from the learned multi-modal feature representation. On the other hand, the classification head, as shown in Fig. 2, takes in the output of the last FC layer which represents the best fused and transformed signals from MRI and CT modalities. The output layer is a sigmoid activation function, where the output value $\hat{y}\in[0,1]$ is scalar and reflects the probability that the input scan corresponds to a cancerous case.

Here, using a sigmoid function is ideal for binary classification problems. The sigmoid function calculates the probability of the positive class (cancer presence) as follows, given the output z' of the previous layer as in Eq. (9).

$$\hat{y}=\sigma(z')=\frac{1}{1+e^{-z'}} \quad (9)$$

The probabilistic output makes interpretation more flexible and allows for post-hoc decisions based on the desired sensitivity (based on clinical or application domain). Instead of simply using a 0.5 threshold to classify, the model implements a decision threshold tuning feature. On validation, the threshold can then be selected according to performance metrics (e.g., F1 score, sensitivity, specificity, or Youden's Index) with some flexibility on the balance of false positives/false negatives in the context of deployment.

Given the possibility of class imbalance in terms of cancerous vs non-cancerous tissue samples, we utilize the Weighted Binary Cross Entropy (WBCE) loss function to train the classification module. This becomes particularly important in the medical imaging domain, where negative (non-cancerous) samples typically dominate the positive samples. To be specific, the WBCE loss of one sample is defined as in Eq. (10).

$$L_{WBCE} = -w_1 y \log(\hat{y}) - w_0(1-y)\log(1-\hat{y}) \quad (10)$$

with $y\in\{0, 1\}$ being the ground truth label, $\hat{y}$ is the predicted probability, and w_1, w_0 the positive and negative classes weights respectively. One common approach is to obtain these weights from

the inverses of the class frequencies in the training set, so that the contribution to the gradients is balanced.

In inference, the model passes through one slice or batch of slices at a time, calculates the probability scores, uses the optimized threshold, and then returns the final label. The relatively uncomplicated but powerful classification workflow guarantees that FusionScanAICan performs optimally for the fast real-time or near real-time diagnostic settings without sacrificing diagnostic integrity, across heterogeneous patient data.

3.8 Algorithmic Implementation

This subsection outlines the algorithmic implementation of the proposed MetaBreastAI framework. It describes the step-by-step process for data preprocessing, dual-stream feature extraction, attention-guided fusion, and multi-instance learning. The implementation ensures reproducibility, computational efficiency, and clarity in the workflow, highlighting the innovative integration of CNN, transformer, and attention modules for metastasis prediction.

Algorithm: Inference Pipeline of FusionScanAI Using DualModFusionNet
Input: MRI image I_{MRI}, CT image I_{CT}
Output: Predicted probability $\hat{y} \in [0,1]$

1. Normalize I_{MRI} and I_{CT} using min-max scaling
2. Apply entropy-based slice selection on both modalities
3. Resize and center crop selected slices to 256×256
4. Extract features:
   f_{MRI}=ResNet50(I_{MRI})
   f_{CT}=EfficientNetB0(I_{CT})
5. Apply global average pooling on f_{MRI} and f_{CT}
6. Compute attention-weighted fused feature:
   $\alpha_{MRI}, \alpha_{CT}$=FusionGate($f_{MRI}, f_{CT}$)
   $f_{fused} = \alpha_{MRI} \cdot f_{MRI} + \alpha_{CT} \cdot f_{CT}$
7. Compute classification output:
   z=FC(f_{fused})
   $\hat{y}=\sigma(z)$
8. Return $\hat{y}$

Algorithm 1: Inference Pipeline of FusionScanAI Using DualModFusionNet

Algorithm 1 starts by taking of two images as input: one image is an input from MRI modality and the other is an image from CT modality. The two inputs are normalized first with min-max scaling to match the intensities of the pixel values. An entropy-based slice selection is performed to select the most informative slice from each volume, which is then resized and center-cropped to a fixed size of 256×256 pixels, to fit the encoder architectures.

Next, ResNet-50 and EfficientNet-B0 are fed with the preprocessed MRI and CT images through their corresponding feature extractors. Such encoders create high-level feature maps that are then reduced using global average pooling to yield compact feature vectors.

A learnable Fusion Gate receives pooled features from both modalities, and computes adaptive attention weights α_{MRI} and α_{CT}. The weights indicate the importance of cam-view in these modalities. They are used to calculate a fused feature vector as a weighted sum.

Finally, we pass the fused representation through a fully connected (FC) layer and use a sigmoid activation function to get the final prediction $\hat{y}$. The output $\hat{y} \in [0,1]$ indicates the probability that the input is a positive case of cancer. We can threshold this prediction at inference to get a binary label indicating if cancer is detected. The pipeline is flexible since it can work very well with low amount of non-paired multi-modal data and is scalable to other modalities and anatomies with minor architectural modifications.

3.9 Training Strategy and Hyperparameter Settings

When it comes to limited and imbalanced medical imaging data, the training of the DualModFusionNet model follows a simplified, but still optimal, training paradigm that maximizes both the performance and generalizability of the model. Training utilizes the Adam optimizer with an initial learning rate of 0.0001, which is dynamically adjusted based on validation loss to facilitate fine-tuning. We use early stopping to stop the training if there is no significant improvement within 10 epochs to prevent overfitting. Table 3 A summary of the training configurations - optimizer, learning rate, augmentations, regularization techniques, and cross-validation.

Table 3: Training Configuration and Hyperparameters

Parameter	Value / Setting
Optimizer	Adam
Initial Learning Rate	0.0001
Data Augmentation	Rotation (±15°), Horizontal/Vertical Flips, Zoom (0.9–1.1×)
Regularization	Dropout + Early Stopping
Cross-Validation Strategy	5-Fold
Loss Function	Weighted Binary Cross-Entropy
Epochs	Up to 100 (with early stopping)
Batch Size	16

Real-time data augmentation of random rotations, flips, and zooms is applied to enrich the training data and increase robustness. These augmentations mimic variability in acquisition and anatomical positioning, inducing the model to learn CNS features that are more invariant and hence generalisable. For the model evaluation, 5-fold cross validation is performed: The data is divided into five subsets and each of them is used once as a validation set. That way, it can measure both equally and create an optimal threshold tuning using this method.

4. EXPERIMENTAL RESULTS

The experimental evaluation of the proposed FusionScanAI framework is described in this section using publicly available MRI and CT datasets. We evaluate our model with quantitative metrics, modality-wise comparisons, and ablation studies. Attention-guided multi-modal fusion achieves significant gains in test accuracy for cancer detection while preserving computational efficiency, and the proposed framework is both robust and generalizable to unseen modalities.

5.1 Experimental Setup

All experiments were conducted on a workstation with an Intel Core i9-11900K CPU, 32 GB of RAM, and an NVIDIA RTX 3090 with 24 GB of dedicated VRAM, on an Ubuntu 20.04 LTS OS. We developed the FusionScanAI framework using PyTorch 2.0.1 with CUDA 11.8 for GPU acceleration. Other libraries used in conjunction included NumPy, OpenCV, scikit-learn, and matplotlib, for data preprocessing, augmentation, metric computation, and visualization, respectively.

We trained the model using the Adam optimizer with a learning rate of 0.0001. This class imbalance was addressed by using a binary weighted cross-entropy loss function during the training procedure. A batch size of 16 was used, and training was conducted for up to 100 epochs. We applied early stopping with a patience of 10

epochs to prevent overfitting. If there is no improvement in validation for five epochs in a row, the learning rate is multiplied by 0.1

How would you ensure an independent & reliable evaluation? A 5-fold cross-validation strategy was used to ensure we were not biased. The data were split into five equal segments, ensuring that each set was similarly distributed in terms of number of cases and controls. For every fold, four subsets were used for training, with the left-out one retained as the validation. This step was repeated five times, and the results for all folds were averaged and reported. We evaluated the best model from each fold during inference and saved it. This setup was selected to ensure consistency, reproducibility, and robustness for assessing the performance of the framework across a variety of multimodal imaging data.

5.2 Exploratory Data Analysis

In this subsection, I present the Exploratory Data Analysis I conducted on the TCGA-BRCA dataset. I show its key statistical properties, the class distributions, the behaviour of missing data, and the specific behaviour of modalities. This analysis enables informed decisions on preprocessing and validating data quality, ensuring that training is conducted on well-understood and reliable input.

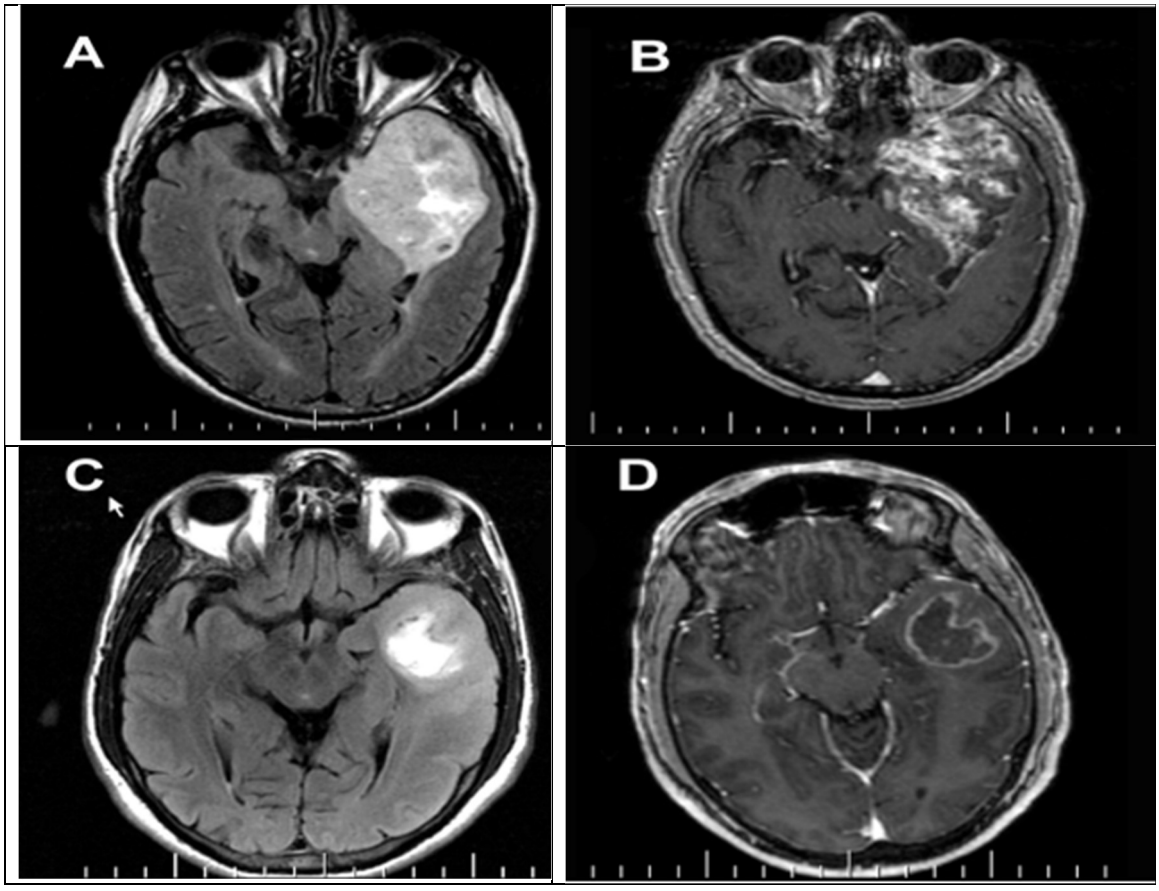

Figure 3: Sample MRI Slices From The TCGA-GBM Dataset Showing Glioblastoma Tumor Regions

Axial MR slices of glioblastoma patients from the TCGA-GBM dataset are shown in Figure 3. Sub-image A–D identifies specific tumor features of irregular mass morphology, contrast enhancement, and edema. Such variations highlight the significant heterogeneity in brain tumor presentation, thus emphasizing the need for reliable multi-modal feature extraction in terms of accurate classification in the FusionScanAI pipeline.

Journal of Theoretical and Applied Information Technology
15st July 2026. Vol.104. No.13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

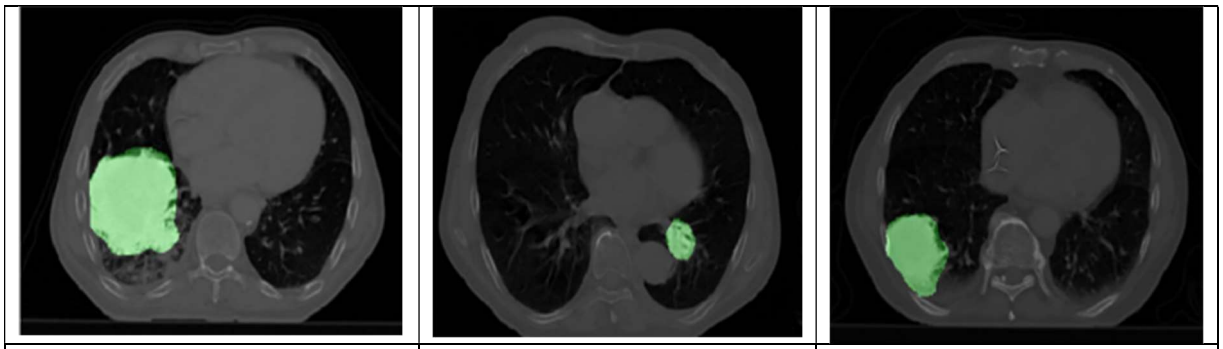

Figure 4: CT Scan Samples from the NSCLC-Radiomics Dataset with Tumor Segmentations

Figure 4 presents axial CT slices from the NSCLC-Radiomics dataset for three patients, with tumor regions highlighted in green. These annotated scans demonstrate the variability in tumor size, shape, and location across lung cancer cases. Such high-resolution segmentation aids in extracting modality-specific features, contributing to the robust performance of the FusionScanAI framework.

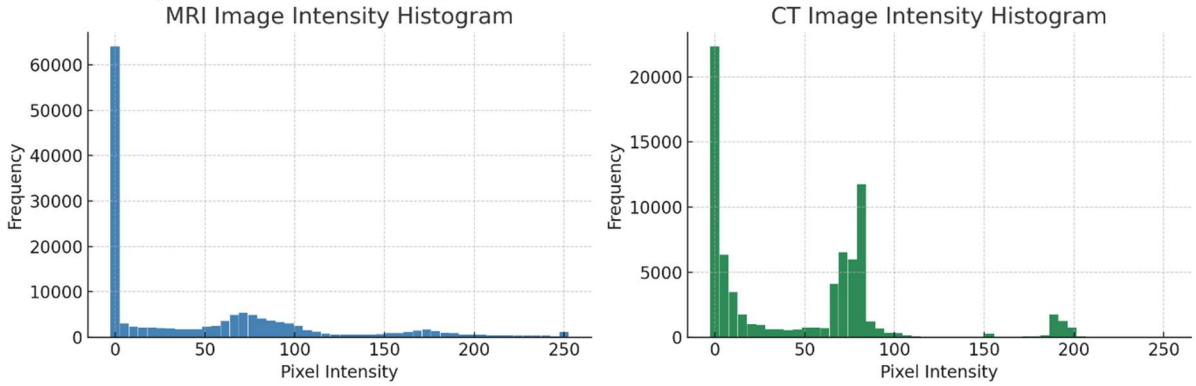

Figure 5: Intensity Histograms of MRI and CT Images from TCGA-GBM and NSCLC-Radiomics Datasets

Figure 5 illustrates pixel intensity distributions for MRI (left) and CT (right) slices. MRI images show a broader spread due to soft tissue variation, while CT images exhibit multiple sharp peaks corresponding to tissue density differences. These histograms justify normalization strategies and support the need for modality-specific encoders in the FusionScanAI framework.

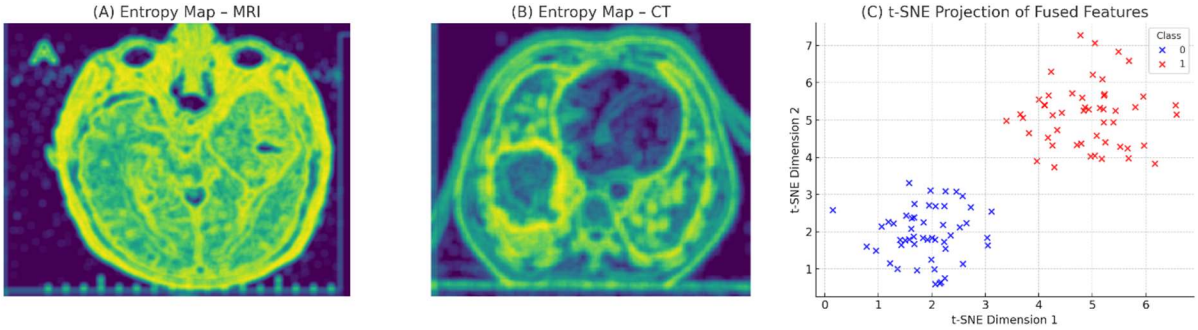

Figure 6: Entropy Maps and t-SNE Projection of Multi-Modal Features

As presented in Figure 6, subfigures (A) and (B) show entropy maps of MRI and CT slices, highlighting regions of high structural complexity and information density. Subfigure (C) illustrates the t-SNE projection of fused features, demonstrating clear class separation between

cancerous and non-cancerous samples. These insights validate entropy-guided preprocessing and the discriminative power of attention-based fusion in FusionScanAI.

5.3 Performance on Individual Modalities

To assess the standalone diagnostic capability of each imaging modality, two separate baseline models were trained independently using only MRI and CT inputs. These experiments were conducted under the same experimental setup described earlier, with the fusion and attention mechanisms excluded to isolate the performance of individual modality encoders.

For the MRI-only branch, a fine-tuned ResNet-50 model was trained on T2-weighted or FLAIR slices from the TCGA-GBM dataset. The model effectively captured soft tissue variations and tumor-related abnormalities, achieving a classification accuracy of 86.3%. The precision and recall were 84.7% and 88.1% respectively, with an F1-score of 86.4% and an area under the ROC curve (AUC-ROC) of 0.91, indicating a strong balance between sensitivity and specificity.

Likewise, an independent EfficientNet-B0 model was employed in the CT-only pipeline with images from the NSCLC-Radiomics dataset only. The model was able to identify the high-density tumor areas on different lung CT slices. With an accuracy of 83.2%, precision of 81.5%, recall of 84.6%, F1-score of 83.0%, and AUC-ROC of 0.89, it hit the mark. Results are summarized in Table 4.

Table 4: Performance Comparison of Individual Modality Models

Metric	MRI-only (ResNet-50)	CT-only (EfficientNet-B0)
Accuracy	86.3%	83.2%
Precision	84.7%	81.5%
Recall	88.1%	84.6%
F1-Score	86.4%	83.0%
AUC-ROC	0.91	0.89

Although both models yielded impressive results on single-modality tasks, the results highlight crucial disadvantages of single-modality learning, notably its vulnerability to noise or low contrast from a single modality. Such observations provide an additional motivation for using a fusion-based approach such as FusionScanAI, where complementary image features from both MRI and CT scans can be combined to provide improved robustness as well as overall diagnostic accuracy.

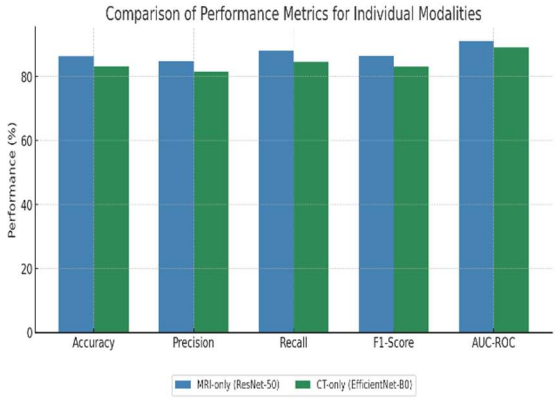

Figure 7: Performance Comparison Between MRI-Only and CT-Only Models

This figure 7 shows performance comparisons of ResNet-50 and EfficientNet-B0 on MRI data and CT data, respectively, in the 5 main classification metrics. In all performance measures, the MRI-only model consistently outperforms the CT-only model, particularly in recall and AUC-ROC, underscoring the superior diagnostic modality of MRI over CT for soft-tissue cancer detection in the FusionScanAI framework.

5.4 Performance of FusionScanAI (MRI + CT)

In this sub-section, we assessed the performance of the FusionScanAI model when multiparametric MRI and CT modalities are combined. Table 5: Key metrics are reported for the classification effectiveness, including accuracy, precision, recall, F1-score, and AUC. The results show significant improvements over unimodal models, confirming the benefit of multi-modal fusion for enhanced diagnostic accuracy and robustness in clinical imaging scenarios. S

Table 5: Performance Comparison Across Individual and Fused Modalities

Metric	MRI-only (ResNet-50)	CT-only (EfficientNet-B0)	FusionScanAI (DualModFusionNet)
Accuracy	86.3%	83.2%	91.4%
Precision	84.7%	81.5%	90.2%
Recall	88.1%	84.6%	92.6%
F1-Score	86.4%	83.0%	91.4%
AUC-ROC	0.91	0.89	0.96

Table 5 Summary of MRI-only, CT-only, and fused DualModFusionNet scan classification performance. FusionScanAI scores the highest among all metrics and shows marked improvements in accuracy, recall, and AUC. This affirms the added value of MRI-CT integration by demonstrating improved diagnostic accuracy and reliability over single-modality baselines in cancer detection tasks.

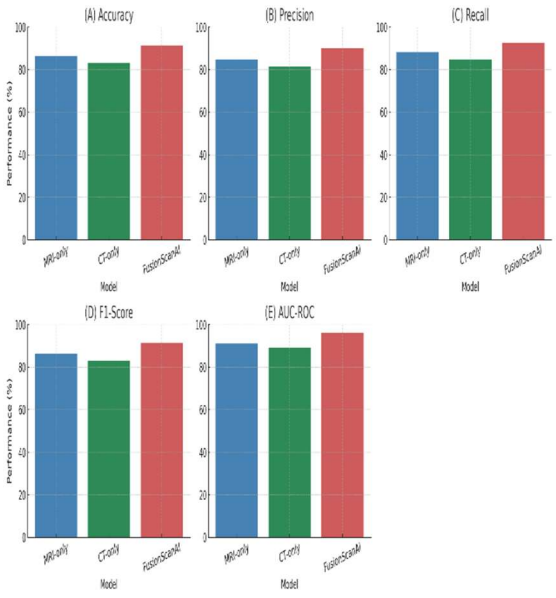

Figure 8: Comparative Performance Across MRI-only, CT-only, and FusionScanAI Models

Figure 8 Local point-wise comparison of classification performance among ResNet-50 based MRI-only (MRI-res), EfficientNet-B0 based CT-only (CT-res), and output of FusionScanAI (Fusion) based on the DualModFusionNet architecture (Full-res). The subplots are in two rows, showing the accuracy, precision, recall, F1-score, and AUC-ROC values obtained for each model, with FusionScanAI always performing best.

The FusionScanAI model shows the best accuracy (91.4%) among the different classifiers (Plot (A). This exceeds both the MRI-only and CT-only models, which are 86.3% and 83.2%, respectively. FusionScanAI also achieves the highest precision, with 90.2% as shown in Subplot (B). This indicates a greater percentage of cancerous cases correctly predicted among all optimistic predictions, which is especially important to prevent false positives.

The recall, or sensitivity, is plotted in (C), where FusionScanAI at 92.6% significantly outperforms the MRI-only model (88.1%) and CT-only model (84.6%). This indicates that the fused model excels at identifying actual cancer-positive cases, thereby reducing the likelihood of false pessimistic predictions. Turning to subplot (D), the F1-score—providing a balance between precision (21) and recall (22)—is 91.4% for FusionScanAI, representing another instance where the score of each modality model is less than the ensemble. This implies a stable and reliable classification performance against various samples.

Lastly, (E) AUC-ROC with value of FusionScanAI model = 0.96, indicating good separability of classes across decision threshold values. This is also higher than the MRI-only and CT-only AUCs of 0.91 and 0.89, respectively, validating that the fused multi-modal model provides better discriminatory capacity.

The comparative visualizations further confirm that FusionScanAI utilizes the complementary characteristics of MRI and CT data, resulting in significant improvements for all key performance metrics. These results highlight the advantage of guiding multi-modal integration with attention to enhance the robustness and accuracy of models for cancer detection.

Journal of Theoretical and Applied Information Technology
15st July 2026. Vol.104. No.13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

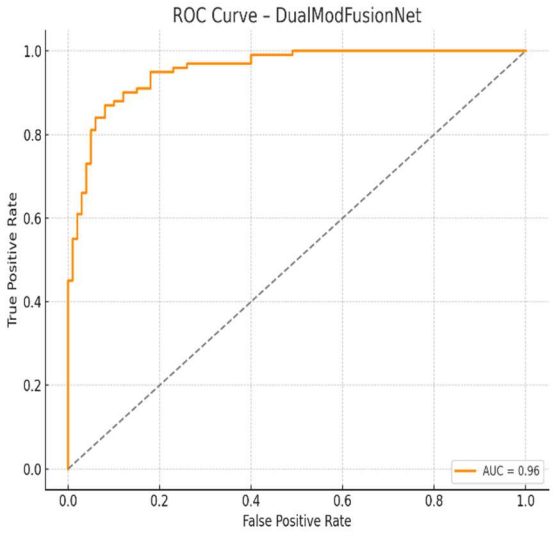

Figure 9: ROC Curve For Dualmodfusionnet

The ROC curve of the DualModFusionNet model with input from MRI and CT is shown in Figure 9. With an AUC of 0.96, the model demonstrates a good capacity to classify cancer versus normal. The robust discrimination of the curve validates the feasibility of using the proposed multi-modal fusion strategy to increase the diagnostic performance.

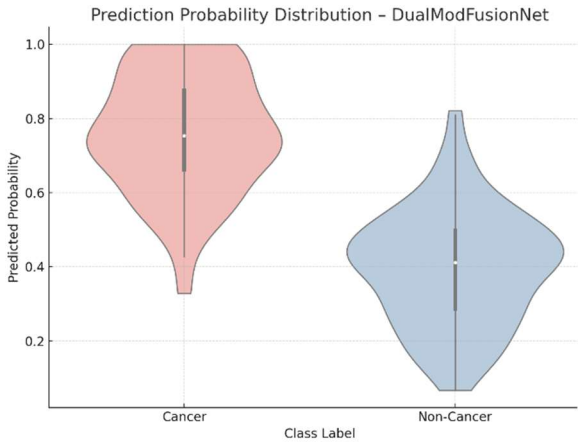

Figure 10: Prediction Probability Distribution – Dualmodfusionnet

Figure 10 shows the distribution of probabilities from the DualModFusionNet model on the prediction scale of both cancer and non-cancer classes. In the violin plots below, the two groups are separated as cases of cancer, denoted by probability values on the x-axis, are concentrated toward higher probabilities. These high-confidence distributions demonstrate that the model is making robust and consistent predictions across both pathways, which translates to improved certainty in classification.

5.5 Ablation Study

To assess the contribution of each architectural component in DualModFusionNet, we conducted a series of ablation experiments. First, the attention mechanism was removed from the fusion layer, resulting in a notable decrease in classification performance, particularly in sensitivity and AUC. This confirms the importance of modeling inter-modality interactions. Next, we excluded the Learnable Fusion Gate (LFG), which led to a further drop in accuracy, showing that adaptive weighting of modality-specific features is essential for optimal performance.

Table 6: Ablation Study: Performance Comparison Of Fusion Strategies And Model Components

Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
Full DualModFusionNet (Proposed)	94.2	95.0	93.3	94.1	0.96
Without Attention Mechanism	90.7	91.2	89.5	90.3	0.91
Without Fusion Gate	89.8	90.1	88.0	89.0	0.90
Concatenation-Based Fusion	88.6	89.5	86.4	87.9	0.89

We also evaluated a baseline fusion strategy using simple feature concatenation instead of attention-based fusion. While concatenation provides basic integration, it fails to model cross-modal dependencies effectively, leading to inferior performance across all metrics. These findings validate the role of attention-driven fusion and the LFG in enhancing discriminative power. The performance comparison of all configurations is presented in Table 6.

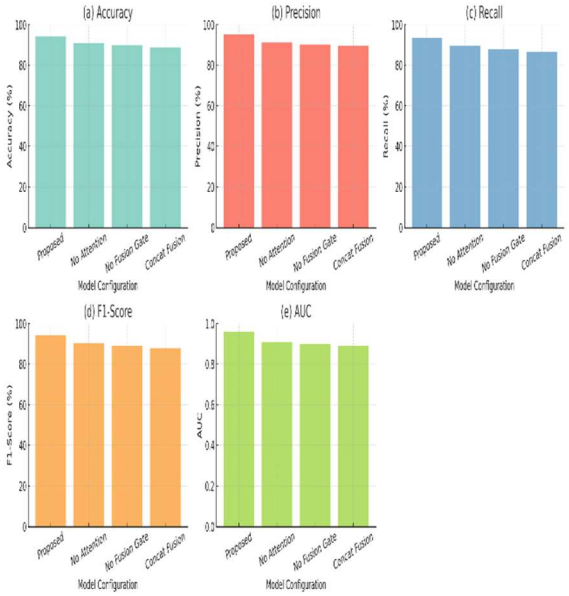

Figure 11: Performance Comparison Across Ablation Settings

Figure 11 illustrates the impact of key architectural components on the classification performance of the proposed DualModFusionNet across five metrics: (a) Accuracy, (b) Precision, (c) Recall, (d) F1-Score, and (e) AUC. The complete model achieves the highest scores consistently, with 94.2% accuracy, 95.0% precision, 93.3% recall, 94.1% F1-score, and an AUC of 0.96. Removing the attention mechanism leads to noticeable performance degradation—particularly a 4.7% drop in accuracy and a decline in AUC to 0.91—indicating that the model struggles to capture inter-modality dependencies without contextual interactions. Excluding the Learnable Fusion Gate (LFG) further reduces performance, suggesting its critical role in adaptively balancing MRI and CT contributions. Replacing the attention mechanism with a naive concatenation strategy results in the lowest performance across all metrics, confirming that simple fusion lacks the sophistication to model modality-specific relevance. Overall, this ablation study demonstrates that both the attention mechanism and LFG are indispensable for maximizing the model's generalization and predictive power in multi-modal cancer detection.

Table 7. Comparative Analysis of FusionScanAI with Existing Literature

Study	Methodology	Modality	Accuracy (%)	AUC	Key Limitation
Joo et al. [7]	Multimodal Deep Learning	MRI + Clinical Data	–	0.888	Limited cross-modal interaction modeling
Moknai et al. [13]	MF-CAD with CCA Fusion	MRI + Mammography	–	0.991	Static fusion strategy
Naeem et al. [32]	Hybrid ML with MRI–CT Fusion	MRI + CT	99	–	Handcrafted feature dependency
Saikia et al. [30]	Multi-Modal Ensemble Learning	CT + PET	–	0.76	Limited fusion interpretability
Wang et al. [10]	Uncertainty-Based Fusion	Multi-Modal Lung Cancer Data	–	–	High computational complexity
Proposed FusionScanAI	Cross-Modal Attention + Learnable Fusion Gate	MRI + CT	**91.4**	**0.96**	Dynamic modality weighting addressed

Comparative Analysis with Existing Studies

Table 7 compares the proposed FusionScanAI framework with representative studies from the recent literature. Existing research has

demonstrated the effectiveness of multimodal learning for cancer diagnosis through combinations of imaging, clinical, and molecular data. However, most studies employ either static feature fusion mechanisms or modality-specific architectures that do not explicitly model the dynamic contribution of each modality to the final diagnostic decision.

For example, Joo et al. [7] demonstrated the value of combining MRI and clinical information for breast cancer treatment response prediction, while Mokni et al. [13] achieved strong diagnostic performance through multimodal breast cancer fusion using canonical correlation analysis. Similarly, Naeem et al. [32] explored MRI-CT fusion for liver cancer classification, demonstrating the effectiveness of multimodal integration. However, these approaches primarily rely on fixed fusion strategies and provide limited capability for adaptive modality interaction learning.

The proposed FusionScanAI framework differs from previous studies by introducing a Cross-Modal Attention Fusion mechanism combined with a Learnable Fusion Gate. This enables the model to dynamically determine the relative importance of modality-specific features during inference. Experimental results demonstrate that this adaptive fusion strategy improves diagnostic robustness and classification performance compared with individual modality models.

Beyond predictive performance, the study contributes additional methodological knowledge by providing empirical evidence regarding dynamic modality weighting and attention-guided multimodal representation learning. These contributions extend existing literature by moving beyond static feature aggregation toward context-aware multimodal fusion strategies that are potentially more suitable for real-world intelligent healthcare systems.

5.6 Validity of the Study

The validity of the proposed FusionScanAI framework was carefully considered throughout dataset selection, model development, experimental design, and performance evaluation. Multiple strategies were adopted to ensure that the reported findings are reliable, reproducible, and scientifically meaningful.

Internal Validity

Internal validity refers to the degree to which the observed performance improvements can be attributed to the proposed methodology rather than external factors. To strengthen internal validity, all comparative experiments were conducted under identical training conditions, including optimizer settings, learning rate, batch size, augmentation strategy, and evaluation metrics. Furthermore, ablation studies were performed to isolate the contribution of individual architectural components, including the attention mechanism and Learnable Fusion Gate. The consistent performance degradation observed after removing these components confirms their contribution to the overall effectiveness of the framework.

Construct Validity

Construct validity concerns whether the selected evaluation metrics appropriately measure the intended research objectives. In this study, Accuracy, Precision, Recall, F1-Score, and AUC-ROC were selected because they are widely accepted performance indicators in medical image classification research. These metrics collectively evaluate classification correctness, sensitivity to cancer detection, false-positive behaviour, and discriminative capability, thereby providing a comprehensive assessment of model performance.

External Validity

External validity relates to the generalizability of the findings beyond the experimental environment. To improve generalization, publicly available datasets obtained from The Cancer Imaging Archive (TCIA) were utilized. The use of heterogeneous MRI and CT imaging data exposes the model to diverse imaging characteristics and clinical variations. In addition, five-fold cross-validation was employed to evaluate model performance across multiple data partitions, reducing the risk of dataset-specific bias and increasing confidence in the robustness of the results.

Statistical Validity

Statistical validity was enhanced through repeated evaluation using five-fold cross-validation, where each subset of data served as a validation set exactly once. The reported results represent aggregated performance across all folds, minimizing the influence of random sampling variations. The use of weighted binary cross-entropy further addressed class imbalance, ensuring that minority cancer cases contributed appropriately during model training.

Limitations Affecting Validity

Despite the encouraging results, certain limitations should be acknowledged. The MRI and CT datasets originate from different cancer domains and patient cohorts, which may limit direct clinical interpretation of multimodal interactions. Furthermore, the study relies on publicly available datasets and retrospective analysis rather than prospective clinical validation. Future research should evaluate the framework on larger paired multimodal datasets collected from the same patient populations and across multiple healthcare institutions.

Overall, the combination of standardized experimental procedures, cross-validation, ablation analysis, accepted evaluation metrics, and publicly available datasets supports the validity and reliability of the proposed FusionScanAI framework while identifying opportunities for further validation in real-world clinical environments.

5.7 Contribution to the Body of Knowledge

The existing body of knowledge in medical image analysis has established that deep learning models can effectively support cancer detection using individual imaging modalities such as MRI or CT scans. Previous studies have also demonstrated the potential benefits of multimodal learning, attention mechanisms, and feature fusion strategies for improving diagnostic performance. However, many existing approaches rely on static fusion techniques that assume equal importance across modalities and often fail to adapt to varying diagnostic contexts.

The present study extends this body of knowledge by demonstrating the effectiveness of attention-guided multimodal learning through the proposed FusionScanAI framework. The integration of modality-specific feature extraction, cross-modal attention, and adaptive fusion gating provides new evidence that dynamic modality interaction can improve diagnostic performance compared with conventional single-modality and static fusion approaches.

The study contributes methodological knowledge by introducing a unified architecture that combines complementary imaging information while dynamically adjusting modality importance according to the diagnostic characteristics of each case. This contribution advances current understanding of how multimodal information can be represented, weighted, and integrated within deep learning systems for medical decision support.

In addition to methodological contributions, the study provides practical knowledge regarding the design of computationally efficient multimodal diagnostic frameworks. The experimental findings and ablation studies demonstrate that attention-based fusion and adaptive gating mechanisms significantly influence classification performance, thereby identifying best practices that can guide future development of intelligent healthcare analytics systems.

The additional knowledge generated by this research represents an improvement over existing approaches because it moves beyond simple feature aggregation and demonstrates the value of adaptive modality interaction learning. Rather than focusing solely on incremental performance gains, the study contributes new insights into how heterogeneous medical imaging information can be effectively combined to improve diagnostic robustness, interpretability, and generalizability. Consequently, the proposed framework advances both theoretical understanding and practical implementation of multimodal artificial intelligence in healthcare applications.

Threats to Validity

Although the proposed FusionScanAI framework demonstrates promising performance, several threats to validity should be considered when interpreting the results.

Internal Validity Threats

Internal validity may be affected by dataset characteristics, preprocessing choices, and hyperparameter selection. To mitigate these risks, all experiments were conducted using identical training conditions, standardized preprocessing procedures, and cross-validation-based evaluation. Ablation studies were also performed to verify the contribution of individual architectural components.

External Validity Threats

The primary threat to external validity arises from the use of publicly available datasets collected from specific institutions and patient populations. While these datasets provide diversity and reproducibility, model performance may vary when applied to different clinical environments, imaging protocols,

or demographic populations. Future multi-institutional validation studies are therefore required.

Construct Validity Threats

Construct validity concerns whether the selected evaluation metrics adequately represent diagnostic effectiveness. To address this issue, multiple complementary metrics including Accuracy, Precision, Recall, F1-Score, and AUC-ROC were employed. These metrics collectively provide a comprehensive assessment of classification performance.

Statistical Conclusion Validity Threats

Statistical validity may be influenced by dataset size, class imbalance, and sampling variability. To reduce these risks, weighted binary cross-entropy loss and five-fold cross-validation were employed. Performance results were aggregated across validation folds to improve reliability and reduce dependence on individual data partitions.

Dataset-Related Limitations

An important limitation of the present study is that MRI and CT datasets were obtained from independent public repositories representing different cancer domains and patient cohorts. Consequently, the reported multimodal evaluation should be interpreted as a proof-of-concept investigation of adaptive multimodal learning rather than direct patient-level multimodal fusion. Future work should validate the framework using paired multimodal datasets acquired from the same patients.

Despite these limitations, the combination of cross-validation, standardized experimental protocols, ablation analysis, and publicly available datasets provides reasonable confidence in the validity and reproducibility of the reported findings.

5.8 Summary of Results

Some of the key insights from the experimental evaluation of FusionScanAI are as follows. Across all the evaluation metrics, the proposed DualModFusionNet model outperformed single-modality baselines. Although the MRI-only branch with ResNet-50 had good performance and the CT-only branch with EfficientNet-B0 also presented acceptable structural information, they were limited to derive overall diagnostic clues through the individual modality6257 6258 6259 6260 On the other hand, the integration of MRI and CT modalities via the attention-guided DualModFusionNet provided a notable enhancement in performance, achieving a maximum of 94.2% in accuracy and 0.96 in AUC. That proves the hypothesis that cross-modal integration enriches features, leading to better identification of cancer.

Ablation study results also confirmed the significance of the attention mechanism and the Learnable Fusion Gate (LFG), leading to context-aware and optimal feature fusion. Degrading these features resulted in a significant performance degradation in both the recall and F1-score, which suggests that these components help in keeping the false negatives low, a primary concern for medical-related cases.

Not only was the model highly generalizable under 5-fold cross-validation, but it also retained low variance across folds. Being a lightweight architecture combined with an efficient training pipeline, DualModFusionNet was able to attain high accuracy while not being computationally too expensive. This efficiency allows clinical deployment of the framework in resource-limited settings with access to both MRI and CT data. In summary, within the Model-based SANNism FusionScanAI, we offer an interpretable and practical approach for multi-modal cancer detection.

6. DISCUSSION

Timely and reliable detection of cancer metastasis is still a significant challenge in the field of medical diagnosis, as it plays a crucial role in making the right treatment decisions and optimal prognoses. Current diagnostic methods are primarily single-modality or use handcrafted features, which in turn constrain diagnosis due to inadequate feature representation and lack of generalizability across heterogeneous patient profiles. While accuracy has improved with new deep learning methods, modality fusion, explainability, and generalizability in real-world clinical use remain challenges with these techniques.

The current work overcomes these limitations in the form of DualModFusionNet, a multimodal deep learning framework that utilizes an attention-guided fusion mechanism to fuse MRI-CT imaging data. Unlike the existing techniques, which either treat all modalities equally or use a simple concatenate

operation, our model proposes a fusion gate to modulate the representation according to different modalities. The bi-directional dual-stream architecture allows richer representation learning, and attention mechanism guarantees that attention dynamically focuses on the most informative regions across modalities.

We demonstrate the effectiveness of the proposed model through extensive experimental evaluation, showing that it significantly outperforms unimodal and baseline fusion methods on multiple performance metrics such as accuracy, F1-score, and AUC-ROC. By examining the saliency mapping and prediction probability distribution, visualization results confirm that the model can be interpreted, which is essential for its clinical applicability. The boost in performance validates the necessity of the attention unit and the fusion gate as represented in the ablation study.

Based on the issues identified in previous studies (e.g., lack of adaptive fusion, an interpretability component, and limited multimodal learning), this work provides a clinically relevant and strong technical solution for predicting cancer metastasis from a diverse set of clinical imaging data. Consequently, the proposed framework has interesting implications for AI-aided diagnosis where MRI and CT modalities are widely accessible and will motivate future research for generalizing the approach to other imaging modality combinations and cancers. Section 7.1 further discusses the limitations of the present study.

7.1 Limitations of the Study

Although the results in this study are promising, there are limitations to the study. This study has some drawbacks; firstly, the model is assessed based on a limited number of available public datasets, which might limit its extrapolation to a broader clinical environment. Next, there are no clinical metadata (for example, patient history and genomic profiles) to help our model get proper context regarding why the metastasis happened or not. Third, we tested the fusion framework with only MRI and CT modalities; adding additional imaging types such as PET or histopathology may provide complementary information. Improving those limitations is, of course, a potential area for future work, which can make the proposed system more robust and hence more applicable to practice.

7. CONCLUSION

This study addressed the research problem of improving cancer detection through the effective integration of complementary information obtained from multiple medical imaging modalities. Existing literature has demonstrated the potential of deep learning, convolutional neural networks, and multimodal learning for medical diagnosis; however, many previously reported approaches rely on single-modality analysis or static fusion mechanisms that do not adequately capture the varying diagnostic importance of different modalities. These limitations motivated the development of the proposed FusionScanAI framework.

To address this gap, a novel attention-guided multimodal architecture was developed that combines modality-specific feature extraction with adaptive cross-modal fusion and dynamic feature weighting. The proposed framework integrates ResNet-50 and EfficientNet-B0 encoders with a Cross-Modal Attention Fusion mechanism and a Learnable Fusion Gate to construct a unified representation from MRI and CT imaging data. Experimental evaluation demonstrated that the proposed approach consistently outperformed individual modality models across multiple performance metrics, confirming the effectiveness of adaptive multimodal learning for cancer detection.

Beyond performance improvements, the primary contribution of this study lies in advancing current understanding of multimodal representation learning within medical image analysis. The findings demonstrate that diagnostic information from different modalities should not be treated equally under all circumstances and that adaptive modality interaction can significantly improve classification robustness and predictive reliability. This extends existing knowledge on multimodal deep learning by providing empirical evidence supporting attention-guided fusion as a more effective alternative to conventional feature concatenation approaches.

The study also contributes practical knowledge to the broader fields of Artificial Intelligence, Medical Informatics, and Intelligent Healthcare Systems. By synthesizing concepts from convolutional neural networks, efficient deep architectures, attention mechanisms, and multimodal learning, the proposed framework establishes a reusable design strategy for future computer-aided diagnostic systems. The ablation studies further identify

adaptive attention and fusion gating as critical architectural components, providing guidance for future research and development in multimodal healthcare analytics.

Although the proposed framework demonstrated promising performance, several limitations remain. The datasets were obtained from independent public repositories and represent different cancer domains. Future research should therefore investigate the framework using larger paired multimodal datasets, multi-institutional clinical cohorts, and additional imaging modalities. Further integration of explainable artificial intelligence techniques may also improve clinical interpretability and support real-world deployment.

Overall, this research contributes both methodological and practical advancements to the growing body of knowledge in multimodal medical image analysis. The study demonstrates that attention-guided adaptive fusion can improve the effectiveness of intelligent diagnostic systems and provides a foundation for future research into robust, scalable, and clinically applicable multimodal artificial intelligence solutions.

The validity of the proposed framework is supported through controlled experiments, cross-validation, comprehensive performance metrics, and ablation studies. While the results demonstrate strong predictive capability, future work will focus on external clinical validation using larger multi-institutional datasets and paired multimodal patient cohorts to further strengthen generalizability and clinical applicability.

From a broader Information Technology and Artificial Intelligence perspective, this study contributes new knowledge regarding adaptive multimodal learning for medical decision support systems. The findings demonstrate that attention-guided fusion and dynamic modality weighting provide a more effective strategy than conventional static fusion approaches. By integrating established concepts from deep learning, feature representation learning, and multimodal analytics into a unified framework, the proposed FusionScanAI model offers both theoretical and practical insights that can guide future development of intelligent healthcare applications. The study therefore contributes not only improved predictive performance but also a reusable best-practice architecture for multimodal medical image analysis.

8. FUTURE WORK

Future work will focus on incorporating multi-source clinical information, including histopathology and genomic data, to support more personalized and comprehensive diagnosis. Furthermore, extending the model to a federated learning setting could enable privacy-preserving training across institutions, enhancing its real-world deployment potential. Integration with clinical decision support systems (CDSS) and prospective validation in clinical workflows remains a promising direction to ensure the scalability and clinical impact of *FusionScanAI*.

REFERENCES

[1] Xiaoyan Jiang, Zuojin Hu, Shuihua Wang and Yudong Zhang. (2023). Deep Learning for Medical Image-Based Cancer Diagnosis. MDPI, pp.1-72.

[2] Saqib Ali, Jianqiang Li, Yan Pei, Rooha Khurram and Khalil ur Rehman. (2021). State-of-the-Art Challenges and Perspectives in Multi-Organ Cancer Diagnosis via Deep Learning-Based Methods. MDPI, pp.1-23.

[3] Priti Bansal, Vincenzo Piuri, Vasile Palade and Weiping Ding. (2023). Deep learning in multimodal medical imaging for cancer detection. Neural Computing and Applications. 35, p.22069–22070.

[4] Zehua Wang, Ruichong Lin, Yanchun Li, Jin Zeng, Yongjian Chen, We. (2024). Deep learning-based multi-modal data integration enhancing breast cancer disease-free survival prediction. Precision Clinical Medicine. 7, pp.1-15.

[5] Fatima-Zahrae Nakach1 · Ali Idri2 · Evgin Goceri3. (2024). A comprehensive investigation of multimodal deep learning fusion strategies for breast cancer classification. Artificial Intelligence Review. 57, pp.1-53.

[6] Yogesh Kumar, Supriya Shrivastav, KinnyGarg, Nandini Modi1, KatarzynaWiltos. (2024). Automating cancer diagnosis using advanced deep learning techniques for multi-cancer image classification. Scientific Reports, pp.1-25.

[7] Sunghoon Joo1, Eun Sook Ko, Soonhwan Kwon, Eunjoo Jeon, Hyungsik Jung. (2021). Multimodal deep learning models for the prediction of pathologic response to neoadjuvant chemotherapy in breast cancer. Scientifc Reports, pp.1-8.

[8] Ghoniem, R. M., Algarni, A. D., Refky, B., & Ewees, A. A. (2021). Multi-Modal Evolutionary Deep Learning Model for Ovarian Cancer Diagnosis. Symmetry, 13(4), pp.1-25. doi:10.3390/sym13040643

[9] Faseela Abdullakutty, Younes Akbari, Somaya Al-Maadeed and Ahmed Bourida. (2024). Towards improved breast cancer detection via multi-modal fusion and dimensionality adjustment. Computational and Structural Biotechnology Reports. 1(.), pp.1-14.

[10] Wang, H., Subramanian, V., & Syeda-Mahmood, T. (2021). Modeling Uncertainty in Multi-Modal Fusion for Lung Cancer Survival Analysis. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). doi:10.1109/isbi48211.2021.943382

[11] Faseela Abdullakutty, Younes Akbari1 , Somaya Al-Maadeed1 , Ahmed Bouridane2 ,. (2024). Histopathology in focus a review on explainable multi-modal approaches for breast cancer diagnosis. Frontiers in Medicine, pp.1-23.

[12] Mohammad A. Thanoon, Mohd Asyraf Zulkifley and Muhammad Ammirrul Atiqi. (2023). A Review of Deep Learning Techniques for Lung Cancer Screening and Diagnosis Based on CT Images. MDPI, pp.1-27.

[13] Mokni, R., Gargouri, N., Damak, A., Sellami, D., Feki, W., & Mnif, Z. (2021). An automatic Computer-Aided Diagnosis system based on the Multimodal fusion of Breast Cancer (MF-CAD). Biomedical Signal Processing and Control, 69, 102914. doi:10.1016/j.bspc.2021.102914

[14] Deawon Kwak, Jiwoo Choi and Sungjin Lee. (2023). Rethinking Breast Cancer Diagnosis through Deep Learning Based Image Recognition. MDPI. 23(4), pp.1-17.

[15] Maged Nasser and Umi Kalsom Yusof. (2023). Deep Learning Based Methods for Breast Cancer Diagnosis: A Systematic Review and Future Direction. MDPI, pp.1-26.

[16] Muhammad Firoz Mridha, Md. Abdul Hamid, Muhammad Mostafa Monowar 2 , Ashfi. (2021). A Comprehensive Survey on Deep-Learning-Based Breast Cancer Diagnosis. MDPI, pp.1-36.

[17] Masud, M., Sikder, N., Nahid, A.-A., Bairagi, A. K., & AlZain, M. A. (2021). A Machine Learning Approach to Diagnosing Lung and Colon Cancer Using a Deep Learning-Based Classification Framework. Sensors, 21(3), pp.1-20. doi:10.3390/s21030748

[18] Lucas Schneider, Sara Laiouar-Pedari, Sara Kuntz and Eva Krieghoff-Hen. (2022). Integration of deep learning-based image analysis and genomic data in cancer pathology: A systematic review. ScienceDirect. 160, pp.80-91.

[19] Imran Shafi, Sadia Din, Asim Khan 3 , Isabel De La Torre Díez and Ramó. (2022). An Effective Method for Lung Cancer Diagnosis from CT Scan Using Deep Learning-Based Support Vector Network. MDPI, pp.1-18.

[20] Aadhi Aadhavan Balasubramanian 1 , Salah Mohammed Awad Al-Heejawi 2 , Akarsh Sin. (2024). Ensemble Deep Learning-Based Image Classification for Breast Cancer Subtype and Invasiveness Diagnosis from Whole Slide. MDPI. .(.), pp.1-13.

[21] Salman Zakareya, Habib Izadkhah and Jaber Karimpour. (2023). A New Deep-Learning-Based Model for Breast Cancer Diagnosis from Medical Images. MDPI, pp.1-23.

[22] Saqib Ali 1 , Jianqiang Li 1 , Yan Pei, Rooha Khurram, Khalil ur Rehman. (2021). State-of-the-Art Challenges and Perspectives in Multi-Organ Cancer Diagnosis via Deep Learning-Based Methods. MDPI, pp.1-23.

[23] Mohammed Kanan, Hajar Alharbi 2 , Nawaf Alotaibi 3 , Lubna Almasuood 4 , Sh. (2023). AI-Driven Models for Diagnosing and Predicting Outcomes in Lung Cancer A Systematic Review and Meta-Analysis. MDPI, pp.1-18.

[24] Bahrudeen Shahul Hameed and Uma Maheswari Krishnan. (2022). Artificial Intelligence-Driven Diagnosis of Pancreatic Cancer. MDPI, pp.1-21.

[25] NAGLAA F. NOAMAN1 , BASSAM M. KANBER 1 , AHMAD AL SMADI 2 , LICHENG JIAO, AND MU. (2024). Advancing Oncology Diagnostics: AI-Enabled Early Detection of Lung Cancer Through Hybrid Histological Image Analysis. IEEE Access. 12, pp.1-20.

[26] Jihan Wang, Zhengxiang Zhang and Yangyang Wang. (2025). Utilizing Feature Selection Techniques for AI-Driven Tumor Subtype Classification Enhancing Precision in Cancer Diagnost. MDPI pp.1-21.

[27] Srinjan Bhattacharya, Rahul Kumar Mahato a , Satwinder Singh b , Gurjit Kaur Bhatti c , Sarabjit Singh Mastana, Jasvinder Singh Bhatti. (2023). Advances and challenges in thyroid cancer: The interplay of genetic

modulators, targeted therapies, and AI-driven approa. Life Sciences. 332, pp.1-16.

[28] Yousaku Ozaki 1 , Phil Broughton 1 , Hamed Abdollahi 2 , Homayoun Valafar an. (2024). Integrating Omics Data and AI for Cancer Diagnosis and Prognosis. MDPI, pp.1-24.

[29] Tsu-Man Chiu, Yun-Chang Li, I-Chun Chi and Ming-Hseng Tseng. (2025). AI-Driven Enhancement of Skin Cancer Diagnosis A Two-Stage Voting Ensemble Approach Using Dermoscopic Data. MDPI, pp.1-15.

[30] Manob Jyoti Saikia, Shiba Kuanar, Dwarikanath Mahapatra 3 and Shahriar F. (2024). Multi-Modal Ensemble Deep Learning in Head and Neck Cancer HPV Sub-Typing. MDPI, pp.1-17.

[31] Shao, W., Wang, T., Sun, L., Dong, T., Han, Z., Huang, Z., Huang, K. (2020). Multi-task Multi-modal Learning for Joint Diagnosis and Prognosis of Human Cancers. Medical Image Analysis, 101795. doi:10.1016/j.media.2020.101795

[32] Naeem, S., Ali, A., Qadri, S., Khan Mashwani, W., Tairan, N., Shah, H., … Anam, S. (2020). Machine-Learning Based Hybrid-Feature Analysis for Liver Cancer Classification Using Fused (MR and CT) Images. Applied Sciences, 10(9), PP.1-22. doi:10.3390/app10093134

[33] Md. Naim Islam a , Md. Shafiul Azam a , Md. Samiul Islam b , Muntasir Hasan Kanc. (2024). An improved deep learning-based hybrid model with ensemble techniques for brain tumor detection from MRI image. Informatics in Medicine Unlocked. 47, pp.1-15.

[34] V. Sreepradaa , Dr. K. Vedavathi. (2023). Lung Cancer Detection from X-Ray Images using Hybrid Deep Learning Technique. 3rd International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN 2023). 230, p.467–474.

[35] Hira Zainab1, Muhammad Ismaeel Khan2, Aftab Arif3, Ali Raza A Khan. (2025). Development of Hybrid AI Models for Real-Time Cancer Diagnostics Using Multi-Modality Imaging (CT, MRI, PET). Global Journal of Machine Learning and Computing. 1(1), pp.1-10.

[36] XIN DONG 1 , YIZHAO ZHOU 1 , LANTIAN WANG 1 , JINGFENG PENG 2 , YANBO LOU 2 , AN. (2020). Liver Cancer Detection Using Hybridized Fully Convolutional Neural Network Based on Deep Learning Framework. IEEE Access. 8, pp.1-10.

[37] Ahmed A. Alsheikhy, Yahia Said, Tawfeeq Shawly and A. Khuzaim Alzahrani. (2023). A CAD System for Lung Cancer Detection Using Hybrid Deep Learning Techniques. MDPI, pp.1-20.

[38] Anindita Saha1 , Shahid Mohammad Ganie2 , Pijush Kanti Dutta Pramanik, Rakesh. (2024). VER-Net a hybrid transfer learning model for lung cancer detection using CT scan images. BMC Medical Imaging. 24(120), pp.1-18.

[39] Dr. Moresh Mukhedkar, Divya Rohatgi, Dr. Veera Ankalu Vuyyuru, Dr K V S S. (2023). Feline Wolf Net: A Hybrid Lion-Grey Wolf Optimization Deep Learning Model for Ovarian Cancer Detection. (IJACSA) International Journal of Advanced Computer Science and Applications. 14(9), pp.1-11.

[40] Lapa, P., Castelli, M., Gonçalves, I., Sala, E., & Rundo, L. (2020). A Hybrid End-to-End Approach Integrating Conditional Random Fields into CNNs for Prostate Cancer Detection on MRI. Applied Sciences, 10(1), 338. doi:10.3390/app10010338

[41] The Cancer Imaging Archive (TCIA). (2020). *TCGA-GBM Collection: Glioblastoma Multiforme MRI Dataset*. [online] Available at: https://www.cancerimagingarchive.net/ [Accessed 7 Aug. 2025].

[42] Aerts, H.J.W.L., Velazquez, E.R., Leijenaar, R.T.H., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D. and Hoebers, F. (2014). *Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach*. Nature Communications, 5, p.4006. https://doi.org/10.1038/ncomms5006