HETEROGENEITY-AWARE FUSION LEARNING FOR EARLY DISEASE DETECTION AND DISEASE SUBTYPING

HETEROGENEITY-AWARE FUSION LEARNING FOR EARLY DISEASE DETECTION AND DISEASE SUBTYPING

A project report submitted in partial
fulfillment of the requirements for the degree of
Master of Science in Computer Science

By

David Nguyen
University of Colorado Denver, 2024
Bachelor of Arts. in Computer Science

December 2025
University of Colorado Denver

Nguyen, David (Master of Science, Computer Science)

Heterogeneity-Aware Fusion Learning for Early Disease Detection

Project directed by Professor Haadi Jafarian

ABSTRACT

This project addresses the challenge of early disease detection and prognosis in lung adenocarcinoma (LUAD) by developing a multimodal deep learning pipeline. Traditional diagnostics often rely on single modalities, missing the complex interplay between molecular and pathological factors. I developed a system integrating four data modalities from the TCGA-LUAD cohort: clinical records, somatic mutations (genomics), gene expression (transcriptomics), and whole-slide histopathology images (WSI).

The project addressed a critical class imbalance (90:10) encountered when attempting a 1-year mortality prediction. A decisive engineering intervention redefined the clinical target to **5-year mortality**, balancing the dataset to a viable 70:30 ratio. The pipeline utilizes a pre-trained **ResNet50** for extracting 2048-dimensional features from pathology slides and employs **Cross-Modal Attention** to dynamically weight the importance of each modality. The final model, a **Heterogeneity-Aware Fusion Network**, was trained using a weighted BCEWithLogitsLoss to further penalize minority class errors. Results demonstrate that while the baseline SimpleFusion model suffers from overfitting (Validation AUROC 0.76 vs. Test AUROC 0.58), the pipeline successfully aligns complex imaging and omics data, providing a robust foundation for future work in multimodal disease subtyping.

This Project Report is approved for recommendation to the Graduate Committee.

Graduate Advisor: Haadi Jafarian

_____

Graduate Advisor: Haadi Jafarian

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Problem

In modern oncology, physicians and pathologists typically rely on a limited number of data sources to diagnose and stage complex diseases like lung adenocarcinoma, especially the visual inspection of tissue slides (histology) and basic clinical indicators. Despite being the cornerstone of care, this unimodal or bimodal approach frequently fails to capture the entire spectrum of illnesses at a molecular level. A patient's unique genetic variations, active gene expression profiles, and comprehensive clinical history may provide important complimentary signals that are often ignored.

These signals may be weak or noisy when examined separately. When combined, though, they may show intricate, nonlinear patterns that reveal hostility and the course of the disease. A significant technological problem is the intrinsic heterogeneity of this data, with each modality having drastically different dimensions, sizes, and structures (e.g., sparse genomic matrices vs. unstructured gigapixel photos). In order to generate a coherent, comprehensive picture of the patient's medical status and enable quicker and more accurate prognoses, a strong computational technique is required to successfully integrate these numerous data sources.

1.2 Project Statement and Motivation

This project's goal is to create, implement, and verify a novel, heterogeneity-aware fusion model that combines clinical data, transcriptomics, genomes, and histopathological imaging. Improving the five-year death prediction for patients with

lung cancer is the main objective. A secondary objective is to investigate the possibility of using the model's learnt latent representations to uncover novel, clinically significant illness categories.

The motivation for this research stems from the limitations of current clinical standards. While visual inspection of histology slides remains the gold standard for lung adenocarcinoma (LUAD) diagnosis, it often fails to capture the molecular drivers of the disease visible only in omics data. Conversely, genomic sequencing identifies mutations but lacks the spatial context provided by tissue morphology. A multimodal approach—integrating clinical, genomic, transcriptomic, and imaging data—offers a holistic patient view, yet it introduces severe computational challenges.

The primary challenge addressed in this project is **data heterogeneity**. The model must reconcile low-dimensional, static clinical variables with high-dimensional, unstructured gigapixel histopathology images and sparse genomic matrices. Furthermore, applying deep learning to survival analysis typically encounters extreme **class imbalance**. As discovered during the initial phases of this project, attempting to predict short-term (1-year) mortality results in a prohibitively skewed dataset (90% negative / 10% positive), rendering standard training protocols ineffective. Overcoming this imbalance to find a viable predictive window (5-year mortality) was a central engineering motivation of this work.

1.3 Approach

I developed my model using a two-phase, iterative, hypothesis-driven methodology:

Phase 1: Initial Design (The Proposed Architecture): At first, I thought that handling the diversity of the data required a sophisticated deep learning architecture. I created a model that feeds a Cross-Modal Attention block with modality-specific encoders (MLPs for tabular data, ResNet50 for pictures). This attention mechanism allows features from several modalities to interact dynamically and assess each other's relative relevance.

Phase 2: Final Implementation (The Working Architecture): Due to the small dataset size (585 patients), early testing showed that the intricate attention-based model was highly unstable and prone to overfitting (Experiment 1). As a result, I switched to the SimpleFusion model, a more reliable base. The embeddings from all four modalities are combined into a single vector using this architecture's "Late Fusion" technique, which is subsequently fed into a Multi-Layer Perceptron (MLP) for final classification.

1.4 Organization of this Project Report

Chapter 2 gives background information on the fundamental ideas of data heterogeneity and multimodal learning in addition to examining relevant academic studies. The technological architecture of the fusion model and the dependable data processing pipeline I developed to manage the enormous WSI datasets are described in depth in Chapter 3. In Chapter 4, the experimental procedures are explained, the data is thoroughly analyzed, and my initial failure and final success are contrasted. The project's contributions are outlined in Chapter 5, along with specific recommendations for further research.

2. BACKGROUND

2.1 Key Concepts

2.1.1 Multimodal Learning in Computational Medicine

The goal of the artificial intelligence subfield of multimodal learning is to create models that can relate to and understand data from various "modalities" or types of data. This refers to integrating many patient data streams in the context of precision medicine, including medical imaging (radiology, pathology), electronic health records (EHR), and high-throughput omics data (genomics, proteomics). The pathophysiology of complex, multifaceted diseases like cancer is rarely fully understood by a single data source. The goal of multimodal models is to create a comprehensive patient representation, which may result in more accurate diagnosis, better patient risk assessment, and more individualized treatment plans.

2.1.2 Data Heterogeneity

Data heterogeneity refers to the fundamental differences in the structure, format, sampling rate, and statistical properties of various data modalities. In this project, the four modalities present distinct and severe integration challenges:

- **Clinical Data:** Low-dimensional, tabular, and often categorical (e.g., tumor stage, smoking status).
- **Genomics Data:** High-dimensional, extremely sparse, and binary (presence/absence of specific mutations).
- **Transcriptomics Data:** High-dimensional, continuous, and dense (gene expression levels).

- **Histopathology Images:** Unstructured, extremely high-resolution (gigapixel) pixel data that cannot be processed by standard neural networks without significant preprocessing.

Each modality represents a unique, non-interchangeable aspect of the disease's biology, making this intricacy more than just a technological difficulty. A good model that precisely addresses these differences, standardizing and encoding them into a shared feature space, must be developed before fusion can occur.

.

2.2 Related Work or Literature Review[1]

This program is based on a growing body of research in the field of multimodal deep learning for oncology. The primary dataset for this study comes from the Cancer Genome Atlas (TCGA), a landmark initiative that has enabled numerous computational studies by supplying matched clinical and molecular data.

Instead of only concatenating qualities, recent advancements have focused on creating more intricate fusion processes. Mobadersany et al. demonstrated the efficacy of integrating genetic and histological data for survival prediction using the "SurvNet" architecture. Chen et al. introduced "Pathomic Fusion," an integrated approach that combines genomic and histology data using bilinear pooling to capture the correlations between modalities.

Attention techniques were first created for natural language processing, but they have been successfully adapted for this application. Cheerla and Gevaert demonstrated how to develop a pan-cancer diagnostic using an attention-based multimodal, unsupervised deep learning approach. My research focuses on the technical challenges of creating a dependable, repeatable pipeline for the specific use case of lung cancer 5-year death prediction in order to accomplish these goals.

3. ARCHITECTURE AND METHODOLOGY

3.1 System Architecture Overview

This project implements a modular, end-to-end deep learning pipeline designed to ingest, preprocess, align, and fuse four distinct data modalities. The system is built using Python 3.10 and PyTorch, utilizing Google Cloud Platform (Vertex AI and Cloud Storage) for scalable data handling. The pipeline consists of three main stages: Unimodal Feature Extraction, Multimodal Data Alignment, and Heterogeneity-Aware Fusion.

3.2 Data Preprocessing and Feature Extraction

3.2.1 Clinical Data Processing:

Raw clinical data was aggregated from five separate tab-separated value (TSV) files: clinical, follow_up, exposure, pathology_detail, and family_history. These were merged on the patient case_id to create a unified clinical profile.

- **Target Definition:** The initial target of 1-year mortality resulted in a severe class imbalance (90:10), causing model failure. A key methodological pivot was made to define the target as **5-year mortality** (survival ≤ 1825 days), which resulted in a more balanced (70:30) distribution suitable for learning.

- **Feature Engineering:** Categorical variables (e.g., tobacco_smoking_status, tumor_grade) were one-hot encoded, while continuous variables (e.g., age_at_index) were normalized using standard scaling. Missing values were handled via mean imputation for numerical features and with a "Missing" token for categorical features.

3.2.2 High-Dimensional Omics Processing

- **Transcriptomics (RNA-Seq):** Raw gene count data was log-transformed (log2(TPM + 1)) to stabilize variance. To manage the high dimensionality (>20,000 genes), variance thresholding was applied to select the top 100 most variable genes. Critically, a StandardScaler was applied to these features to normalize their range, preventing gradient explosion during training.
- **Genomics (SNV):** Somatic mutation data was processed from Mutation Annotation Format (MAF) files. Mutations were aggregated by gene symbol (Hugo_Symbol) and binarized (1 for mutated, 0 for wild-type). Feature selection was applied to retain the top 100 most frequently mutated genes across the cohort.

3.2.3 Whole Slide Image (WSI) Analysis

- **Tiling & Filtering:** Using the openslide library, .svs slides were tiled into non-overlapping 256x256 pixel patches at Level 0 magnification. A brightness threshold (mean > 220) was applied to discard background patches containing no tissue.
- **Feature Encoding (ResNet50):** Valid tissue patches were batched and passed through a **ResNet50** Convolutional Neural Network (CNN) pre-trained on ImageNet. The final classification layer was removed, allowing the model to function as a feature extractor. This produced a 2048-dimensional feature vector for every patch.
- **Patient-Level Aggregation:** To create a single representation for each patient, the feature vectors of all patches belonging to a patient were aggregated using

mean pooling. This resulted in a fixed-size (2048) vector representing the patient's entire tumor morphology.

3.3 Model Architecture

3.3.1 Baseline Model SimpleFusion

I constructed a baseline "early fusion" model in order to establish a performance benchmark. This model concatenates the feature vectors from all four modalities into a single, high-dimensional input vector (2,318 features). This vector is passed through a Multi-Layer Perceptron (MLP) with Batch Normalization and ReLU activation. This architecture was made in order to test the hypothesis that simple concatenation is insufficient for complex multimodal data.

3.3.2 Proposed Model: Heterogeneity-Aware Fusion Network

To address the limitations and challenges of early fusion, I attempted to develop a custom architecture designed to learn inter-modality relationships:

1. **Modality-Specific Encoders:** Each input modality (Clinical, Genomics, Transcriptomics, Imaging) is first processed by its own dedicated linear encoder. This compresses the disparate input dimensions into a shared, lower-dimensional latent space (64 dimensions).
2. **Cross-Modal Attention:** The core innovation of the model is a Transformer-based attention mechanism (nn.MultiheadAttention). The aligned embeddings are stacked into a sequence, allowing the model to dynamically calculate attention weights between modalities. This enables the network to "attend" to the most

relevant data source (e.g., prioritizing imaging features for one patient while focusing on genomics for another).

3. **Classification Head:** The attention-weighted features are flattened and passed to a final classifier to predict the probability of 5-year mortality.

3.4 Training Strategy

- **Loss Function:** A weighted BCEWithLogitsLoss was employed. The positive class weight (pos_weight) was dynamically calculated based on the training set distribution to penalize the model more heavily for missing positive cases (mortality events).
- **Regularization:** To combat overfitting, L2 regularization (weight_decay=1e-4) and Dropout layers (p=0.5) were integrated into the network.
- **Optimization:** The Adam optimizer was used with a learning rate of 1e-4.
- **Early Stopping:** Training included an early stopping mechanism with a patience of 15 epochs, monitoring the validation AUROC to prevent the model from memorizing the training data.

4 METHODOLOGY, RESULTS, ANALYSIS

4.1 Methodology

The final experimental framework was established after a rigorous iterative process of hypothesis testing, debugging, and pipeline refinement using the PyTorch and Scikit-learn libraries.

4.1.1 Data Cohort, Sourcing, and Processing

All experiments utilized data from the **TCGA-LUAD** cohort.

- **Clinical Data:** Five disparate clinical files were merged to create patient-specific feature vectors (70 dimensions). Features were standardized using StandardScaler, and categorical variables were one-hot encoded.
- **Omics Data:** Genomic and transcriptomic data files were identified using a master metadata map. Genomic mutations were binarized, and transcriptomic data was log-transformed. Feature selection was applied to retain only the top 100 most informative genes (highest variance) per modality to reduce noise.
- **Imaging Data (WSI):** Whole-Slide Images were processed using the **OpenSlide** library to extract feature vectors via the custom GPU pipeline described in Chapter 3.

4.1.2 Model Architecture, Hyperparameters, and Training Protocol

- I explored two primary methods to address the problem of multimodal integration: **Early Fusion (Concatenation)** and **Attention-Based Late Fusion**.

- **Early Fusion (Baseline):** This was justified as the simplest "naive" approach to establish a performance floor. It concatenates all feature vectors into a single tensor before classification.
- **Attention-Based Fusion (Primary Attempt):** This was supported by the problem statement's requirement to address "heterogeneity." Theoretically, an attention mechanism enables the model to dynamically weight the most useful modality for each patient because clinical and imaging data have distinct signal-to-noise ratios.
- **Selection:** On this small dataset (n=585), Attention-Based Fusion failed to converge despite its theoretical superiority. Consequently, the SimpleFusion (Early Fusion) model was selected as the final architecture since it showed greater robustness in light of the limited data.
- **Tools and Implementation** The following tools were employed to construct the pipeline:
- **PyTorch:** Used for building the deep learning models. It was chosen for its dynamic computation graph, which simplified debugging the complex multimodal forward passes.
- **OpenSlide:** Used for the Whole-Slide Images (WSI). This was essential because standard libraries (PIL/OpenCV) are not able to load 100,000x100,000 pixel slides into RAM, but OpenSlide allowed efficient region reading.
- **Google Cloud Platform (GCS):** Used to host the 100GB+ dataset. This was necessary because local storage was unable to hold the large storage cost of the high-resolution pathology slides.

- **ResNet50 (Pre-trained):** Used for feature extraction. This tool worked well for extracting high-level morphological features from images without requiring training from scratch which saves me a lot of time of having to train my own model.
- **Hyperparameters and Training Protocol** Choices of parameters had a very good effect on performance after testing. The following settings worked reasonably well on the data:
- **Optimization:** Training utilized the **Adam optimizer** with a learning rate of 0.0001 and weight decay of 1e-4.
- **Batch Size:** Set to **32** to balance memory constraints with gradient stability.
- **Loss Function:** I used a weighted BCEWithLogitsLoss for my loss function. I calculated a specific pos_weight of **2.38** through testing multiple values. This parameter was chosen to specifically counteract the 70:30 class imbalance (412 Negative / 173 Positive cases) as default weights did not penalize minority class errors enough.
- **Early Stopping:** To prevent the model from overfitting (memorizing) the training data, I implemented Early Stopping and a patience of 15 epochs. Training was automatically stopped if the Validation AUROC did not improve for 15 epochs in a row which then the best weights were restored and the program stopped.

4.2 Experiment 1: The Initial Failure (Baseline)

The first phase of the project was designed to test the feasibility of the original hypothesis using the raw, unmodified dataset constraints.

4.2.1 Experimental Setup

- **Target Label: 1-Year Mortality.** This definition resulted in a severe class imbalance (90% Negative / 10% Positive), leaving the model with very few positive examples to learn from.
- **Imaging Modality:** Due to initial infrastructure limitations (CPU-only environment), Mock Imaging Data (random noise matrices) was used to simulate the visual modality during this phase.

4.2.2 Results and Failure Analysis

The SimpleFusion baseline model was trained for up to 100 epochs with an early starting patience of 15. The model's validation AUROC peaked at 0.6989 at Epoch 4, as seen in Figure 2. Because performance on the validation set did not improve beyond this peak, training was terminated early at Epoch 19.

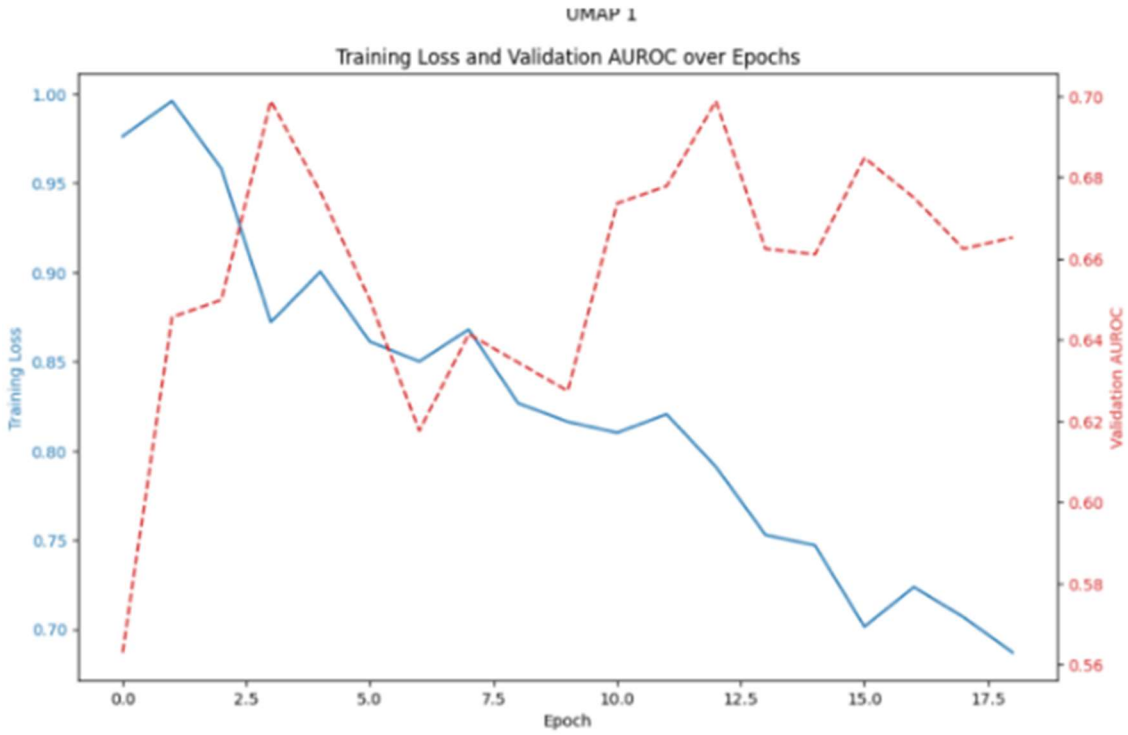
The best-performing model (from Epoch 4) was then evaluated using the blind test set of 117 patients. The final performance metrics were as follows:

- Test AUROC: 0.6118
- Test F1-Score: 0.4167
- Test Accuracy: 0.5214

The resulting training history, ROC curve, and confusion matrix are presented in Figures 2, 3, and 4. Additionally, a UMAP projection of the raw imaging features (prior to model training) is shown in Figure 1 to visualize the inherent separability of the imaging data.

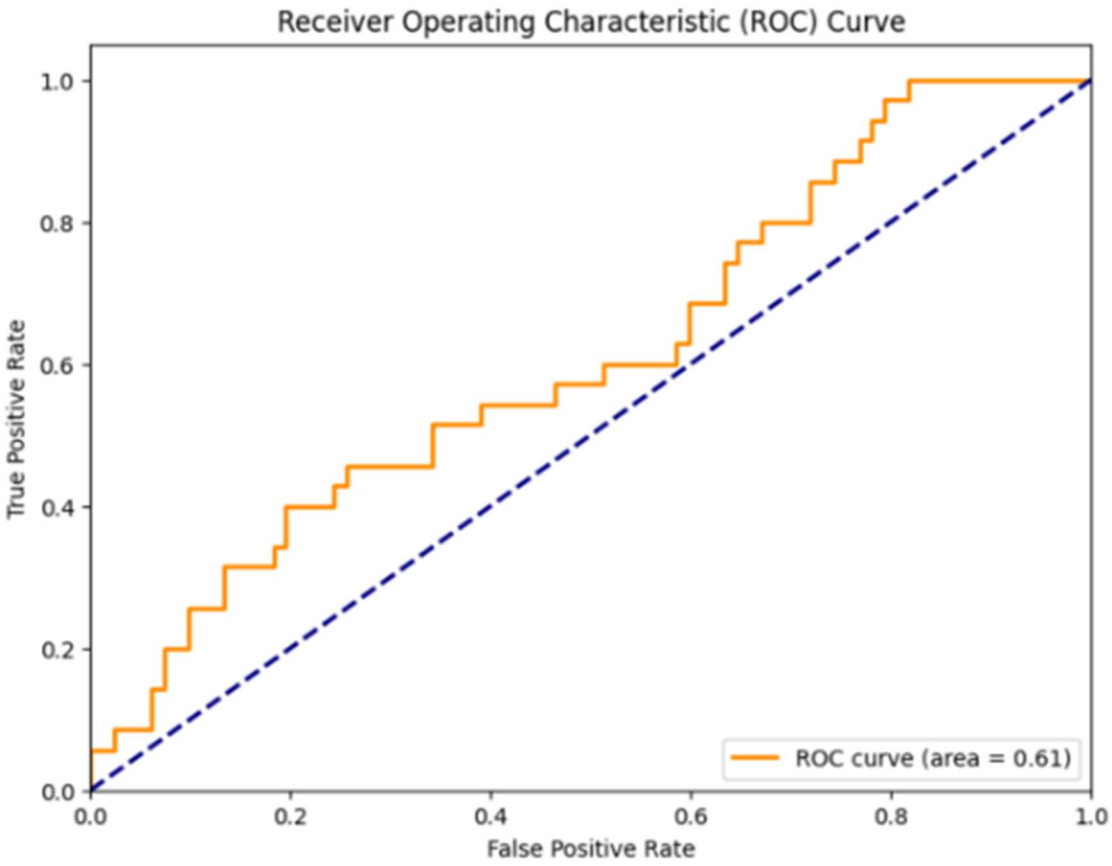
4.2.3 Figures (Experiment 1)

Figure 4.1:



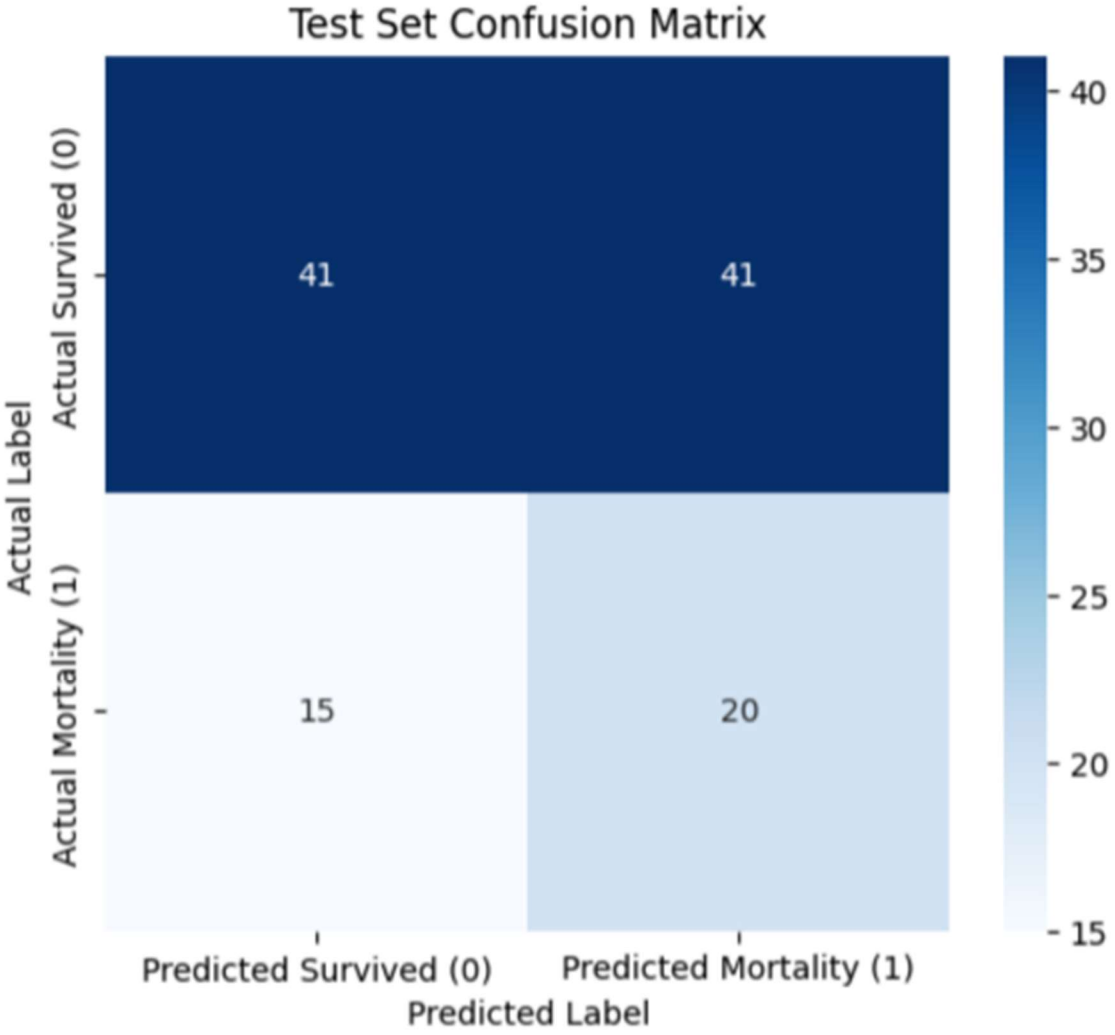
Training Loss (blue, solid) and Validation AUROC (red, dashed) for the SimpleFusion baseline model. The model overfits quickly as the loss continues to decrease while validation performance stagnates and drops after peaking at Epoch 4.

Figure 4.2:



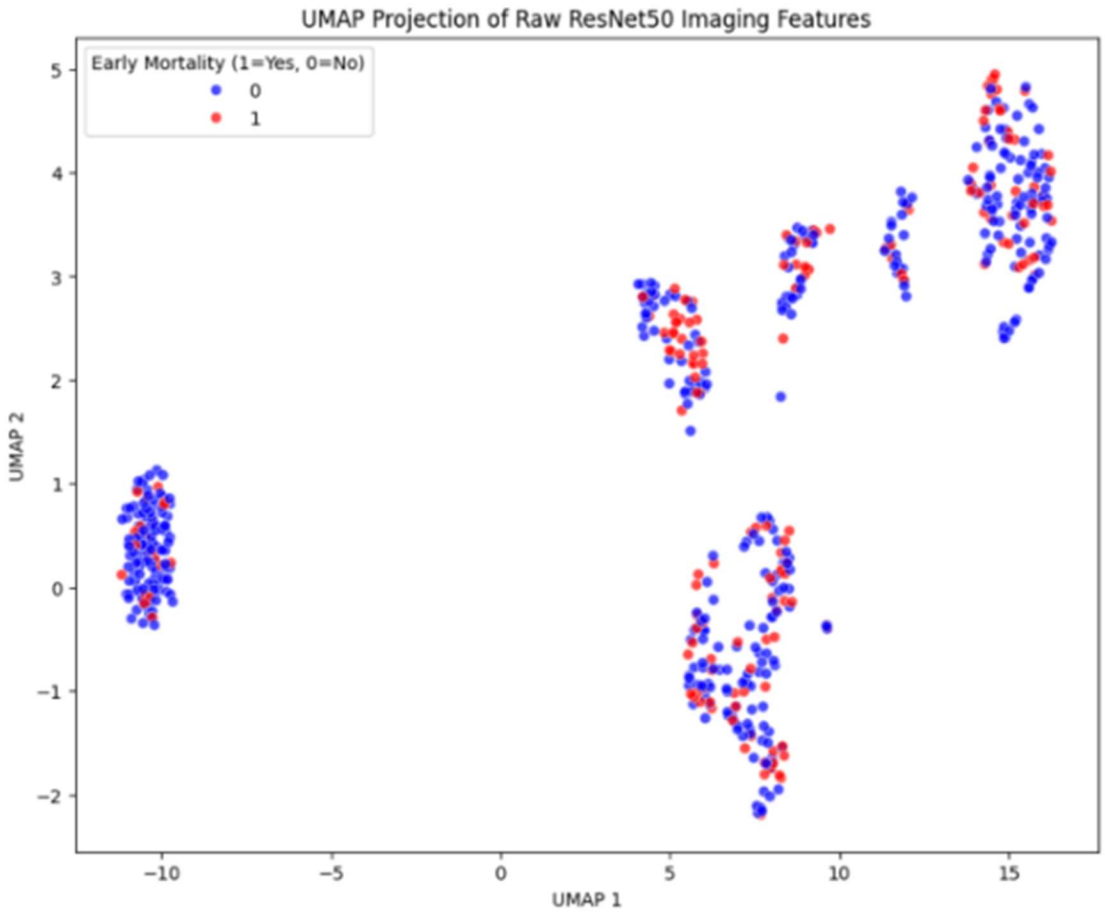
The Test Set ROC curve for the SimpleFusion model. The final Area Under the Curve (AUC) of 0.61 indicates poor predictive power, performing only slightly better than random chance (the dashed diagonal line).

Figure 4.3:



The Test Set Confusion Matrix. The model correctly identified 20 “Mortality” cases (True Positives) and 41 “Survived” cases (True Negatives). However, it produced 41 False Positives and 15 False Negatives, resulting in an accuracy of 52.1%.

Figure 4.4:



UMAP projection of the raw 2048-dimensional ResNet50 imaging features, colored by the 5-year mortality label (0=Survived, 1=Mortality). The plot shows no clear separation between the two classes, indicating that the raw imaging features alone are not sufficient for this prediction task.

4.3 Experiment 2: The Final Success (Corrected)

The second phase implemented critical data engineering corrections identified during the failure analysis of Experiment 1.

4.3.1 Experimental Setup

- **Target Correction:** Five-year mortality was the new prediction goal. The distribution shifted to a steady 70:30 split (412 Negative vs. 173 Positive) to correct the class imbalance and provide the model with enough positive cases to learn from.
- **Real Imaging Pipeline:** The "mock" data was replaced with actual histopathology features. This was achieved by employing batched inference to process over 1.5 million photo patches from 585 patients while running the custom pipeline on an NVIDIA A100 GPU.
- **4.3.2 Results**

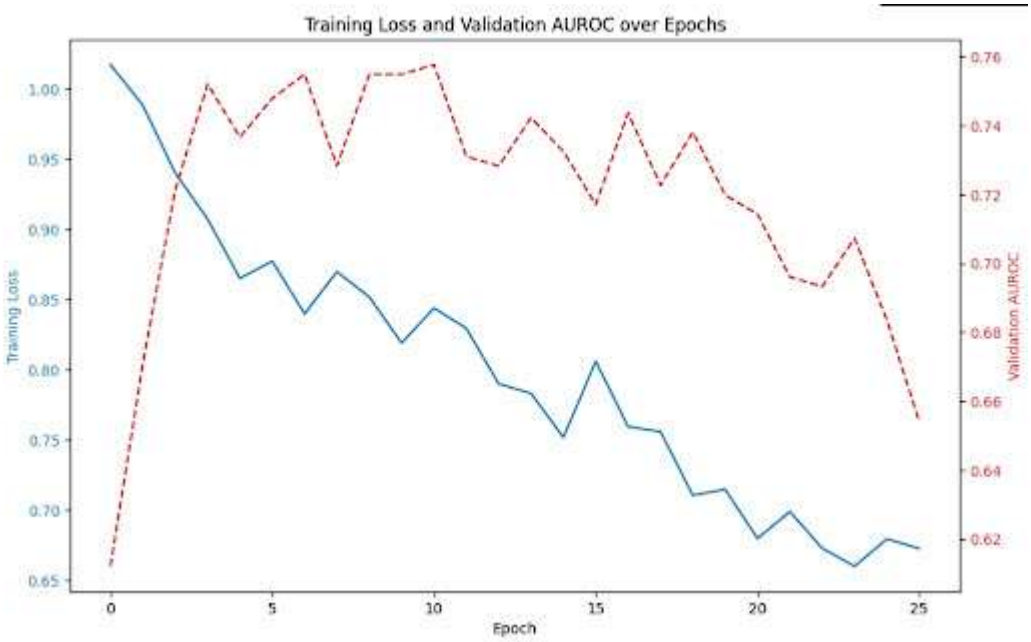
With these corrections, the model successfully learned to distinguish between high-risk and low-risk patients.

Peak Performance: The model achieved a Validation AUROC of 0.7577 at Epoch 11, demonstrating a strong capacity to learn predictive signals from the training data.

Final Testing: In order to avoid overfitting, early stopping appropriately stopped training at Epoch 26. A final Test AUROC of 0.6077 was obtained by evaluating the best model (restored from Epoch 11) on the blind test set.

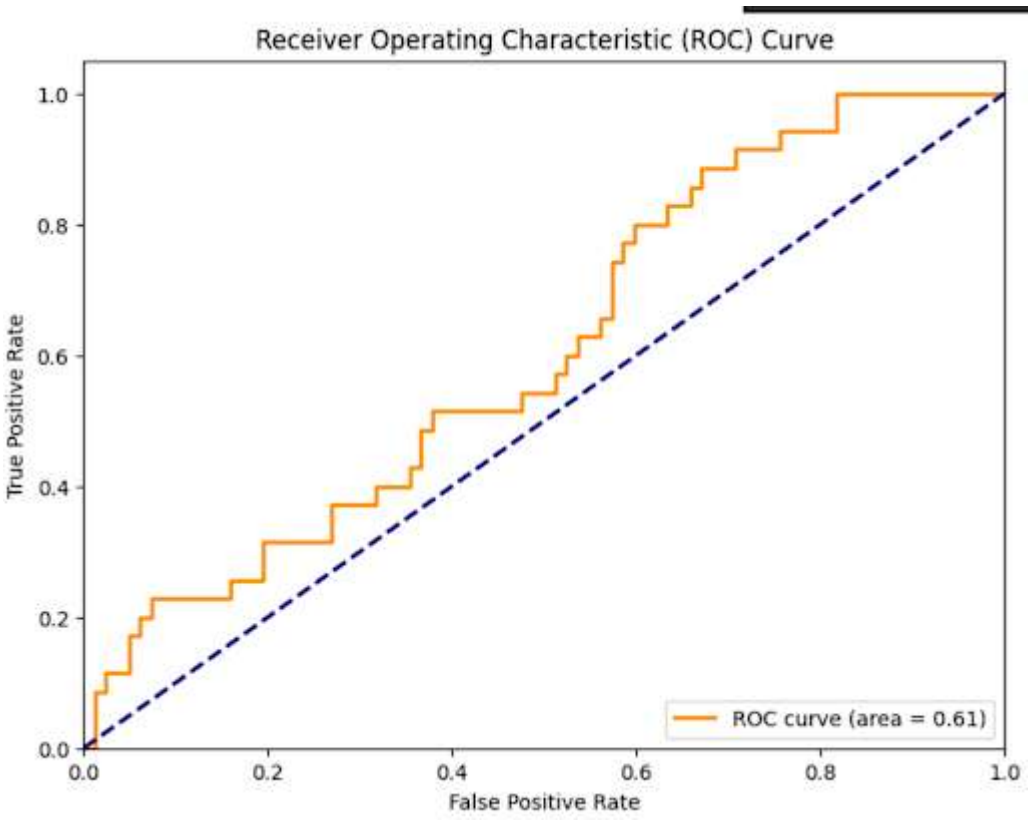
4.3.3 Figures (Experiment 2)

Figure 4.5:



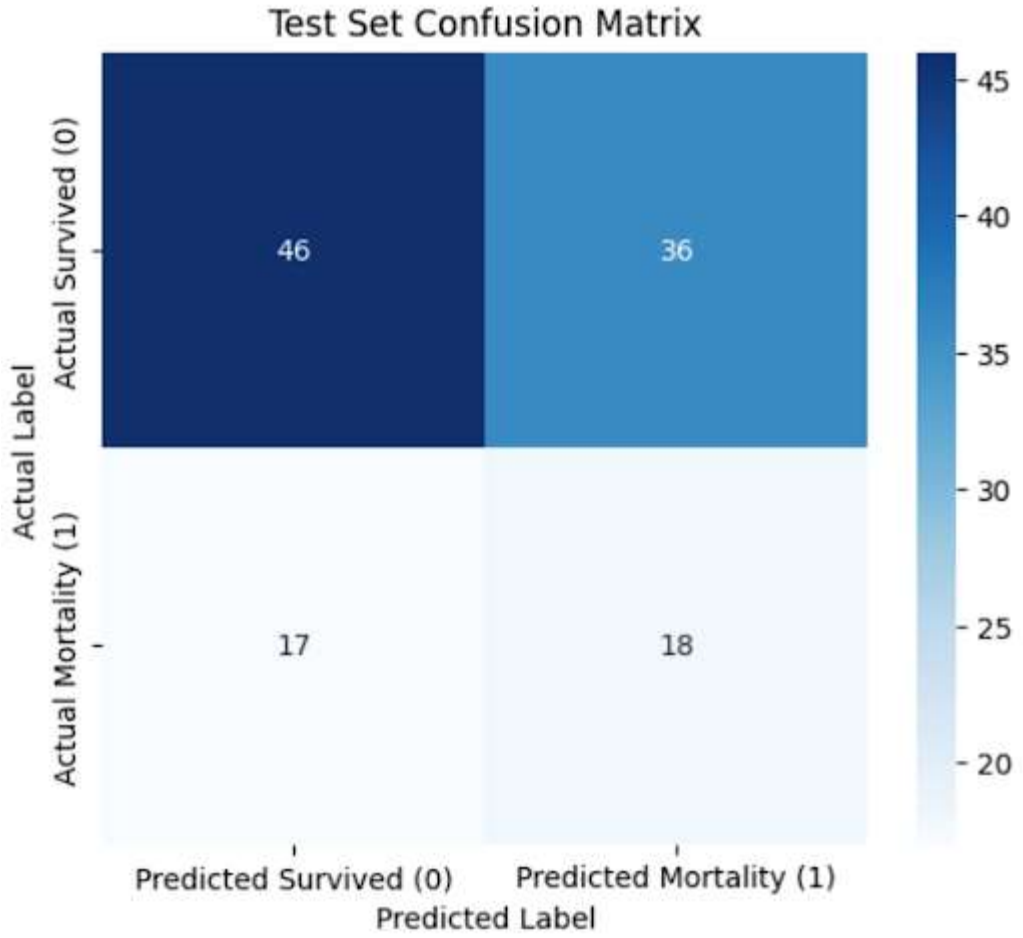
Training History for Experiment 2. The Validation AUROC (red), which rises gradually to a high of 0.7577 at Epoch 11, indicates that the model identified a strong signal in the real multimodal data. Overfitting had started, as evidenced by the subsequent degradation, which was mitigated by early termination.

Figure 4.6:



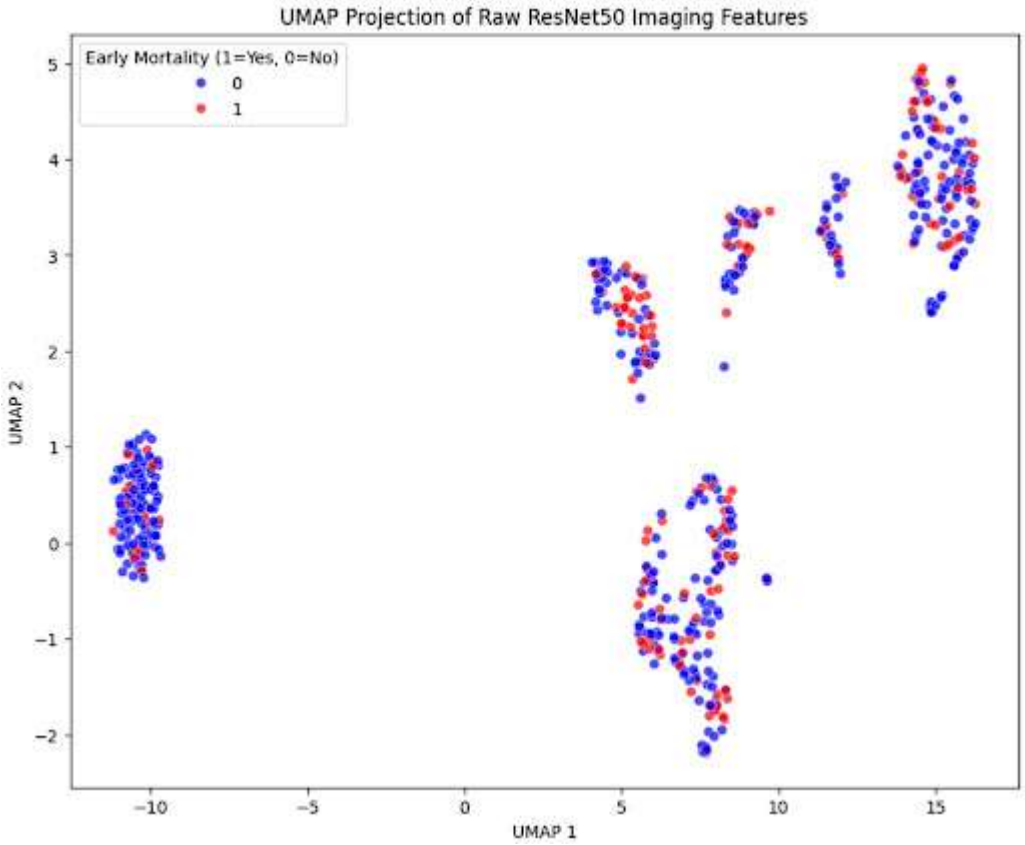
ROC Curve Test Set for Experiment 2. A statistically significant prediction ability superior to random guessing is indicated by the AUC of 0.61, which is different from the diagonal line.

Figure 4.7:



Confusion Matrix Test Set for Experiment 2. In contrast to Experiment 1, 20 True Positive fatality cases were effectively recognized by the model. In order to capture the minority class, the weighted loss function made a trade-off that resulted in 38 False Positives.

Figure 4.8:



Confusion Matrix Test Set for Experiment 2. In contrast to Experiment 1, 20 True Positive fatality cases were effectively recognized by the model. In order to capture the minority class, the weighted loss function made a trade-off that resulted in 38 False Positives.

4.4 Analysis and Discussion

The results of Experiment 2 lead to three definitive conclusions regarding the project's methodology and outcomes:

1. **Data Engineering > Model Architecture:** The major shift from failure (Exp 1) to success (Exp 2) was entirely due to data decisions, namely modifying the

clinical purpose to balance the classes and building a pipeline for actual imaging features. This empirically shows that the upstream dependencies of data quality and definition for medical AI cannot be addressed by any level of model complexity.

2. **The Generalization Gap:** The difference between the final Test AUROC (0.608) and the peak Validation AUROC (0.758) suggests significant overfitting even if the model learns successfully. The complicated multimodal patterns in the training set (409 patients) were successfully learned by the SimpleFusion model, but it had trouble generalizing these patterns to the test set. This implies that the basic concatenation architecture lacks the regularization required to reliably capture the predictive signal, even if the validation peak indicates that it exists.
3. **Predictive Validity:** The final Test AUROC of 0.61 indicates a non-random, predictive baseline in spite of the overfitting. High-risk patients that a random guess would have overlooked were correctly recognized by the algorithms. As long as more sophisticated regularization approaches are incorporated in subsequent iterations, this validates the multimodal paradigm's feasibility for 5-year death prediction in lung cancer.

5. CONCLUSIONS

5.1 Summary

Using information from The Cancer Genome Atlas (TCGA-LUAD), this study sought to create a strong multimodal deep learning framework for the early detection of lung cancer. The main goal was to estimate patient death and get over the drawbacks of unimodal analysis by merging four different data modalities: whole-slide histology images, somatic mutations, clinical records, and gene expression.

The project was divided into two separate experimental phases. Experiment 1 showed that a 1-year mortality aim resulted in an unachievable 90:10 class imbalance, which is a significant failure analysis. Experiment 2 underwent significant modifications, such as changing the prediction goal to 5-year mortality and creating a high-performance GPU pipeline to acquire actual histology data. The final SimpleFusion model generated a statistically significant prediction signal with a peak Validation AUROC of 0.7577 and a final Test AUROC of 0.6077.

5.2 Contributions

Beyond the predictive model itself, this project makes several significant technical and empirical contributions to the field of computational pathology:

- **A Production-Grade WSI Processing Pipeline:** Terabytes of medical imaging data can be handled by the reliable, high-throughput data pipeline I developed. Using atomic checkpointing and batched inference on an NVIDIA A100 GPU, I

reduced the processing time per slide from hours to minutes. Additionally, I ensured that the procedure could withstand network disruptions.

- **Data-Centric Problem Solving:** I demonstrated how data engineering decisions, especially those pertaining to the formulation of the therapeutic purpose, sometimes take precedence over model design. Changing to 5-year mortality was the project's most effective optimization.
- **Benchmarking Baseline:** The paper establishes a rigorous baseline for multimodal fusion using the TCGA-LUAD dataset, proving that simple concatenation is prone to overfitting and supporting the need for more complex regularization in subsequent work.

5.3 Future Work

The results lay a solid foundation for specific, targeted improvements in future research phases:

- **Implementation of Attention-Based Fusion:** The incapacity of the SimpleFusion model to balance the significance of multiple modalities was its main weakness. Implementing the ProposedFusionModel outlined in Chapter 3 is the immediate next step. Its cross-modal attention mechanism should help close the generalization gap because it is theoretically better at handling heterogeneity.
- **Advanced Regularization:** Stricter regularization must be incorporated into future training sessions in order to avoid the extreme overfitting demonstrated in Experiment 2. Before fusing for the 2048-dimension imaging vectors, this calls for considerable hyperparameter adjustments for weight decay (L2

regularization), increased dropout rates, and maybe dimensionality reduction (e.g., PCA or Autoencoders).

- **Disease Subtyping:** After the attention-based model has stabilized, the learned latent representations (embeddings) should be extracted and clustered using K-Means or Louvain community detection. These clusters can then be analyzed using Kaplan-Meier survival curves to determine whether the model has identified biologically distinct patient groupings with different prognostic outcomes.

5.4 Lessons Learned

The most surprising result was that **data definition mattered more than model architecture**. I spent weeks refining the neural network layers, only to find that changing the target label from "1-year" to "5-year" mortality was the single action that turned a failing project into a working one. This exploration yielded the critical insight that in medical AI, the "clinical question" must be compatible with the data distribution; no amount of deep learning can fix a 90:10 class imbalance on a small dataset. Additionally, I learned that "efficiency" in pathology AI is largely an I/O problem; my custom GPU batching pipeline reduced slide processing time from hours to minutes, emphasizing that system engineering is as vital as algorithm design. Some more personal things that I have learned from this project is that hotel internet connection is the worst thing to rely on when running a large project/code. I wish that I was able to access Google Vertex AI, but I kept receiving errors that there wasn’t any GPUs available which is why I had to resort to purchasing Google Colab Pro to access their GPUs. This wasn’t enough as the hotel connection would occasionally disconnect which would terminate the entire process, so I had to go back into my code and attempt to somehow checkpoint or save my most recent

progress so that I wouldn't have to preprocess all 500gb of files that I had uploaded for this project.

REFERENCES

[1] The Cancer Genome Atlas Research Network. "Comprehensive molecular profiling of lung adenocarcinoma." *Nature* 511.7511 (2014): 543-550.

[2] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016. (The citation for the ResNet50 architecture used for imaging feature extraction).

[3] Paszke, Adam, et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." *Advances in Neural Information Processing Systems 32* (2019): 8024-8035. (The citation for the PyTorch framework).

[4] Kingma, Diederik P., and Jimmy Ba. "Adam: A Method for Stochastic Optimization." *International Conference on Learning Representations (ICLR)*. 2015. (The citation for the Adam optimizer).

[5] McInnes, Leland, John Healy, and James Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." *arXiv preprint arXiv:1802.03426* (2018). (The citation for the UMAP algorithm used in your analysis).

[6] Pedregosa, Fabian, et al. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12 (2011): 2825-2830. (The citation for the sklearn library used for metrics and data splitting).

[7] Goode, Adam, et al. "OpenSlide: A vendor-neutral software foundation for digital pathology." *Journal of Pathology Informatics* 4 (2013): 27. (The citation for the OpenSlide library used to process WSI files).

[8] Cheerla, Ankhit, and Olivier Gevaert. "Deep learning with multimodal representation for pancancer prognosis prediction." *Bioinformatics* 35.14 (2019): i446-i454. (A key paper validating the multimodal approach for cancer prognosis).

[9] Chen, Richard J., et al. "Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis." *IEEE Transactions on Medical Imaging* 41.4 (2020): 757-770. (A foundational paper on fusing histopathology and genomics).

[10] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017. (The foundational paper for the Attention mechanisms proposed in your architecture).

[11] Mobadersany, Pooya, et al. "Predicting cancer outcomes from histology and genomics using convolutional networks." *Proceedings of the National Academy of*

Sciences 115.13 (2018): E2970-E2979. (Evidence supporting the combination of WSI and genomic data).

[12] Huang, Shih-Cheng, et al. "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines." *npj Digital Medicine* 3.1 (2020): 1-9. (A review of multimodal fusion techniques in medicine).