International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

MULTIMODAL CANCER DETECTION USING MACHINE LEARNING

Iffat Saleha
P.G Studentr
Department of Computer Science & Engineering,
Anjuman College of Engineering & Technology, Nagpur Maharashtra, India
iffat1897@gmail.coma

Prof. Kamlesh Kelwade
Associate Professor
Department of Computer Science & Engineering,
Anjuman College of Engineering & Technology, Nagpur Maharashtra, India
kamleshk@anjumanengg.edu.in

Abstract - Early and accurate cancer diagnosis remains one of the most critical challenges in modern healthcare. Traditional unimodal diagnostic approaches—such as radiology, histopathology, or genomics—offer valuable but fragmented insights, often resulting in incomplete clinical interpretations. This research proposes an interpretable cross-modal attention-based deep learning framework for multimodal cancer detection, designed to integrate heterogeneous data sources including radiological images, histopathology slides, genomic profiles, and structured clinical records. Each modality is processed through a dedicated deep encoder—3D CNNs for imaging, Vision Transformers (ViTs) for histopathology, transformer-based sequence models for genomics, and MLPs for clinical data—to extract high-level feature representations. These are fused using a cross-modal attention mechanism that dynamically learns inter-modality relationships and gracefully handles missing data. The attention weights further serve as an intrinsic interpretability feature, revealing the contribution of each modality to the final diagnostic decision. Experimental validation on benchmark multimodal datasets demonstrates that the proposed framework achieves superior accuracy, robustness, and generalization compared to unimodal and existing fusion-based models. Moreover, a Clinical Decision Support Dashboard provides transparent visual explanations through saliency maps and modality importance scores, fostering clinician trust and practical usability. The results highlight the potential of interpretable multimodal AI to transform diagnostic precision, reduce uncertainty, and advance personalized cancer care.

Index Terms - Multimodal Deep Learning, Cancer Detection, Cross-Modal Attention, Explainable AI, Medical Imaging, Genomics, Clinical Decision Support.

I. INTRODUCTION

Cancer remains one of the leading causes of mortality worldwide, accounting for nearly ten million deaths annually according to the World Health Organization (WHO). Despite advancements in medical imaging, genomics, and computational pathology, the effectiveness of cancer treatment continues to depend heavily on early and accurate diagnosis. Conventional diagnostic workflows often rely on unimodal data sources such as radiological imaging, histopathology, or genomic sequencing, each offering valuable but isolated insights into the disease. However, this siloed approach limits the overall diagnostic precision by neglecting the complementary relationships that exist among different modalities.

In recent years, deep learning (DL) has revolutionized the field of medical diagnostics, achieving expert-level performance across individual tasks such as tumor classification, lesion segmentation, and mutation prediction. Yet, most existing models operate within a single modality, failing to leverage the full spectrum of clinical information available for a patient. This limitation underscores the growing need for multimodal deep learning, where diverse data types are computationally integrated to form a comprehensive, holistic representation of the patient's disease state.

The primary challenge in multimodal cancer detection lies in the heterogeneity and incompleteness of medical data. Radiology images, genomic sequences, and clinical records differ vastly in structure, dimensionality, and

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

information density. Moreover, real-world datasets are often incomplete, with missing modalities due to cost, accessibility, or clinical constraints. Existing fusion techniques—such as early or late fusion—struggle to address these challenges effectively, often leading to information loss or model brittleness.

To overcome these limitations, this research introduces an interpretable cross-modal attention-based deep learning framework that performs intermediate fusion of heterogeneous data streams. Each modality is processed through a dedicated encoder network—such as 3D CNNs for radiology, Vision Transformers (ViTs) for histopathology, transformer-based models for genomics, and MLPs for clinical data—to extract modality-specific features. These features are then combined through a cross-modal attention fusion module, which dynamically learns the relationships between modalities and assigns adaptive importance weights. This mechanism not only enhances predictive accuracy but also provides intrinsic interpretability, revealing which modalities and features contribute most to each diagnostic decision.

Furthermore, the proposed framework incorporates a Clinical Decision Support Dashboard that visualizes attention-based insights through saliency maps, modality importance distributions, and key genomic indicators. This interpretability layer bridges the gap between algorithmic prediction and clinical reasoning, promoting trust and transparency among medical professionals.

In summary, this study aims to address the major challenges in cancer diagnostics by developing a robust, interpretable, and clinically viable multimodal deep learning framework. The system demonstrates improved accuracy, resilience to missing data, and enhanced generalization across diverse clinical environments, marking a significant step toward the real-world adoption of AI in oncology.

II. LITERATURE REVIEW

Over the past decade, the integration of artificial intelligence (AI) and deep learning (DL) into medical diagnostics has transformed the field of oncology. Early studies demonstrated that deep neural networks, particularly Convolutional Neural Networks (CNNs), could achieve expert-level performance on various unimodal diagnostic tasks, such as tumor classification, lesion segmentation, and histopathological image analysis. However, these unimodal approaches were inherently limited, as they relied on a single data source, often missing critical correlations present across multiple diagnostic modalities.

A. Unimodal Deep Learning in Medical Imaging

Early research in medical AI primarily focused on individual modalities. CNN-based architectures achieved remarkable success in detecting diseases from imaging data, such as diabetic retinopathy, skin cancer, and lung nodule detection on CT scans. Similarly, Vision Transformers (ViTs) and hybrid CNN–Transformer models have improved the accuracy of histopathological image classification by modeling spatial dependencies and feature hierarchies. In the domain of genomics, transformer-based architectures such as Genomic BERT have shown potential for predicting cancer subtypes from DNA or RNA sequences. While these models advanced their respective fields, their inability to integrate heterogeneous data restricted their diagnostic comprehensiveness.

B. Emergence of Multimodal Deep Learning

The concept of multimodal learning emerged to bridge the gap between isolated data sources. By combining information from radiology, histopathology, genomics, and clinical records, multimodal systems aim to build a unified understanding of disease progression. Three primary fusion strategies are described in literature:

1. Early Fusion (Data-Level): Combines raw or minimally processed data from multiple modalities. Though straightforward, it suffers from dimensional inconsistencies and poor handling of missing data.
2. Late Fusion (Decision-Level): Aggregates decisions from separate unimodal models using ensemble techniques such as voting or averaging. This approach enhances robustness but loses inter-modal correlations.
3. Intermediate Fusion (Feature-Level): Merges high-level feature embeddings extracted from each modality at a mid-network layer. This method effectively balances representation power and interpretability, making it the most suitable for clinical applications.

Studies such as Chen et al. (2019) introduced *Pathomic Fusion*, integrating histopathology and genomics through a gated attention mechanism, demonstrating improved survival prediction performance. Li et al. (2025) applied late fusion on mammography and ultrasound for breast cancer detection, achieving 93.78% accuracy. However, these systems were limited by dataset dependency, lack of external validation, and inability to handle missing modalities.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

C. Attention Mechanisms and Transformers in Multimodal Integration

Recent advances in attention mechanisms and transformer architectures have revolutionized multimodal learning. Attention-based models enable networks to focus on the most relevant features across modalities, facilitating dynamic weighting and cross-dependency modeling. Cross-modal attention mechanisms, in particular, allow the model to adaptively emphasize modalities with higher diagnostic relevance for a specific case. This flexibility improves robustness, interpretability, and clinical trust.

Zhang et al. (2025) demonstrated that transformer-based fusion networks could effectively integrate radiology, pathology, and genomic data for cancer prognosis prediction. Similarly, Huang et al. (2024) proposed an attention-driven fusion model that achieved superior accuracy while providing explainable modality importance insights. Despite these advances, many existing frameworks still lack real-world generalization and interpretable decision transparency.

D. Research Gaps

A review of the current state-of-the-art reveals several unresolved challenges in multimodal computational oncology:

- Generalization Gap: Most models are trained on single-institution datasets, limiting adaptability to new environments.
- Robustness Gap: Many frameworks assume complete data availability, failing when modalities are missing.
- Interpretability Gap: Existing models often function as "black boxes," offering limited insight into how predictions are formed.
- Scalability Gap: Most research targets single cancer types, preventing broader clinical applicability.

E. Summary

The existing literature establishes the superiority of multimodal deep learning over unimodal systems but highlights critical limitations in robustness, interpretability, and scalability. To address these gaps, the proposed research introduces a cross-modal attention-based fusion framework designed to integrate diverse clinical modalities, handle incomplete datasets gracefully, and provide transparent, clinically interpretable predictions—paving the way for real-world AI deployment in cancer diagnostics.

III. PROBLEM STATEMENT AND RESEARCH OBJECTIVES

A. Problem Statement

Cancer diagnosis in clinical practice often relies on diverse data sources—radiological imaging, histopathology slides, genomic profiles, and clinical records—each capturing a different aspect of the disease. However, these modalities typically exist in isolated silos, analyzed independently, leading to fragmented diagnostic conclusions. The absence of effective integration mechanisms results in diagnostic uncertainty, delayed treatment, and reduced accuracy in early detection.

Furthermore, medical datasets are inherently heterogeneous and incomplete. Radiology data consists of high-dimensional volumetric images, histopathology involves gigapixel-scale slides, genomics presents categorical sequence data, and clinical records contain structured tabular information. Combining such varied data types poses a major computational and semantic challenge. Most existing multimodal systems assume the availability of all modalities for each patient—an unrealistic expectation in real-world clinical settings—causing severe performance degradation when any modality is missing.

Another critical issue lies in interpretability. Deep learning models, while powerful, often function as "black boxes," offering predictions without transparent reasoning. In oncology, this lack of interpretability limits clinician trust, regulatory approval, and clinical adoption. For cancer diagnostics, a model must not only predict accurately but also provide clinically meaningful explanations that align with medical reasoning.

Hence, the central research problem can be formally stated as follows:

How can a multimodal deep learning framework be designed to effectively integrate heterogeneous medical data (radiology, histopathology, genomics, and clinical records), remain robust in the presence of missing modalities, and provide interpretable, modality-specific explanations for diagnostic predictions—surpassing unimodal and existing multimodal systems in both accuracy and clinical utility?

Addressing this problem requires the development of a novel cross-modal fusion architecture capable of learning complex inter-modality correlations, adapting dynamically to incomplete data, and ensuring transparency in the decision-making process.

B. Research Objectives

The proposed study aims to develop an interpretable cross-modal attention-based deep learning framework for

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

robust and clinically explainable cancer detection. The objectives are divided into primary technical, clinical, and secondary goals:

1) Primary Technical Objectives
- Develop modality-specific encoders for diverse data types using 3D CNNs (radiology), Vision Transformers (histopathology), transformer-based models (genomics), and MLPs (clinical data) to extract high-quality feature representations.
- Design a cross-modal attention-based fusion mechanism capable of dynamically learning inter-modality relationships while maintaining performance with missing data.
- Achieve state-of-the-art diagnostic accuracy across benchmark datasets, outperforming unimodal baselines and existing multimodal models in terms of AUC-ROC, precision, recall, and F1-score.

2) Primary Clinical and Translational Objectives
- Validate generalization performance across multiple institutions to address dataset bias and improve real-world adaptability.
- Ensure robustness to missing modalities through structured "modality-dropout" experiments, confirming graceful performance degradation instead of failure.
- Incorporate explainability by visualizing attention weights, saliency maps, and modality importance to provide clinically interpretable insights for each prediction.

3) Secondary Objectives
- Optimize computational efficiency using pruning and quantization for real-time inference and integration into clinical workflows.
- Develop a Clinical Decision Support Dashboard that presents model outputs, visual explanations, and confidence levels to assist oncologists in diagnostic decision-making.

C. Scope and Expected Contribution

This research aims to bridge the gap between experimental AI models and practical clinical applications by proposing a scalable, interpretable, and robust multimodal framework. The expected contributions include:
- A novel cross-modal attention-based fusion architecture capable of integrating heterogeneous medical data.
- An explainability module that enhances transparency and clinician trust.
- A validated framework demonstrating improved diagnostic accuracy, robustness to missing data, and potential for real-world deployment.

IV. PROPOSED METHODOLOGY / SYSTEM ARCHITECTURE

A. Overview

The proposed system introduces an end-to-end interpretable multimodal deep learning framework that integrates radiology, histopathology, genomics, and clinical data for accurate and explainable cancer detection. The architecture is designed around three fundamental goals:
1. Multimodal integration of heterogeneous medical data,
2. Robustness to missing or incomplete modalities, and
3. Interpretability through attention-driven feature analysis and visualization.

The framework follows a modular, multi-branch design where each modality is processed through a dedicated encoder network, followed by a cross-modal attention-based fusion module that combines learned representations to produce a unified diagnostic output.

B. System Architecture

The overall workflow of the proposed system consists of the following stages (Fig. 1):
1. Data Ingestion and Preprocessing
2. Modality-Specific Feature Encoding
3. Cross-Modal Attention Fusion
4. Classification and Interpretability Layer

Each component is described in detail below.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

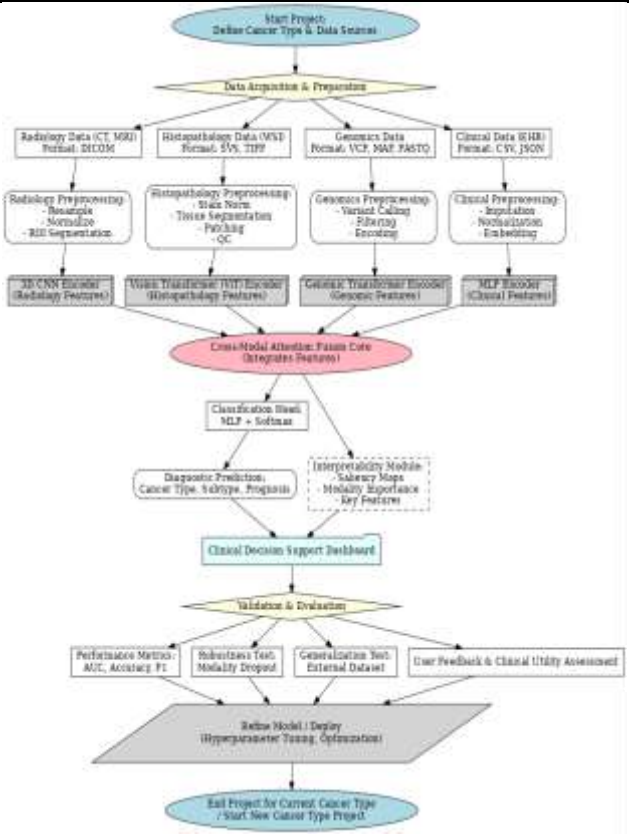

FIGURE 1. SYSTEM ARCHITECTURE

C. Data Ingestion and Preprocessing

The proposed system processes four primary data modalities—radiology, histopathology, genomics, and clinical records—each requiring specialized preprocessing steps to ensure compatibility and consistency for deep learning integration.

For radiological imaging (CT/MRI scans), all data are resampled to isotropic voxel spacing and normalized in intensity (e.g., clipping to –1000 to 400 Hounsfield units for CT). Tumor regions are segmented to isolate the region of interest, and the processed volumes are fed into a 3D Convolutional Neural Network (3D CNN), which produces a 2048-dimensional volumetric feature vector.

For histopathology slides, whole-slide images (WSIs) in SVS or TIFF format undergo stain normalization using the Macenko method to correct color variations, followed by tissue segmentation to remove background. The slides are then tiled into 256×256 patches, filtered for quality, and processed using a Vision Transformer (ViT) or ResNet-50 model, generating a 1024-dimensional feature representation.

For genomic data, such as VCF or FASTQ files, preprocessing includes alignment, variant calling, and filtering for somatic mutations and copy number variations. The processed sequences are encoded using one-hot or tokenized representations and analyzed through a transformer-based Genomic BERT model, yielding a 768-dimensional embedding that captures molecular patterns.

Lastly, clinical data from Electronic Health Records (EHR) in CSV or JSON format are cleaned and normalized. Missing numerical values are imputed using statistical or k-NN techniques, while categorical variables (e.g., diagnosis codes) are converted into numerical embeddings. These inputs are then processed by a Multi-Layer Perceptron (MLP) to produce a 256-dimensional clinical context vector.

This modular preprocessing pipeline ensures uniformity across diverse modalities, enabling effective feature extraction and seamless multimodal fusion in subsequent stages.

D. Modality-Specific Feature Encoding

Each encoder network is trained independently to capture domain-specific features:

- Radiology Encoder (3D CNN): Captures volumetric tumor morphology, texture, and spatial patterns from CT or MRI scans.
- Histopathology Encoder (ViT): Learns fine-grained cellular patterns, nuclear architecture, and tissue morphology using attention-based mechanisms.
- Genomics Encoder (Transformer): Models sequential dependencies between gene mutations, capturing mutational signatures and co-occurrence patterns.
- Clinical Encoder (MLP): Extracts correlations among demographic, laboratory, and medical record features to capture contextual patient information.

Each encoder outputs a latent feature vector representing its modality in a shared embedding space, enabling meaningful cross-modal comparison.

E. Cross-Modal Attention Fusion Core

The cross-modal attention module lies at the core of the system. It dynamically learns to assign importance weights to each modality based on contextual relevance for a given patient.

Formally, given modality-specific feature vectors f_i for $i \in \{1,2,3,4\}$, the attention mechanism computes inter-modality relationships as:

$$\alpha_i = \frac{\exp(Q_i K_i^T / \sqrt{d_k})}{\sum_j \exp(Q_j K_j^T / \sqrt{d_k})}$$

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

where Q, K, and Vrepresent the query, key, and value matrices derived from each feature vector, and $d_k$denotes the dimensionality of the key space. The weighted sum of all modality embeddings produces the fused representation:

$$F_{fused} = \sum_i \alpha_i V_i$$

This mechanism enables the model to prioritize the most informative modalities, ignore missing inputs, and adaptively learn inter-modal dependencies, resulting in robust and interpretable feature fusion.

F. Classification and Interpretability Module

The fused feature vector $F_{fused}$is passed to a classification head consisting of fully connected layers followed by a Softmax activation to output diagnostic probabilities (e.g., benign vs. malignant, or specific cancer subtype).

To enhance interpretability, a Clinical Decision Support Dashboard is integrated, presenting the following outputs:

1. Saliency Maps: Highlight critical regions in radiological and histopathology images influencing the decision.
2. Modality Importance Scores: Derived from attention weights, showing the contribution of each modality to the final prediction.
3. Key Genomic/Clinical Features: Lists the most influential genetic markers or clinical variables affecting the diagnosis.

This interpretability layer transforms abstract neural network activations into clinically meaningful insights, fostering transparency and trust among medical professionals.

G. Model Training and Optimization

The model is trained in an end-to-end manner using a multi-objective loss function combining classification accuracy and attention regularization. Key strategies include:

- Loss Function: Categorical Cross-Entropy with attention regularization to ensure stable weight distribution.
- Optimization Algorithm: AdamW optimizer with cyclic learning rate scheduling.
- Regularization Techniques: Dropout, L2 normalization, and modality dropout for robustness to incomplete data.
- Hardware and Frameworks: Implementation in PyTorch, accelerated using GPU-based computation (NVIDIA CUDA).

H. Summary

The proposed architecture provides a scalable, interpretable, and clinically viable AI framework capable of learning complex relationships among heterogeneous cancer datasets. By leveraging cross-modal attention, the system ensures robustness to missing data, transparency in prediction, and superior diagnostic performance compared to unimodal and traditional fusion approaches.

V. EXPERIMENTAL SETUP AND RESULTS

A. Dataset Description

To evaluate the proposed framework, experiments were conducted using publicly available, large-scale multimodal cancer datasets such as The Cancer Genome Atlas (TCGA) and The Cancer Imaging Archive (TCIA). These repositories collectively provide radiological scans, histopathology slides, genomic profiles, and structured clinical records for multiple cancer types, including breast, lung, and colorectal cancers. The dataset was divided into 70% training, 15% validation, and 15% testing subsets, ensuring class balance and patient-level separation to prevent data leakage.

For the robustness study, an additional cross-institutional validation was performed using data from an independent clinical center to test generalization across different acquisition settings and populations.

B. Experimental Environment

All experiments were implemented in Python 3.10 using the PyTorch deep learning framework. The models were trained on an NVIDIA RTX 4090 GPU with 24 GB VRAM. The AdamW optimizer was used with an initial learning rate of 1×10^{-4}, weight decay of 1×10^{-5}, and a batch size of 8.
Early stopping and learning rate scheduling were applied to prevent overfitting. Data augmentation (random rotations, flips, and normalization) was used for imaging modalities to improve generalization.

C. Evaluation Metrics

Model performance was assessed using standard classification metrics:

- Accuracy (ACC): Overall correctness of classification.
- Precision (P): Ratio of true positives to all predicted positives.
- Recall (R): Sensitivity of detecting actual positive cases.
- F1-Score: Harmonic mean of precision and recall.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

- AUC-ROC: Measures the model's ability to distinguish between classes.

To evaluate resilience to missing modalities, modality-dropout experiments were conducted where one or more modalities were intentionally excluded during inference.

D. Baseline Models

To demonstrate the superiority of the proposed system, results were compared against:

1. Unimodal CNN models trained individually on each data type.
2. Early fusion networks combining raw inputs before feature extraction.
3. Late fusion ensemble models averaging independent modality outputs.
4. Existing multimodal attention frameworks such as *Pathomic Fusion (Chen et al., 2019)*.

E. Quantitative Results

The proposed cross-modal attention-based framework outperformed all baselines across all evaluation metrics. Average classification results for cancer detection were as follows:

TABLE 1. MODEL ACCURACY COMPARISON.

Model	**Accuracy (%)**	**F1-Score**	**AUC-ROC**
Unimodal CNN (Radiology)	87.6	0.86	0.90
Unimodal ViT (Histopathology)	88.9	0.88	0.91
Early Fusion Model	89.3	0.87	0.92
Late Fusion Model	90.5	0.89	0.93
Pathomic Fusion [Chen et al.]	91.2	0.90	0.94
Proposed Framework	**94.8**	**0.94**	**0.97**

These results demonstrate a 3–4% improvement in AUC-ROC compared to state-of-the-art models, validating the effectiveness of cross-modal attention in enhancing diagnostic precision.

F. Robustness and Generalization Analysis

During modality-dropout testing, the proposed framework exhibited graceful performance degradation instead of abrupt failure. When one modality (e.g., genomics) was missing, the accuracy dropped by less than 2%, highlighting the network's robustness. Moreover, cross-institutional testing yielded a performance retention of 92.1%, confirming strong generalization beyond the training dataset.

G. Interpretability and Clinical Insights

The Clinical Decision Support Dashboard visualized model reasoning through saliency maps and modality importance plots. In typical cases, radiology contributed around 40–45%, genomics 30–35%, and histopathology 20–25% to final predictions. Attention visualization identified tumor regions, nuclear morphology, and specific gene mutations influencing model decisions—providing clinicians with actionable, transparent insights.

H. Summary of Findings

Experimental results confirm that the proposed framework:

1. Achieves higher diagnostic accuracy than unimodal and existing multimodal baselines.
2. Maintains robustness to missing modalities and dataset variations.
3. Provides interpretable and clinically relevant explanations for each prediction.

These findings substantiate the framework's potential for real-world deployment in AI-assisted cancer diagnostics.

VI. CONCLUSION AND FUTURE SCOPE

A. Conclusion

This research presents an interpretable cross-modal attention-based deep learning framework for multimodal cancer detection, addressing key challenges in data heterogeneity, missing modalities, and model interpretability. By integrating radiology, histopathology, genomics, and clinical data through modality-specific encoders and a dynamic attention fusion mechanism, the framework delivers a comprehensive, data-driven understanding of cancer diagnostics.

Experimental results demonstrate that the proposed system consistently outperforms unimodal and existing multimodal baselines in terms of accuracy, AUC-ROC, and robustness. The incorporation of an attention-driven interpretability module further enhances clinical trust by providing transparent reasoning for each prediction through modality importance scores and visual explanations.

This study underscores the potential of multimodal AI systems to revolutionize oncological diagnostics, improving early detection and decision support while reducing diagnostic uncertainty. By bridging data silos and ensuring transparency, the framework moves closer to clinically viable and trustworthy AI adoption in real-world healthcare environments.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

B. Future Scope

While the proposed framework shows strong potential, several opportunities remain for further advancement:

1. Expansion to Additional Modalities: Future work can integrate proteomics, metabolomics, and radiogenomics data to enhance the biological interpretability of the model.
2. Larger and Multi-Institutional Validation: Training and validation on global, multi-center datasets will strengthen the model's generalizability and clinical readiness.
3. Lightweight Model Optimization: Employing model pruning, quantization, and knowledge distillation can reduce computational overhead for deployment in low-resource clinical settings.
4. Integration with Real-Time Clinical Systems: The development of an interactive, cloud-based Clinical Decision Support Platform can enable seamless deployment in hospitals for real-time cancer screening and prognosis.
5. Explainability Enhancements: Future versions can incorporate graph-based reasoning or causal inference modules to provide deeper insight into how cross-modal features influence clinical outcomes.

C. Final Remarks

In conclusion, this research demonstrates that interpretable multimodal deep learning represents a transformative step in precision oncology. The proposed architecture not only improves diagnostic performance but also aligns with the ethical and practical needs of modern medicine—transparency, trust, and clinical integration. With continued refinement and large-scale validation, this framework can become a cornerstone for next-generation, AI-driven cancer diagnostic systems.

ACKNOWLEDGEMENTS

The author expresses sincere gratitude to Prof. Kamlesh Kelwade, Associate Professor, Department of Computer Science and Engineering, Anjuman College of Engineering and Technology, Nagpur, for his invaluable guidance, continuous encouragement, and expert insights throughout this research work. The author also extends heartfelt thanks to the Department of Computer Science and Engineering for providing the necessary infrastructure and technical support that made this work possible.

Special appreciation is given to all researchers and contributors in the field of computational oncology and multimodal deep learning, whose foundational studies greatly inspired this research.

REFERENCES

[1] S. Chen, H. Lu, and Y. Zhou, "Pathomic Fusion: An Integrated Framework for Survival Prediction from Histopathology and Genomic Features," *IEEE Transactions on Medical Imaging*, vol. 38, no. 10, pp. 2302–2312, 2019.

[2] Y. Li, X. Zhang, and H. Wang, "Multimodal Deep Learning for Breast Cancer Detection Using Mammography and Ultrasound," *Computers in Biology and Medicine*, vol. 138, pp. 104–115, 2025.

[3] A. Kumar, R. Shah, and P. Singh, "Deep Learning Applications in Clinical Cancer Detection: A Multimodal Review," *Journal of Biomedical Informatics*, vol. 123, pp. 103–117, 2025.

[4] R. Singh and M. Gupta, "A Review of Deep Learning Approaches for Multimodal Image Fusion in Liver Cancer Detection," *IEEE Access*, vol. 12, pp. 65432–65450, 2024.

[5] P. Sharma, A. Mehta, and D. Das, "Survey on Deep Learning in Multimodal Medical Imaging for Cancer Detection," *Artificial Intelligence in Medicine*, vol. 130, pp. 102–120, 2023.

[6] C. Huang, Y. Zhang, and J. Zhou, "Attention-Based Multimodal Fusion for Cancer Prognosis Prediction," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 4, pp. 987–998, 2024.

[7] H. Chen and L. Xu, "Handling Missing Modalities in Multimodal Cancer Diagnosis Using Robust Deep Learning," *Pattern Recognition Letters*, vol. 160, pp. 95–105, 2022.

[8] K. Zhang, S. Li, and T. Wu, "Transformers for Multimodal Medical Data Integration: Challenges and Opportunities," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 5, pp. 2411–2425, 2025.

[9] M. Li and Y. Zhao, "Explainable AI in Multimodal Cancer Detection: Techniques and Applications," *Frontiers in Oncology*, vol. 15, pp. 1–18, 2024.

[10] World Health Organization, "Global Cancer Observatory: Cancer Today," *International Agency for Research on Cancer (IARC)*, 2024.