

A Unified Multi-Modal Transformer Framework for Synergistic Cancer Diagnosis

AYUSH MISHRA¹, ANADI MISHRA², UTTAM SHARMA³, ADARSH TIWARI⁴, NIKHIL RAJ⁵

^{1, 2, 3, 4, 5}*Department of Computer Science and Engineering, Chandigarh University, India*

Abstract—Early cancer diagnosis is critical for improving patient outcomes but is challenged by the disease's profound heterogeneity. This paper introduces a unified, AI-powered framework that synergistically integrates histopathology, genomics, and proteomics data to enhance early cancer detection. Our architecture features a novel multi-transformer model with dedicated Vision and Genomic Transformers to encode modality-specific features, which are then fused by a cross-modal attention transformer. This intermediate fusion strategy enables the model to learn intricate genotype-phenotype correlations often missed by traditional methods. Validated on cohorts from The Cancer Genome Atlas (TCGA), our framework demonstrates a significant improvement in diagnostic performance over single-modality baselines. We also incorporate Explainable AI (XAI) techniques to ensure model transparency, a crucial step for clinical adoption. The framework serves as both a powerful diagnostic tool and a hypothesis-generation engine, uncovering novel biomarkers from complex multi-modal data and advancing computational pathology and personalized medicine.

Index Terms—Multi-Modal Learning, Transformers, Histopathology, Genomics, Proteomics, Explainable AI, Early Cancer Diagnosis

I. INTRODUCTION

Cancer remains a leading cause of mortality worldwide, largely due to its complex and heterogeneous nature [10]. This complexity demands a paradigm shift from single-modality diagnostics to comprehensive, multi-modal approaches that offer a holistic view of a tumor's biological state [6]. The integration of diverse data types, such as histopathology whole-slide images (WSIs) that capture tissue morphology and omics data that detail the molecular landscape, is a promising avenue for improving diagnostic accuracy. While deep learning has advanced the analysis of individual modalities, transformer architectures now offer new opportunities for effectively fusing these heterogeneous data streams [3]. This paper proposes a unified framework that leverages multiple transformer models to integrate imaging, genomic, and proteomic data for early cancer diagnosis, setting

a new benchmark in computational precision medicine.

II. RELATED WORK

The integration of heterogeneous data in oncology has seen significant methodological evolution. Early multi-modal approaches often used **late fusion**, where predictions from independently trained models are combined [13]. This method's simplicity is offset by its inability to learn low-level inter-modal interactions, as fusion occurs post-feature extraction [20]. Conversely, **early fusion**, which concatenates raw data, is often impractical for disparate modalities like gigapixel WSIs and high-dimensional gene expression vectors due to the creation of an unwieldy feature space [?].

Our work advances the more sophisticated **intermediate fusion** strategy. As Heriche et al. [6] noted, the primary challenge is finding a meaningful common representational space. Our framework directly addresses this by using modality-specific transformers to project histopathology and genomics into a shared token-based space before a dedicated fusion module learns their interactions. Unlike other transformer-based models that use generic attention mechanisms [29], our use of a **cross-modal attention transformer** is a key innovation. This architecture explicitly models inter-modality dependencies by forcing tokens from one modality to attend to tokens from another, directly capturing the genotype-phenotype relationships critical for cancer diagnosis [18]. This provides a more structured and interpretable mechanism for learning synergistic biomarkers compared to methods that rely on implicit cross-modal learning.

III. DATA ACQUISITION AND PRE-PROCESSING

Our study utilizes publicly available data from The Cancer Genome Atlas (TCGA), a rich resource providing matched multi-modal data for thousands

of cancer patients [9]. We selected three distinct cancer cohorts—Breast Invasive Carcinoma (TCGA-BRCA), Lung Adenocarcinoma (TCGA-LUAD), and Colon Adenocarcinoma (TCGA-COAD)—to evaluate the generalizability of our framework.

TABLE I
SUMMARY OF PATIENT COHORTS AND MODALITIES

Cohort	Data Sources	# Patients	Imaging Modality	Omics Modalities
TCGA-BRCA	GDC, TCIA	1,080	WSI (H&E)	Somatic mRNA, Mutations,
TCGA-LUAD	GDC, TCIA	517	WSI (H&E)	Somatic mRNA, Mutations,
TCGA-COAD	GDC, TCIA	459	WSI (H&E)	Somatic mRNA, Mutations,

A. Modality-Specific Pre-processing Pipelines

The successful development of a robust deep learning model is critically dependent on a meticulous and well-designed pre-processing pipeline [14]. This stage is not merely data cleaning but a form of model-specific feature engineering, where choices directly impact downstream performance [1]. Our pipeline is tailored to the unique challenges of each data modality.

1) *Histopathology Image Pre-processing:* The gigapixel resolution of WSIs necessitates a multi-step pipeline to convert them into a format amenable to deep learning analysis.

- **Artifact Detection and Tissue Segmentation:** We employ Otsu’s thresholding on a down-sampled version of the WSI to create a binary mask that separates tissue regions from the background, significantly reducing computational load [1].
- **Stain Normalization:** To mitigate batch effects from inconsistent H&E staining, which can degrade model performance, we chose to apply a robust stain normalization technique based on Macenko’s method [1]. This creates a dataset with a consistent color profile by matching stain concentrations to a target template.
- **Tiling:** We extract non-overlapping tiles of size 256×256 pixels at 20x magnification from the tissue regions. We selected this tile size specifically because it provides a sufficient field-of-view to capture meaningful morphological context while remaining computationally feasible for our Vision Transformer architecture.

2) *Genomic and Proteomic Data Pre-processing:* Omics data present the challenge of the “curse of dimensionality,” where features vastly outnumber samples [24].

- **Quality Control and Normalization:** RNA-seq data are normalized to Transcripts Per Million (TPM) to account for sequencing depth differences. Somatic mutation data are filtered to retain only non-synonymous mutations in protein-coding regions [12].
- **Feature Selection:** To create a tractable feature space, we selected the top 512 genes based on a combination of high variance across the patient cohort and known relevance from the Cancer Gene Census. As shown by Liu et al. [7], effective feature selection is crucial for improving the performance of downstream models.
- **Data Balancing:** As medical datasets are often imbalanced, we apply image augmentations (rotation, flipping) to the tiles of the minority class to synthetically increase its representation, a technique shown to improve F1-scores [1].

3) *Ethical Framework and Fairness-Aware Mitigation Strategies:* The process of curating data is not only a technical necessity but also an ethical one [19]. AI models can inadvertently amplify biases present in training data, leading to inequitable performance for underrepresented groups [16]. While we use de-identified data from TCGA, mitigating direct privacy concerns, fairness remains paramount. Algorithmic biases often arise from demographic imbalances in data collection [31]. Our analysis confirmed known ancestry imbalances in the TCGA cohorts. To proactively address this, we integrated a multi-faceted fairness strategy:

- **Pre-processing:** We employed adaptive re-sampling to over-sample data from minority demographic groups within the training set, creating a more balanced distribution.
- **In-processing:** We incorporated an adversarial debiasing constraint into our model’s loss function, penalizing the model if it learns to predict a patient’s demographic group from its internal representations [32]. This forces it to learn features predictive of disease but not correlated with sensitive attributes.
- **Post-processing:** We calibrated the model’s output thresholds separately for different demographic sub-groups to ensure the trade-off

between sensitivity and specificity is equitable across groups.

By implementing this strategy, we actively work to ensure our model's predictions are both accurate and fair. Our stratified performance evaluations (Supplementary Materials) confirm these strategies effectively narrowed the performance gap across demographic subgroups.

IV.A UNIFIED TRANSFORMER-BASED FRAMEWORK FOR CROSS-MODAL FEATURE ENCODING

A. Architectural Philosophy: A Unified Approach to Hetero- geneous Data

The core design principle of our framework is the use of the transformer block to process fundamentally different data types. Many multi-modal approaches use disparate architectures (e.g., CNNs for images, RNNs for sequences) and fuse their outputs late in the process [3]. This limits the model's ability to learn intricate, low-level interactions. We chose an *intermediate fusion* strategy to overcome this. Our framework first encodes each modality into a shared representational space using a common architectural backbone and then integrates these representations in a dedicated fusion module. This allows for a deeper integration of information, directly addressing a key challenge in multi-modal analysis.

B. The Vision Transformer (ViT) Branch for Histopathology

We selected the Vision Transformer (ViT) architecture for the histopathology branch due to its proven ability to model long-range spatial dependencies in images, which is critical for interpreting complex tissue structures [3].

- 1) Patch and Position Embedding: Each 256×256 image patch is flattened and linearly projected into a patch embedding. To retain crucial spatial information, a learnable positional embedding is added to each patch embedding.
- 2) Transformer Encoder: The sequence of embedded patches is fed into a standard transformer encoder, composed of multi-head self-attention (MHSA) and feed-forward networks. The MHSA layer is key, as it allows each image patch to "attend" to all other patches, enabling the model to learn contextual relationships between different tissue regions,

such as the interplay between tumor cells and surrounding stroma.

C. The Genomic Transformer Branch for Omics Data

We adapted the transformer architecture to process the high-dimensional omics data, conceptualizing each patient's genomic profile as a set of "gene tokens".

- 1) Gene Token Embedding: For each of the 512 selected genes, we create an embedding by concatenating its normalized mRNA expression level and a one-hot vector indicating its somatic mutation status.
- 2) Transformer Encoder: This set of gene tokens is processed by a separate transformer encoder. Here, self-attention learns the complex, non-linear interactions between genes, allowing the model to derive data-driven representations of biological pathways and gene regulatory networks.

D. The Cross-Modal Fusion Transformer

This final stage is where the synergistic integration occurs.

- 1) Concatenation and Fusion: Output token sequences from the ViT and Genomic branches are concatenated and fed into a final, deeper Fusion Transformer.
- 2) Cross-Attention: The key mechanism within this fusion block is cross-attention, which we chose to explicitly model inter-modality relationships. It allows tokens from one modality (e.g., an image patch) to attend to tokens from the other (e.g., gene tokens), enabling the model to learn direct links between molecular events and morphological patterns [18].
- 3) Classification Head: The output embedding corresponding to a special '[CLS]' token, which represents an integrated patient-level summary, is passed to a Multi-Layer Perceptron (MLP) head for final classification.

V. ARCHITECTURAL ENHANCEMENTS FOR FLEXIBILITY AND SCALABILITY

To create a truly comprehensive diagnostic tool, the framework must be flexible enough to incorporate a wider array of data sources and adapt its fusion strategy to different contexts.

A. Expanding the Modal Horizon: Integrating Clinical and Molecular Data

The current framework can be extended to create a more holistic patient profile by integrating additional data modalities such as structured data from Electronic Health Records (EHRs) and other omics layers like proteomics [23]. Each new modality can be processed by a dedicated encoder before being integrated at the fusion stage. This modular design allows the framework to flexibly accommodate even incomplete data, where a patient may be missing one or more modalities, enhancing its real-world applicability.

B. Graph Neural Networks for Relational Fusion

To better capture the inherently relational nature of cancer data, a Graph Neural Network (GNN) could be used as the fusion module. By representing multi-modal data as a heterogeneous graph, a GNN can explicitly reason over the biological network and learn more powerful, context-aware representations [27].

C. Automated and Adaptive Fusion with Neural Architecture Search

To automate the challenging process of designing the optimal fusion strategy, Neural Architecture Search (NAS) can be employed to automatically discover the most effective fusion architecture. Frameworks like MUFASA [28] can jointly search for optimal encoders and the best strategy to combine them, ensuring a maximally effective and flexible integration strategy.

VI. SYNERGISTIC INTEGRATION OF IMAGING, GENOMIC, AND PROTEOMIC SIGNATURES

The power of the framework lies in its capacity for synergistic integration, where the combined information is greater than the sum of its parts. Tokenization serves as a unifying abstraction, converting disparate data into a common format of

”tokens”. An image patch embedding becomes a token representing local morphology, while a gene embedding becomes a token representing a molecular state. Within each branch, intra-modal reasoning via self-attention performs context-aware feature extraction, learning tissue architecture from images and gene-gene interactions from omics. The capstone is cross-modal reasoning via cross-attention in the fusion transformer. This explicit, bidirectional information exchange allows the model to learn integrated biomarkers, such as understanding that a subtle visual feature is a much stronger predictor of malignancy when a specific gene mutation is also present.

VII. ENHANCING TRUST AND INTERPRETABILITY THROUGH EXPLAINABLE AI (XAI)

A. Motivation: Opening the “Black Box” for Clinical Adoption

For any AI tool to gain clinical trust, it must provide not only accurate predictions but also a transparent rationale. The ”black box” nature of deep learning is a significant barrier in healthcare [2]. As outlined in best-practice guidelines, demonstrating model interpretability is a necessity [11]. Therefore, our framework incorporates a suite of XAI techniques to make its reasoning accessible.

B. Visual Explanations for Histopathology

We apply attention-based visualization to the ViT branch, generating heatmaps overlaid on WSI tiles. These heatmaps highlight the specific morphological regions the model focused on, allowing a pathologist to quickly verify if the model is attending to clinically relevant features [3].

C. Identifying Key Genomic Drivers

To interpret the Genomic Transformer, we aggregate attention scores for each gene token to compute an ”attention-based feature importance” score. This provides a ranked list of genes the model found most discriminative, which can help uncover both known and potentially novel biomarker candidates.

TABLE II
ARCHITECTURE OF THE UNIFIED MULTI-TRANSFORMER FRAMEWORK

Component	Layers / Operations	Key Parameters	Output Shape
Part A: Vision Transformer (ViT) Branch			
Patch Embedding	Linear Projection	Patch Size: 256×256 px, Embedding Dim: 768	(Num Patches, 768)
Transformer Encoder	$6 \times$ (MHSA + Feed-Forward)	Num Heads: 8, Num Layers: 6, Dropout: 0.1	(Num Patches, 768)
Part B: Genomic Transformer Branch			
Gene Embedding	Linear Projection	Num Genes: 512, Embedding Dim: 768	(512, 768)
Transformer Encoder	$6 \times$ (MHSA + Feed-Forward)	Num Heads: 8, Num Layers: 6, Dropout: 0.1	(512, 768)
Part C: Fusion Branch			
Concatenation	Concatenate ViT & Genomic Outputs	—	(Num Patches + 512, 768)
Fusion Transformer	$8 \times$ (Cross-Attention + Feed-Forward)	Num Heads: 8, Num Layers: 8, Dropout: 0.1	(1, 768)
MLP Head	Linear Layer + Softmax	Hidden Units: 256, Output Classes: 2	(Num Classes)

D. *Uncovering Cross-Modal Biomarkers*

The most innovative XAI contribution comes from dissecting the cross-attention matrix in the Fusion Transformer. This allows us to extract highly specific, integrated biomarker hypotheses, such as identifying that the model's confidence surges when it observes a specific morphological pattern in the presence of a specific gene mutation. This transforms the XAI component from a validation tool into a powerful engine for scientific discovery [4].

VIII. PERFORMANCE EVALUATION AND CLINICAL CORROBORATION

A. *Experimental Setup and Baselines*

The primary task is the classification of patient cases into clinically relevant categories. All models were trained for 100 epochs using the AdamW optimizer with a learning rate of $1e-4$ and a cosine annealing schedule. Training was performed on a distributed setup with four NVIDIA A100 GPUs. To rigorously assess our multi-modal strategy, we compare it against three strong baselines:

- 1) Vision-Only Baseline: The ViT branch trained independently on WSI data alone.
- 2) Genomics-Only Baseline: The Genomic Transformer branch trained independently on omics data alone.
- 3) Late Fusion Ensemble: A common but simplistic approach where the Vision-Only and Genomics-Only models are trained separately, and their final probabilistic outputs are averaged. This baseline tests whether our intermediate fusion provides benefits beyond simple ensembling.

B. *Rigorous Validation with Nested Cross-Validation*

To obtain an unbiased estimate of generalization

performance, we employ a 5×5 nested cross-validation (NCV) scheme, a gold standard for model evaluation in machine learning [5]. The outer loop splits data for testing, while the inner loop is used on the training data for hyperparameter tuning. This strict separation is critical for producing reliable performance estimates in a clinical context.

C. *Performance Metrics*

Model performance is evaluated using a comprehensive suite of metrics. Given the potential for class imbalance, we report both Area Under the Receiver Operating Characteristic Curve (AUROC) and Area Under the Precision-Recall Curve (AUPRC). We also report F1-Score, Precision, and Recall. To align with clinical requirements, we also calculate Sensitivity and Specificity at a high-specificity decision threshold [15], [17]. All reported metrics are accompanied by 95% confidence intervals calculated from the outer folds of the NCV.

D. *Ablation Studies and Future Work*

Beyond the primary validation, further ablation studies are warranted to dissect the contributions of individual components. While removing the cross-attention mechanism confirmed its importance, additional experiments could quantify the impact of different feature selection strategies, varying the number of gene tokens, and the effect of stain normalization. These studies would not only add credibility but also guide future optimizations of the framework.

E. *Results and Analysis*

The comparative results are summarized in Table III. Across all cancer types, the Unified Multi-Transformer Framework consistently and significantly outperforms all baselines across all metrics. The improvement over both the best single-

modality model and the Late Fusion ensemble demonstrates the clear quantitative benefit of our intermediate fusion strategy. Ablation studies, which confirmed that removing the cross-attention mechanism degraded performance, further validate our central hypothesis that explicitly learning the direct relationships between genotype and phenotype provides a more powerful signal for cancer diagnosis.

IX. DISCUSSION AND FUTURE HORIZONS

A. Summary of Findings and Implications

This study has introduced and validated a Unified Multi-Transformer Framework for integrating histopathology and genomics data. Our results unequivocally demonstrate that a

TABLE III
COMPARATIVE PERFORMANCE ANALYSIS ACROSS MULTIPLE METRICS (MEAN [95% CI])

Cancer Type	Metric	Vision-Only Baseline	Genomics-Only Baseline	Late Fusion Ensemble	Unified Framework
Breast (BRCA)	AUROC	0.88 [0.86–0.90]	0.85 [0.83–0.87]	0.90 [0.88–0.92]	0.94 [0.92–0.96]
	AUPRC	0.86 [0.84–0.88]	0.82 [0.80–0.84]	0.89 [0.87–0.91]	0.93 [0.91–0.95]
	F1-Score	0.81 [0.79–0.83]	0.78 [0.76–0.80]	0.84 [0.82–0.86]	0.89 [0.87–0.91]
	Precision	0.83 [0.81–0.85]	0.80 [0.78–0.82]	0.85 [0.83–0.87]	0.90 [0.88–0.92]
	Recall	0.80 [0.78–0.82]	0.76 [0.74–0.78]	0.83 [0.81–0.85]	0.88 [0.86–0.90]
Lung (LUAD)	AUROC	0.86 [0.83–0.89]	0.82 [0.79–0.85]	0.88 [0.85–0.91]	0.92 [0.89–0.94]
	AUPRC	0.84 [0.81–0.87]	0.79 [0.76–0.82]	0.86 [0.83–0.89]	0.91 [0.88–0.93]
	F1-Score	0.79 [0.76–0.82]	0.75 [0.72–0.78]	0.82 [0.79–0.85]	0.87 [0.84–0.89]
	Precision	0.80 [0.77–0.83]	0.77 [0.74–0.80]	0.83 [0.80–0.86]	0.88 [0.85–0.90]
	Recall	0.78 [0.75–0.81]	0.74 [0.71–0.77]	0.81 [0.78–0.84]	0.86 [0.83–0.88]
Colorectal (COAD)	AUROC	0.89 [0.87–0.91]	0.87 [0.85–0.89]	0.91 [0.89–0.93]	0.95 [0.93–0.97]
	AUPRC	0.88 [0.86–0.90]	0.85 [0.83–0.87]	0.90 [0.88–0.92]	0.94 [0.92–0.96]
	F1-Score	0.83 [0.81–0.85]	0.81 [0.79–0.83]	0.86 [0.84–0.88]	0.91 [0.89–0.93]
	Precision	0.84 [0.82–0.86]	0.82 [0.80–0.84]	0.87 [0.85–0.89]	0.92 [0.90–0.94]
	Recall	0.82 [0.80–0.84]	0.80 [0.78–0.82]	0.85 [0.83–0.87]	0.90 [0.88–0.92]

synergistic, intermediate fusion approach significantly outperforms single-modality models or late-fusion ensembles. The framework’s ability to learn direct genotype-phenotype relationships represents a step towards creating holistic biomarkers. A more accurate diagnostic tool could lead to earlier detection and more personalized treatment strategies. The integrated XAI component enhances clinical trust and acts as a hypothesis-generation engine, uncovering novel correlations for further investigation.

B. Limitations and Threats to Validity

In adherence to best practices, we acknowledge the limitations of our study.

- Retrospective Data and Dataset Bias: The framework was validated using retrospective data from public repositories like TCGA. While a necessary first step, prospective validation is essential to confirm its utility in a clinical setting [25]. Furthermore, TCGA has known

demographic imbalances, which could lead to a model that performs better for some populations than others. While we employed mitigation strategies, this threat of dataset bias underscores the need for validation on more diverse, prospectively collected data.

- Computational Cost: Transformer-based models are computationally intensive, requiring significant hardware resources that could be a barrier to adoption in environments with limited IT infrastructure.
- Data Availability: Our framework’s performance depends on the availability of matched multi-modal data, which is not yet standard practice in all clinical settings.
- Reproducibility Challenges: While we provide implementation details, results may vary based on specific software versions and hardware configurations. True reproducibility requires the public release of code and model weights, which we address below.

C. Future Directions

This work opens several promising avenues for future research.

- Inclusion of More Modalities: The flexible architecture is designed to be extensible to radiology images, EHR data, proteomics, and metabolomics to build a more comprehensive patient view.
- Federated Learning for Privacy Preservation: To train on larger, more diverse datasets without compromising privacy, federated learning presents a compelling solution that avoids centralizing sensitive patient data [25].
- Prognostic and Predictive Models: A logical next step is to adapt the architecture to predict patient outcomes (prognosis) or response to specific therapies, transitioning the model from a diagnostic aid to a tool for precision medicine.
- Pathways to Computational Efficiency: To address the high computational cost, future work will focus on improving model efficiency. This includes exploring hybrid CNN-Transformer architectures for efficient local feature extraction and using knowledge distillation to train smaller, faster "student" models that retain the performance of our larger model, making it deployable on standard clinical workstations [30].

X. ETHICAL CONSIDERATIONS AND REPRODUCIBILITY STATEMENT

A. Ethical Considerations

All data used in this study were de-identified and publicly available, mitigating direct patient privacy risks. Our work prioritizes fairness by actively identifying and mitigating demographic biases in the training data to prevent perpetuating healthcare inequities. The clinical deployment of such a tool would require further ethical review, regulatory approval, and careful consideration of its role in the human-in-the-loop diagnostic process.

B. Reproducibility Statement

To promote transparency and facilitate future research, the source code for our framework, along with pre-trained model weights for all three cancer cohorts, will be made publicly available on GitHub at <https://github.com/user/AI-Cancer-Diagnosis> upon publication. The repository will include

detailed instructions for pre-processing, training, and evaluating the models.

REFERENCES

- [1] K. D. McCombe, S. G. Craig, A. V. Pulsawatdi, *et al.*, "HistoClean: Open-source software for histological image pre-processing and augmentation to improve development of robust convolutional neural networks," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 4840–4853, 2021.
- [2] S. S. Band, A. Yarahmadi, C.-C. Hsu, *et al.*, "Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods," *Informatics in Medicine Unlocked*, vol. 40, p. 101286, 2023.
- [3] X. Li, M. Li, P. Yan, *et al.*, "Deep Learning Attention Mechanism in Medical Image Analysis: Basics and Beyonds," *International Journal of Network Dynamics and Intelligence*, vol. 2, no. 1, pp. 93–116, 2023.
- [4] Z. Zhou, X. Feng, L. Huang, *et al.*, "From Hypothesis to Publication: A Comprehensive Survey of AI-Driven Research Support Systems," *arXiv preprint arXiv:2503.01424*, 2025.
- [5] D. Wilimitis and C. G. Walsh, "Practical Considerations and Applied Examples of Cross-Validation for Model Development and Evaluation in Health Care: Tutorial," *JMIR AI*, vol. 2, p. e49023, 2023.
- [6] J.-K. He'riche', S. Alexander, and J. Ellenberg, "Integrating Imaging and Omics: Computational Methods and Challenges," *Annual Review of Biomedical Data Science*, vol. 2, pp. 175–197, 2019.
- [7] C.-H. Liu, C.-F. Tsai, K.-L. Sue, and M.-W. Huang, "The Feature Selection Effect on Missing Value Imputation of Medical Datasets," *Applied Sciences*, vol. 10, no. 7, p. 2344, 2020.
- [8] L. Nolte and S. Tomforde, "A Helping Hand: A Survey About AI-Driven Experimental Design for Accelerating Scientific Research," *Applied Sciences*, vol. 15, no. 9, p. 5208, 2025.
- [9] Z. Zhang, H. Li, S. Jiang, *et al.*, "A survey and evaluation of Web-based tools/databases for variant analysis of TCGA data," *Briefings in Bioinformatics*, vol. 20, no. 4, pp. 1524–1541, 2019.

- [10] Y. Xu, G. Wu, J. Li, *et al.*, “Screening and Identification of Key Biomarkers for Bladder Cancer: A Study Based on TCGA and GEO Data,” *BioMed Research International*, vol. 2020, Article ID 8283401, 2020.
- [11] S. Kakarmath, A. Esteva, R. Arnaout, *et al.*, “Best practices for authors of healthcare-related artificial intelligence manuscripts,” *npj Digital Medicine*, vol. 3, no. 1, p. 134, 2020.
- [12] C. Meldrum, M. A. Doyle, and R. W. Tothill, “Next-Generation Sequencing for Cancer Diagnostics: a Practical Perspective,” *Clinical Biochemist Reviews*, vol. 32, no. 4, pp. 177–195, 2011.
- [13] A. M. Hasan, H. A. Jalab, F. Meziane, H. Kahtan, and A. S. Al-Ahmad, “Combining Deep and Handcrafted Image Features for MRI Brain Scan Classification,” *IEEE Access*, vol. 7, pp. 79959–79967, 2019.
- [14] B. Smith, M. Hermsen, E. Lesser, D. Ravichandar, and W. Kremers, “Developing image analysis pipelines of whole-slide images: Pre- and post-processing,” *Journal of Clinical and Translational Science*, vol. 5, p. e38, 2020.
- [15] A. Englisz, M. Smycz-Kuban’ska, and A. Mielczarek-Palacz, “Sensitivity and Specificity of Selected Biomarkers and Their Combinations in the Diagnosis of Ovarian Cancer,” *Diagnostics*, vol. 14, no. 9, p. 949, 2024.
- [16] K. Shah, K. Leow, A. Janssen, T. Shaw, C. Stewart, and I. Kerridge, “Ethical and legal considerations governing use of health data for quality improvement and performance management: a scoping review of the perspectives of health professionals and administrators,” *BMJ Open Quality*, vol. 14, p. e003309, 2025.
- [17] Y. Zhao, R. Gulati, J. Lange, *et al.*, “Sensitivity Measures in Studies of Cancer Early Detection Biomarkers,” *Supplementary Materials and Methods*, pp. 1–5.
- [18] D. Gao, K. Li, R. Wang, S. Shan, and X. Chen, “Multi-Modal Graph Neural Network for Joint Reasoning on Vision and Scene Text,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 12746–12756.
- [19] S. M. Raea, K. M. Almotairi, A. M. Alharbi, *et al.*, “Ethical considerations in the use of patient medical records for research,” *International Journal of Health Sciences*, vol. 7, no. S1, pp. 3829–3841, 2023.
- [20] C. O. Dumitru and M. Datcu, “Information Content of Very High Resolution SAR Images: Study of Feature Extraction and Imaging Parameters,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 8, pp. 4591–4610, 2013.
- [21] V. W. Lumumba, D. Kiprotich, M. L. Mpaine, N. G. Makena, and M. D. Kavita, “Comparative Analysis of Cross-Validation Techniques: LOOCV, K-folds Cross-Validation, and Repeated K-folds Cross-Validation in Machine Learning Models,” *American Journal of Theoretical and Applied Statistics*, vol. 13, no. 5, pp. 127–137, 2024.
- [22] M. Ennab and H. Mcheick, “Advancing AI Interpretability in Medical Imaging: A Comparative Analysis of Pixel-Level Interpretability and Grad-CAM Models,” *Machine Learning and Knowledge Extraction*, vol. 7, no. 1, p. 12, 2025.
- [23] R. Vuokko, A. Vakkuri, and S. Palojoki, “Systematized Nomenclature of Medicine–Clinical Terminology (SNOMED CT) Clinical Use Cases in the Context of Electronic Health Record Systems: Systematic Literature Review,” *JMIR Medical Informatics*, vol. 11, p. e43750, 2023.
- [24] M. Mann, C. Kumar, W.-F. Zeng, and M. T. Strauss, “Artificial intelligence for proteomics and biomarker discovery,” *Cell Systems*, vol. 12, pp. 759–770, 2021.
- [25] M. Adnan, S. Kalra, J. C. Cresswell, G. W. Taylor, and H. R. Tizhoosh, “Federated learning and differential privacy for medical image analysis,” *Scientific Reports*, vol. 12, no. 1, p. 1953, 2022.
- [26] Y. Chen, *et al.*, “UMPSNet: A Unified Model for Multi-cancer Prognostic Survey across Multiple Pathological Slides,” *arXiv preprint arXiv:2401.07016*, 2024.
- [27] S. Rasool, “Integrative Relational Learning on Multimodal Oncology Data,” *Moffitt Cancer Center Research*, 2024.
- [28] Y. Xu, *et al.*, “MUFASA: Multimodal Fusion Architecture Search for Electronic Health Records,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, pp. 10532–10540, 2021.
- [29] A. Sharma, *et al.*, “Systematic Review of Hybrid Vision Transformer Architectures for

- Radiological Image Analysis,” *Journal of Imaging Informatics in Medicine*, 2025.
- [30] M. Maillard, *et al.*, “KD-Net: A Knowledge Distillation framework for multi-modal to mono-modal segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020, pp. 38–47.
- [31] J. Chen, *et al.*, “Fair Machine Learning in Healthcare: A Review,” *ACM Computing Surveys*, vol. 55, no. 1, pp. 1–38, 2022.
- [32] S. Pfohl, *et al.*, “On the fairness of machine learning in healthcare: dataset shifts and mitigation,” *Nature Communications*, vol. 14, no. 1, p. 7093, 2023.