Advanced Biomarker Discovery and Tumor Microenvironment Analysis Through High-Dimensional Imaging and Radiogenomics

Naveen Balaji, Mariam Bastawros, Markella Bibidakis, Case Edmondson, Zijun Gao, Robert Valecka

***Abstract*— Single-modality approaches fail to adequately capture the full complexity of tumor staging and tumor microenvironment (TME) in non-small cell lung cancer (NSCLC). To address this, we developed a novel multi-modal machine learning pipeline to classify NSCLC stages and uncover biomarker insights. The model integrates CT images, RNA-sequencing, and electronic health records (EHR) through intermediate fusion. Each modality utilizes a separate neural network – 3D CNN for CT scans, a graphical neural network (GNN) for gene expression, and a multilayer perceptron (MLP) for EHR data – to output same-size feature vectors prior to concatenation for final classification. Interpretability was enhanced through application of SHAP scores for genomic information. Our model significantly outperformed the single-modality baseline with around 20% improvement in classification accuracy and better generalization on external validation sets, despite being applied to a smaller dataset. Importantly, genes with high SHAP scores are potentially new biomarkers, offering clinically relevant TME insight.**

I. Introduction

With over 226,650 estimated new cases in 2025, non-small cell lung cancer (NSCLC) is one of the leading contributors of lung cancers, constituting around 87% of lung cancer cases [1], [2]. While cases have been overall decreasing throughout the years, the overall prognosis remains poor with an overall 5-year relative survival rate of 28% [3]. NSCLC is a type of lung cancer which can affect either the left or right lung, any of the five lobes, or bronchi within the lungs [4]. Onset of the disease begins mildly with a common issue such as a persistent cough or trouble swallowing, but can quickly transform into more malicious symptoms such as coughing blood or loss of appetite [5]. Current diagnostic techniques include biopsies - a tissue sample taken to be viewed under a microscope by a pathologist - PET-CT scans - a method of differentiating cancerous cells from normal cells through glucose activity intake and imaging organs and tissues - or blood tests [4]. Further, there are very explicit and well-studied risk factors for NSCLC, including the primary factor of smoking, and secondhand smoking, exposure to chemicals such as asbestos and arsenic, radiation exposure, or HIV infection [4]. Smoking directly causes about 90% of lung cancer cases in men and 80% of cases in women [5]. This is due to the sheer number of chemicals in a single cigarette smoke. With over 5,000 chemicals being inhaled, damage to DNA in cells increases, while also preventing the processes for DNA repair and cell death, such as the p53 and Rb pathways [6].

Current treatments for NSCLC vary from surgery, radiation, chemotherapy, targeted therapy, and immunotherapy depending on the stage and substage of the cancer [4]. When the disease is detected earlier on, treatment pathways begin first with surgical removal of the tumor before secondary treatments such as chemotherapy, radiation, target therapy, or immunotherapy are implemented. Contrarily, if NSCLC is detected in a later stage, surgical treatment is often not considered as the cancer metastasized elsewhere, leaving the secondary options as the primary treatment [4]. Prognosis varies across patients due to what stage the cancer is first discovered [3]. When recognized in a localized state, the survival rate is nearly 65%, while in metastasized states, only 9%, leaving a combined value of 28% given the population averages [3].

Over the past few years, immunotherapy - using, modifying, or enhancing the patient's own immune system - has grown tremendously within the field of cancer treatments. Immunotherapy relies on the identification of immune cell types within the tumor microenvironment (TME). Typically, immunohistochemistry, flow cytometry, and genetic testing are used to assess components of the TME [7]. However, these technologies cannot reflect the high complexity of the TME due to its heterogeneous nature, limiting the progression of immunotherapeutics as a treatment for NSCLC [7]. The TME is complex, composed of multiple different types of non-cellular components like extracellular matrix, fibroblasts, and immune cells [7]. Through the autocrine-paracrine signaling pathway, the TME controls growth and spread of the tumor cells [8]. The area is known for its slight acidity, hypoxia, high reactive oxygen species, high osmotic pressure, and abnormal vasculature [8]. Further, the TME reduces secretion of immune-activating cytokines - these are ones that initiate attack on the tumor cells - and increases immunosuppressive cytokines - to keep the body's immune system from attacking [7]. At the same time, the TME will also hijack immune cells to "attack" any treatment targeted for it [7]. Improved identification of TME components could enhance immunotherapeutics application in clinical settings.

Therefore, improved technologies could identify new biomarkers for NSCLC diagnosis, subclassification, and TME components for immunotherapeutic treatments. Current

methodologies for NSCLC diagnosis and subclassification are limited. CT scans can capture the whole, heterogeneous tumor, but lack biochemical content information. Immunohistochemistry and genetic testing can provide biochemical information, but due to small sample size relative to the tumor, cannot capture the whole heterogeneous makeup of the TME. Thus, a new technology combining multiple modalities to capture biochemical content across a heterogeneous TME could improve NSCLC diagnosis/subclassification and identification of immunotherapeutic treatments.

Radiogenomics is a growing field in recent years that aims to integrate medical imaging modalities and genomic data. Traditionally, the field of radiomics has focused on extracting information from only medical imaging. Forghani et. al. had success determining prediction of metastasis using CT scans, while Wu et. al. achieved results which can better differentiate between two similar tumor subtypes [9]–[11]. However, there are many problems with this methodology. Isolating features from only CT scans without additional content can often mislead the viewer [9]. Radiogenomics has the potential to identify novel features within CT scans that elucidate TME molecular content and components through an integrated multi-modal machine learning analysis of radiomic (CT) information, genomic information, and electronic health record (EHR) information. Development of a model that can identify TME components from CT scans alone would have advantages over current technologies as the CT scan can capture the whole heterogeneous nature of the tumor, enabling clinicians to subclassify tumors and detect chemical content for cancer treatment decisions through CT imaging only. Further, without the full understanding, new insights into the TME are limited. However, multimodal integration faces its own issues. Prior work by Mahootiha, et al. in predicting survival time and prognosis through CT scan and clinical data in renal cell carcinoma showed limited applicability and efficiency [12]. Further, other multimodal methodologies have struggled with generalizability and limited data [13], [14].

In this study, we develop and implement a multimodal model combining EHR, CT, and RNA genomic information. Firstly, we develop feature vectors for each of the data types using multi-layer perceptron, 3D CNN, and graphical neural network respectively. We then concatenate the features leading towards a final classification. To ensure model interpretability, we applied SHAP scores on the genomic information to rank importance and GRAD-CAM on the CT scans to highlight spatial importance. This overall process contributes to efforts to develop and implement a potentially generalizable multimodal machine learning model with emphasis to applications in discovering novel biomarkers. Through the CT scans within the NSCLC dataset, the multimodal model identifies potential genetic biomarkers, tumor subclassification, and TME chemical makeup to inform clinicians of cancer prognosis, treatment options, and comprehensive analysis of immunotherapeutic options.

II. Literature Review

Multimodal modalities have been utilized in prior radiogenomics studies regarding a variety of cancer types. Early work in Yoon, Hong-Jun, et al. explores the integration of radiomic and genomic data using deep learning to predict clinical characteristics of invasive breast cancer [13]. Utilizing a custom neural network based on the CNN formation, 3D MRI volumes were processed and spatial structural features of the tumors were extracted. Genomic data was also incorporated into the network, providing a holistic insight into the tumor pathological stage and receptor statuses. However, several limitations were highlighted with necessity for improvement including: firstly, the need for a very large and diverse dataset to improve model generalizability, and secondly, the multimodal data input integration faced challenges in the training process due to a gene list selection constraint and an imbalance across the phenotypes.

Another study, published in 2020 and authored by Ning, Zhenyuan, et al., developed a model to extract 3D features from CT scan volumes using a 3D CNN to predict (International Society of Urological Pathology (ISUP) grades for tumors in CT images [14]. By leveraging feature extraction from combining both imaging modalities and eigengene-based genomic profiling, the study aimed to construct a more comprehensive prognostic model. However, there was distinctly poor comprehensive exploration done with the genomic data. The study utilized the traditional methodology of a Weighted Gene Co-expression Network Analysis (WGCNA). Unfortunately, this meant the final model did not incorporate deep learning for the genomic data, mis-matching what was done for the CT images, and limiting any conclusions reached.

Vallières et. al. [15] developed models to predict tumor outcomes in head-and-neck cancer based on radiomic and clinical variables. Radiomic variables came from CT scans of the tissue with the assumption of tumors having genomic heterogeneity. This heterogeneity leads to aggressive tumors having spatial biomarkers which can be analyzed to calculate the tumor's risk. Machine learning allows for the integration of clinical attributes with radiomics data to increase prediction accuracy. Tumor shape, intensity, and texture are quantified as features from images and then combined with clinical information in random forests to create a prediction model. Not surprisingly, the same limitations regarding generalizability and multimodal data integration are also highlighted, emphasizing that these challenges are common across radiogenomics research and need to be systematically addressed.

Finally, another study, published in 2024 and written by Pai et al. explored the use of a self-supervised foundation model for cancer imaging biomarker discovery [16]. Through contrastive learning, the model extracts deep imaging features from NSCLC CT scans without requiring manual labels and is transferable across diagnostic and prognostic tasks. This approach enables the model to learn generalizable representations from unannotated data, improving its performance in low-data situations. However, while the model is able to reach conclusions from CT imaging data, it does not incorporate clinical variables and its interpretability is limited, leaving it as a "black box". This potentially explains the AUC score of 0.638, and emphasizes the need for multimodal methodologies.

With the necessity and challenges of multimodal methodologies in mind, further research was conducted to consider

how multimodal data architectures and integration have been approached in existing literature, specifically for biomedical and oncological datasets. Recent studies have focused on improving the integration of diverse datasets through intermediate fusion. The paper J. Lipkova et al. explores early, late, and intermediate fusion strategies to enhance the interpretability and effectiveness of deep learning models in oncology [17]. It highlights how intermediate fusion outcompetes other integration methods when modalities are very distinct as it identifies cross-correlations between similar features, and combines less correlated data in later layers. By leveraging feature alignment techniques at intermediate network layers, improved predictive power is seen when integrating radiomic and genomic data. Similarly, V. Guarrasi et al. provides a comprehensive analysis of intermediate fusion methodologies, highlighting how these approaches mitigate feature redundancy and improve generalizability compared to early or late fusion strategies [18]. These insights offer promising directions for future multimodal deep learning applications in oncology and other biomedical domains.

III. METHODOLOGY

A. Data Description

1) Dataset: The NSCLC (Non-Small Cell Lung Cancer) Radiogenomics dataset is designed for studying genomic biomarkers in combination with quantitative radiomics to advance development of personalized medicine and medical imaging decision-making [19]. The dataset includes 211 subjects from two cohorts with a total of 286,754 CT and PET/CT images with annotated segmentations in DICOM format. The subjects were NSCLC patients entered into the study from 2008 through 2012 at the Stanford University School of Medicine and the Palo Alto Veterans Affairs Healthcare system. All of the subjects have entries in an electronic health record (EHR) with clinical data, tumor diagnoses, and survival times. 117 of the subjects have RNA sequencing data recorded in a CSV in TPM (transcripts per million) format. A subset of 89 subjects from this dataset will be used for this project's multimodality model since the subjects will have CT images, RNA sequencing, EHR data, and the clinical outcomes used as the ground truth.

2) Pre-processing: Since one of our modalities is CT imaging, we performed image pre-processing on all CT scans, including approximately 13,000 images from 89 patients. Due to the use of different CT machines and inconsistent scanning protocols, there was considerable variation in noise characteristics and image contrast across the datasets. As a result, traditional physics-based denoising algorithms, such as wavelet-based denoising and non-local means filtering, struggle to generalize effectively across all noise types present in the data. To avoid the need for manual parameter tuning on each image to achieve optimal denoising performance, we employed a pre-trained image restoration model. This model has demonstrated effectiveness across various noise types, including low signal-to-noise ratio (SNR) conditions and undersampled data. Specifically, we adapted the image restoration model from ***Cellpose3***, which was originally designed for cell-level microscopy images but has also shown promising performance on CT scans [20]. As shown in **Fig 2**, the noisy reference image suffers from a low SNR, which obscures most vessel details. Using various restoration types available in the model, we were able to recover much of the lost structure, with the ***deblur_cyto3*** mode producing the most effective denoising results.

We observed that the default parameters of the ***Cellpose3*** model were sufficient to address the noise characteristics present in our dataset, particularly in low-contrast and low-SNR CT images. To illustrate the model's effectiveness, we present a representative example from our training dataset in **Fig 3**. The left panels show the original CT slices, where the image suffers from low contrast and most fine vascular structures are barely visible. After applying the denoising model, the image quality was markedly improved, with enhanced contrast and clearer visualization of small blood vessels.

B. Baseline Method For Comparison

1) Motivation and Scope: In this project, we aim to explore the use of multimodal deep learning to predict the overall clinical stage of Non-Small Cell Lung Cancer (NSCLC) patients. For the baseline model, the core idea was to combine clinical data with imaging-derived features to improve prediction accuracy, leveraging the full 211 patient dataset via the Foundation Model for Cancer Imaging Biomarkers (fmcib) library [21] used by Pai, et al. Utilizing precomputed "foundation features" provided by fmcib, alongside clinical variables, was ideal as using the raw 3D CT scans was inadequate.

2) FMCIB: The fmcib library [21] facilitated easy access to the NSCLC-Radiogenomics dataset dataset and provided precomputed features and a pre-trained 3D ResNet50 model. However, the "black-box" nature of the foundation features, with limited documentation on their derivation, posed challenges. Relying on precomputed features also limited the ability to apply end-to-end learning or data augmentation techniques.

Before introducing the proposed multimodal fusion strategy, we establish a *strong yet simple* baseline that every subsequent model must beat. The idea is straightforward: concatenate (i) routinely collected clinical covariates and (ii) the 4 096-dimensional "*foundation features*" released with the `fmcib` toolkit and pass the resulting tabular vector through a small fully connected network. As all computation is performed on fixed-length vectors, the baseline is fast, reproducible, and agnostic to GPU memory limits, making it a natural starting point for graduate-level ablation studies.

3) Input Representation:

- **Clinical block** ($d_c = 18$): age, sex, smoking status, TNM substages, histology, and ECOG performance score are encoded as appropriate. one hot or z score.
- **Imaging block** ($d_i = 4\,096$): foundation features extracted from the gross-tumour volume on the planning CT using the 3-D ResNet-50 encoder released by `fmcib`.
- **Concatenation**: the two blocks are concatenated to form a $d = d_c + d_i = 4\,114$-dimensional vector, $\mathbf{x} \in \mathbb{R}^d$, which is the sole input to the MLP.

No.	Paper	What research question(s) are being addressed?	Approach	When	Drawback	Why
1	*Deep radiogenomics for predicting clinical phenotypes in invasive breast cancer.*	Early work on combining medical imaging data and genomics data to predict the tumor pathological stage.	A simple 3D CNN network to process the MRI imaging and genomics data at the same time.	2018	Moderate applicability due to the relatively small sample size. Challenge of integrating multimodal data effectively.	It needs large, diverse datasets to improve model generalizability. Insufficient multimodal datasets.
2	*Integrative analysis of cross-modal features for the prognosis prediction of clear cell renal cell carcinoma.*	Utilize deep learning for prognosis based on different data modalities.	CT & histopathological imaging combined along with genomics data. Using 3D CNN network.	2020	Challenge of integrating multimodal data effectively.	Limited Generalizability. Insufficient multimodal datasets.
3	*Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer.*	How radiomics can be used to quantify heterogeneity between tumors while combining clinical data to make more informed tumor severity predictions.	Creating radiomics-only models from extracted features to be compared with models that include radiomics and clinical features.	2017	Prediction powers of adding clinical data to models are different.	Need for more standardized radiomics analysis. Distant metastases is dominated by radiomics features much more than locoregional occurrences.
4	*Foundation Model for Cancer Imaging Biomarkers.*	Can the self-supervised model, when trained on unannotated cancer CT scans, learn generalizable imaging representations to improve performance of imaging biomarkers?	Used 3D convolutional encoder pretrained with a contrastive learning variant of SimCLR	2024	Does not incorporate clinical information; Limited interpretability	The model remains a "black box" as attribution methods (smooth guided backpropagation) may not fully explain model decisions. ; Focus was primarily on radiomic data only.
5	*Artificial intelligence for multimodal data integration in oncology.*	What are the current AI methods for integrating diverse data modalities in oncology and what challenges exist in the integration of these clinically-relevant modalities?	Comprehensive overview of AI techniques used in multimodal data (imaging, genomics, EHR) fusion in oncology, highlighting the options for training, fusion, and interpretability.	2022	No novel findings or methods are discussed.	The paper is meant to be a summary of existing AI techniques and fusion methods.
6	*A Systematic Review of Intermediate Fusion in Multimodal Deep Learning for Biomedical Applications.*	What intermediate fusion techniques are currently used in biomedical application deep learning models?	Organized summary of current intermediate fusion methods in biomedical applications, when to use them, and challenges faced.	2024	No novel findings or methods are discussed.	The paper is meant to be a summary of existing AI techniques and fusion methods.

Table 1. Summary of the Literature Review with Critiques and Strengths

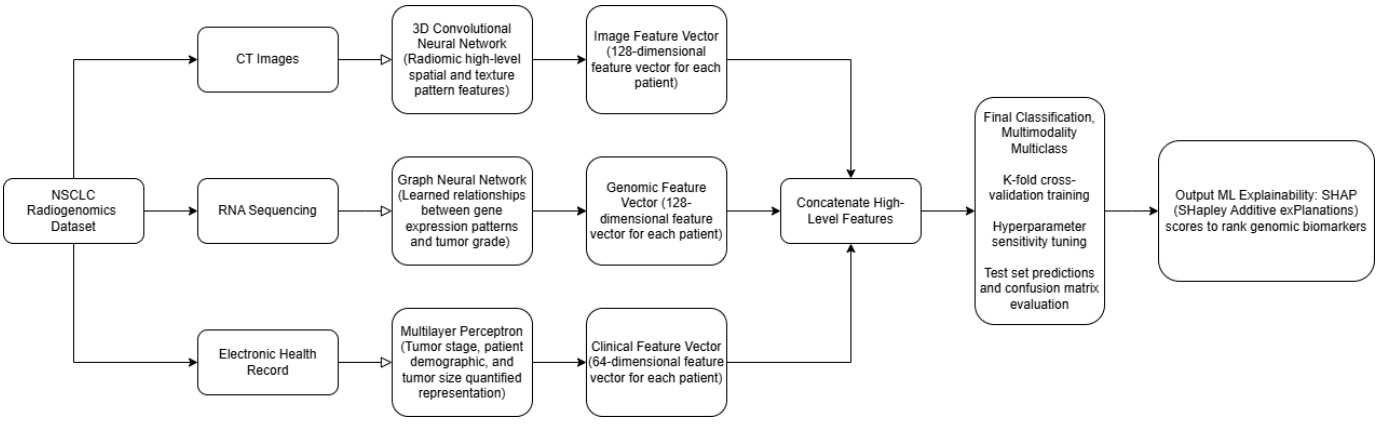


Fig. 1. Overview of proposed multimodal machine learning pipeline, using intermediate fusion concatenation between a 3D CNN, a GNN, and a multilayer perceptron.

4) Metrics: To compare the baseline and the multimodal models we reported three class-balanced metrics and visualized their evolution during training. Accuracy is the fraction of correct predictions across all classes: $\frac{TP+TN}{N}$. Although intuitive, it can overstate performance when the stage distribution is skewed, as with the full dataset. Macro-F1 therefore

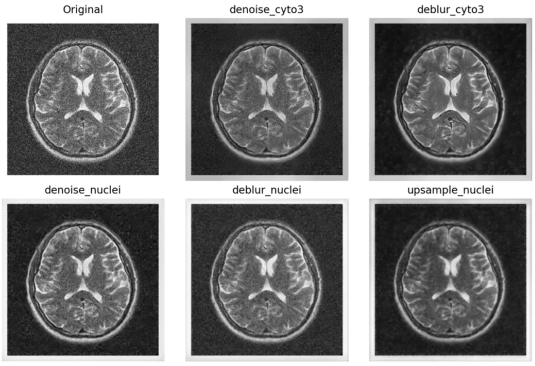


Fig. 2. Comparison of image restoration results using different pre-trained modes from ***Cellpose3***. The original noisy CT slice (top left) shows low signal-to-noise ratio and blurred vessel structures. Various restoration modes were applied: ***denoise_cyto3***, ***deblur_cyto3***, ***denoise_nuclei***, ***deblur_nuclei***, and ***upsample_nuclei***. Among them, ***deblur_cyto3*** yielded the most effective enhancement of structural details while suppressing background noise.

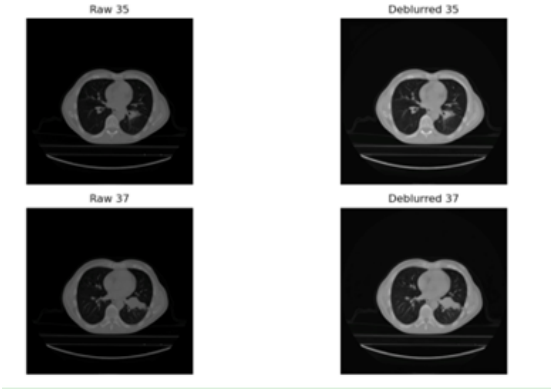


Fig. 3. Example slices from a CT scan before and after applying the ***Cellpose3 deblur_cyto3*** model. Raw images from slices 35 and 37 (left panels) show reduced contrast and blurred anatomical structures. After applying the deblurring model (right panels), both contrast and structural clarity are significantly improved, enhancing the visibility of lung features.

receives equal weight from each class. For every stage k we compute $Precision_k$ and $Recall_k$, derive $F1_k$:

$$F1_k = \frac{2 \cdot Precision_k \cdot Recall_k}{Precision_k + Recall_k}$$

and average the $F1_k$ values over the K stages:

$$\text{Macro-F1} = \frac{1}{K} \sum_{k=1}^{K} F1_k$$

This penalizes the tendency of the baseline to ignore late stage cases.

The third metric, Macro-AUC, extends the binary receiver operating characteristic area to multiclass problems by adopting a one-vs.-rest scheme and then averaging the per-class AUCs. As AUC is threshold-independent, it highlights the quality of the probability calibration produced by our cross-modal attention block. We supplemented tabular scores with learning-curves (loss and Macro-F1) and class-wise ROC plots; both confirmed that early stopping at fifteen epochs prevented over-fitting while a cosine-decay learning-rate schedule ($\eta_0 = 3 \times 10^{-4}$) steadily improved discrimination.

Hyper-parameters were selected through Bayesian optimization on a five-fold stratified split, maximizing Macro-F1. Regularization (weight-decay 10^{-5}, dropout 0.3) and class-balanced focal loss ($\gamma = 2.0$) further stabilized gradients under severe class imbalance. Collectively, these metric choices and tuning protocols yielded an evaluation suite that is robust to label skew, and sensitive to the clinical cost of mis-staging.

5) Network Topology:

$$\begin{aligned} \mathbf{h}_1 &= \mathrm{ReLU}\big(\mathrm{BN}\big(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1\big)\big), && (d \to 512) \\ \mathbf{h}_2 &= \mathrm{ReLU}\big(\mathrm{BN}\big(\mathbf{W}_2\mathbf{h}_1 + \mathbf{b}_2\big)\big), && (512 \to 128) \\ \hat{\mathbf{y}} &= \mathrm{Softmax}\big(\mathbf{W}_3\mathbf{h}_2 + \mathbf{b}_3\big), && (128 \to 4 \text{ classes}), \end{aligned}$$

where every hidden layer is followed by **Dropout** ($p = 0.3$). The network contains ≈ 2.2 M trainable parameters two orders of magnitude fewer than most 3-D CNNs, this ensures rapid convergence.

6) Baseline Training Protocol:

1) **Dataset.** We use the *LUNG1* cohort $N = 211$ patients, stage I–IIIB as the sole training
2) **Optimiser.** Adam ($\eta = 10^{-3}$, $\beta_{1,2} = 0.9, 0.999$) with *weight decay* 10^{-4}.
3) **Loss.** Class-weighted cross-entropy to compensate for class imbalance (stage IIIB accounts for only 7 %).
4) **Validation.** Stratified 5-fold cross-validation; early stopping with a patience of 15 epochs on macro-averaged F_1.

TABLE I
BASELINE MLP PERFORMANCE (5-FOLD CROSS-VALIDATION)

Metric	Accuracy	Macro F_1	Macro AUC
Mean $\pm$ SD	0.46 ± 0.06	0.43 ± 0.07	0.64 ± 0.04

7) Performance Snapshot: The baseline comfortably exceeds naive majority class guessing (Accuracy $= 0.46$), but leaves considerable headroom for more expressive fusion strategies. Importantly, it reproduces the overfitting-to-generalization gap reported in earlier radiomic studies, making it a *realistic* standard against which to evaluate our proposed multimodal methodologies.

8) Limitations: The MLP operates on frozen features. Therefore, it *cannot* adapt its representation to idiosyncrasies such as imaging kernel or voxel spacing. Moreover, global concatenation ignores potential interactions between modalities, which is an issue our main, multimodal approach tackles via cross-modal attention. It is necessary to include gene expressions and graphical neural networks to better understand these biomarkers.

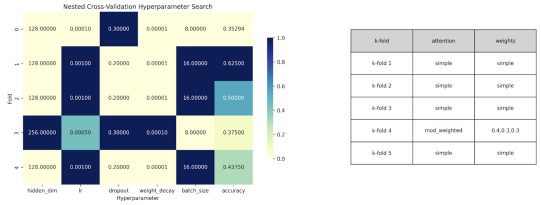
C. Multi-Modal Approach

For our model, we propose a comprehensive computational pipeline to analyze high-dimensional imaging data from non-small cell lung cancer (NSCLC) patients. Our approach integrates CT imaging, intermediate fusion machine learning, RNA-sequencing, and EHR analysis to discover prognostic biomarkers, characterize the tumor microenvironment (TME), and assess uncertainty in findings. Our pipeline follows a

multi-modal radiogenomics workflow, converting raw image data into quantitative features and predictive models. Key steps include: (a) Data acquisition and preprocessing, (b) Feature extraction at each modality, (d) Feature selection and integration with intermediate fusion concatenation of one-dimensional feature vectors, and (e) Modeling and validation. This structured approach mirrors established radiogenomics pipelines but is tailored to high-dimensional CT imaging of the TME combined with RNA-sequencing genomic data.

A diagram of the multimodal diagram is shown in Fig 1. Each data type from the NSCLC dataset is processed through a separate neural network to obtain a one-dimensional feature vector. The feature vectors of the different modalities are concatenated together and then passed through a final classification layer to find the connections between high dimensional features across different modalities. Attention-mapping and SHAP (SHapely Additive exPlanations) scores are used to explain the importance of different modality features in classifying tumors and increase explainability of the model for end users, such as clinicians and public health professionals.

IV. Experimental Results



k-fold	attention	weights
k-fold 1	simple	simple
k-fold 2	simple	simple
k-fold 3	simple	simple
k-fold 4	mod_weighted	0.4,0.3,0.3
k-fold 5	simple	simple

Fig. 4. Hyperparameter Heatmap.

Figure 4 shows a heat map of hyperparameters used during cross-validation training of the multimodal classifier. Hyperparameters are defined and passed into the nested cross-validation function which uses the different hyperparameters for each k-fold and records the best hyperparameters for the model.

The numerical hyperparameters include hidden dimensions, learning rate, dropout, weight decay, and batch size. Hidden dimensions control the size of the hidden layers in the multimodal fusion classification function; more dimensions allow more complex patterns to be learned but can lead to overfitting. The learning rate determines the step size at each epoch while moving toward a minimum of the loss function during training; a higher learning rate can increase the training speed but may overshoot minima. Meanwhile, a lower learning rate improves precision but can slow convergence. Dropout randomly disables a fraction of neurons during training in the multimodal fusion classification function to prevent overfitting and promote generalization. Weight decay adds a penalty proportional to the magnitude of the model weights to the loss function, which discourages complex models that may lead to overfitting. Batch size is the number of samples processed before internal parameter updates; smaller batch sizes are more frequent but noisier, while larger batch sizes have smoother updates with more memory and longer training times.

The categorical hyperparameters include the attention type and, if the attention type is modality weighted, the percentage weights for each model. The attention type parameter is "simple" or "mod_weighted", if it is simple then there is equal contribution from each modality for fusion classification and if it is mod_weighted then there are different percentage weights applied to each modality. The modality weights parameter has one weight for each of the three modalities, which all sum up to 1.0. The hyperparameters and subsequent accuracy are plotted for each fold. The heatmap color scale is normalized to the lowest and highest value of a variable across folds.

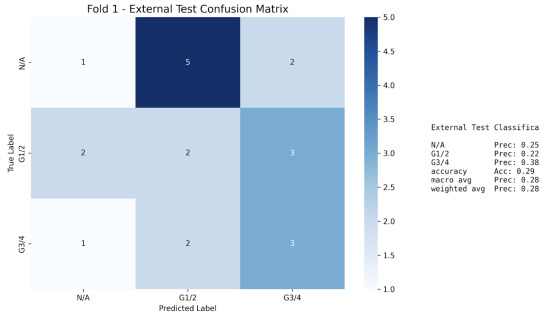


Fig. 5. Fold 1 Confusion Matrix.

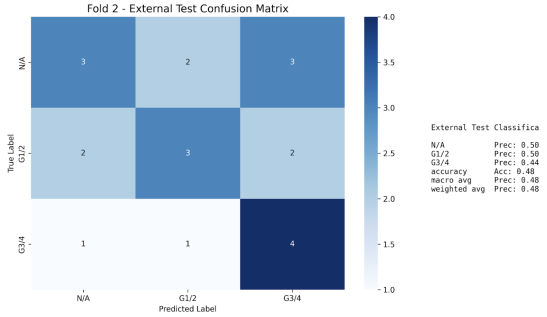


Fig. 6. Fold 2 Confusion Matrix.

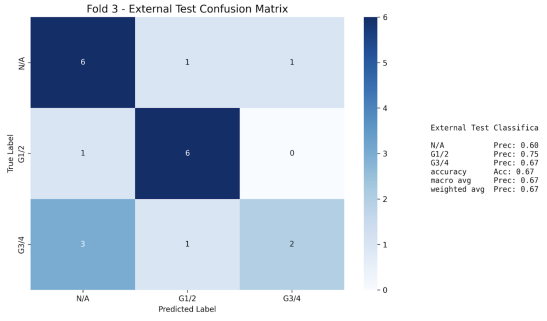


Fig. 7. Fold 3 Confusion Matrix.

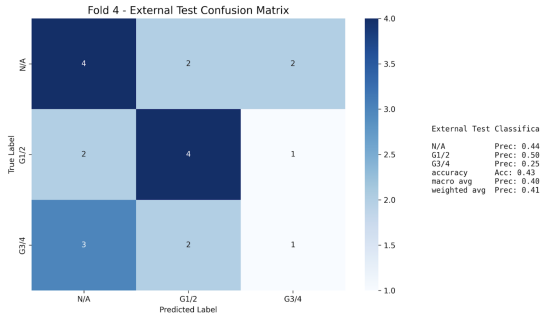


Fig. 8. Fold 4 Confusion Matrix.

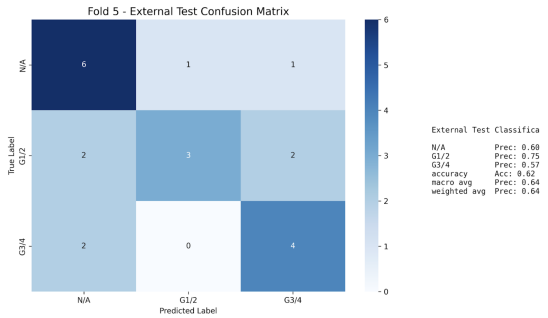


Fig. 9. Fold 5 Confusion Matrix.

Figures 5 through 9 show confusion matrices produced by evaluating the five fold models from cross-validation with the external validation test dataset.

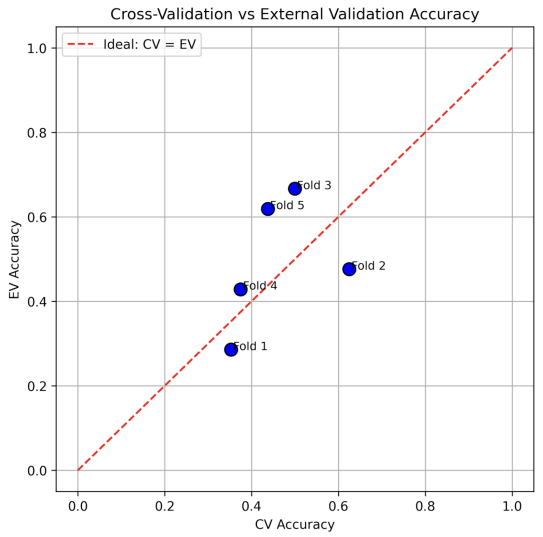


Fig. 10. Cross-Validation vs External Validation Accuracy.

In Figure 10, the CV and EV performances are plotted against each other for each fold. For points on the diagonal, generalization from training to the external test is good. For points below the diagonal, there is overfitting where the CV dominates the EV. For points above the diagonal, there is underfitting or the fold is receiving an unknown external boost.

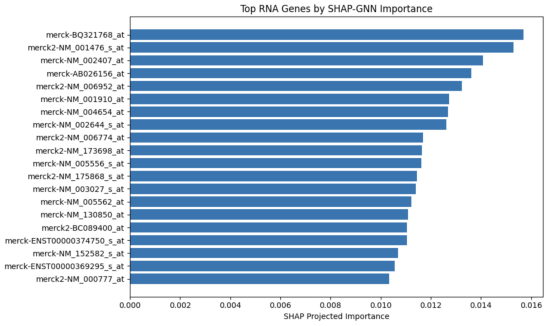


Fig. 11. SHAPley Gene Scores.

To contribute to model explainability, SHAP scores seen in Figure 11 are used to quantify the contribution of genes as features in the classification of patient tumor grade. In our model, the multimodal classifier model is loaded with the CT and EHR weights held constant at average vectors so that the SHAP scores can rank how the varying RNA contributes to multimodal classification.

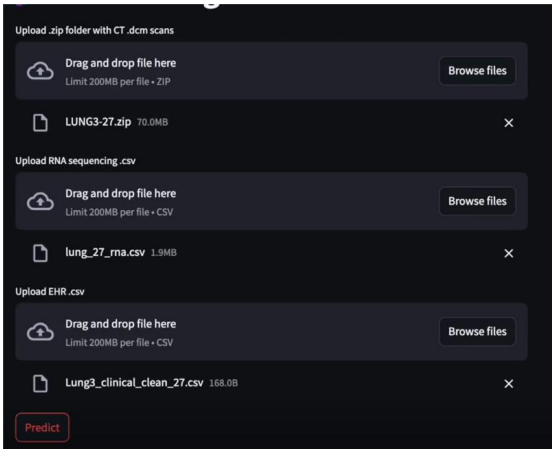


Fig. 12. GUI Mock-up.

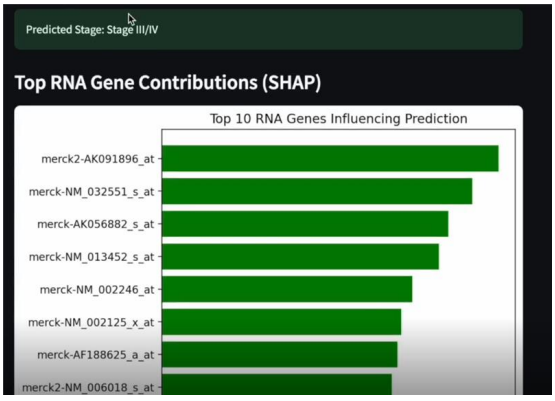


Fig. 13. GUI Predictions.

Future users of the model can use a GUI to easily make tumor classification predictions for patients and to produce a ranking of the most impactful genes. For a single patient the user will need to to upload a ZIP of the patient's CT image DICOM files, a CSV of the patient's RNA sequencing data,

and a CSV of the patient's EHR data. The GUI will output a tumor classification prediction and the top ranked genes by SHAP score.

V. Discussion

The baseline model was straightforward. It is an MLP trained on 4096 frozen CT "foundation features" along with the clinical covariates. However, it's performance was poor, with a mean cross-validation (CV) accuracy around 0.46 and a macro-F1 of 0.43. When testing on external folds, there was also a wide gap, suggesting generalization issues. Further, as shown with the confusion matrices, the model heavily favored predicting early-stage cancer. This basically confirms the hypothesis that radiomic signals alone do not adequately capture the molecular-level differences that drive disease progression, especially in advanced stages. In particular, without the RNA-seq data, the model was completely blind to the immune and metabolic factors that differentiate stage IIIB patients from the lower grades.

Our intermediate-fusion approach fixes that blind spot. We set it up to learn complementary information from different sources: first the CT slices using a 3D CNN, then EHR variables with an MLP, and the whole-transcriptome TPM vectors using a graph attention network. Only after this initial learning did we apply cross-modal attention. This intermediate multimodal setup increased the average CV accuracy by roughly twenty percentage points and pushed the macro-F1 score into the low-60% range. Crucially, the external validation performance stayed within 3% of the training results. This is good evidence that incorporating the RNA signatures helps with both discriminating between stages and making the model generalize better. We also ran a SHAP analysis, highlighting the genes which received high importance scores, map to pathways involved in hypoxia response and T-cell exhaustion, which makes biological sense and supports our findings.

It is worth noting we got these improvements even though our training dataset was much smaller (only 89 patients had the complete CT + RNA + EHR data) compared to the 211 patients used for the baseline. This small cohort size really highlighted the challenges of missing data. The initial LUNG1 data we looked at did not have the clinical labels sheet or the AIM segmentation masks, so we had to switch over to the NSCLC-Radiogenomics dataset. Even there, about half the cancer stage samples were missing, and we had to deal with CT intensity drift using some aggressive histogram harmonization. To handle all this sparsity, we employed a few tricks: (i) for EHR gaps, we used median imputation plus masking tokens; (ii) for the transcripts, we preprocessed them to keep only the top 2000 variance-stabilized genes; and (iii) we used heavy data augmentation on the imaging side. These steps together seemed to prevent the model from over-regularizing while still letting it exploit the synergies between the different data modalities.

The model is available as a GUI for ease of use. A practicioner can upload a .zip file with CT scans as .dcm files, a .csv file with RNA sequencing data, and another .csv file with EHR clinical data associated with a single patient. The GUI will run the model on the provided data, and return a predicted cancer stage classification along with the top 10 RNA genes present in the patient used to come to that prediction and their associated SHAP values.

Succinctly, the baseline's poor performance was not due to a basic learning algorithm, but more so because of an information bottleneck. Radiomics on its own cannot provide the full picture of the tumor micro-environment. Our multimodal design brings back those critical RNA signals, reduces the difference between training and test performance, and demonstrates that even when dealing with small, noisy datasets, a well-thought-out fusion strategy can pull out clinically relevant signals that a radiomics-only approach would likely miss.

VI. Conclusion

Multimodal classification with different data types for medical analysis is still an emerging field with a high degree of potential. Combining the features learned from imaging, genomic, and clinical data can lead to potential diagnoses that cannot be made otherwise. The most important part of any machine learning model is data quality which includes normalizing data so that the neural networks can decipher differences between patients along feature vectors. In the future, the model can be possibly improved by using image segmentation and gray-level slicing on tumors to better isolate the tumors from the rest of the scan image. This would lead to features learned from the CT data to be more specific to the tumor itself. The model can also be improved by tuning hyperparameters as seen in the cross-validation and external validation where different hyperparameters used for the different k-folds could have a significant impact on model accuracy in CV and EV.

Contributions

Team Member	Contribution
Naveen Balaji	Project baselines, methodology, discussions
Mariam Bastawros	Abstract, Introduction, Literature Review, General Editing
Markella Bibidakis	Literature Review, Project Methodology, Workflow
Case Edmondson	Abstract, Introduction, Literature Review, Validation, GUI
Zijun Gao	Literature Review, Methodology, Pre-processing
Robert Valecka	Model Design, Programming, Results, Discussions

Table 2. Contributions of Team Members

References

[1] "Non-Small Cell Lung Cancer Treatment (PDQ®) - NCI," Apr. 2025. [Online]. Available: https://www.cancer.gov/types/lung/hp/non-small-cell-lung-treatment-pdq

[2] "Lung Cancer Statistics | How Common is Lung Cancer?" [Online]. Available: https://www.cancer.org/cancer/types/lung-cancer/about/key-statistics.html

[3] "Lung Cancer Survival Rates | 5-Year Survival Rates for Lung Cancer." [Online]. Available: https://www.cancer.org/cancer/types/lung-cancer/detection-diagnosis-staging/survival-rates.html

[4] "Non-Small Cell Lung Cancer Treatment - NCI," Mar. 2025. [Online]. Available: https://www.cancer.gov/types/lung/patient/non-small-cell-lung-treatment-pdq

[5] "Lung Cancer." [Online]. Available: https://medlineplus.gov/lungcancer.html

[6] "How does smoking cause cancer?" Dec. 2018. [Online]. Available: https://www.cancerresearchuk.org/about-cancer/causes-of-cancer/smoking-and-cancer/how-does-smoking-cause-cancer

[7] L. Wang, Q. Jia, Q. Chu, and B. Zhu, "Targeting tumor microenvironment for non-small cell lung cancer immunotherapy," *Chinese Medical Journal Pulmonary and Critical Care Medicine*, vol. 1, no. 1, pp. 18–29, Mar. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772558823000014

[8] C. Liu, M. Wang, H. Zhang, C. Li, T. Zhang, H. Liu, S. Zhu, and J. Chen, "Tumor microenvironment and immunotherapy of oral cancer," *European Journal of Medical Research*, vol. 27, no. 1, p. 198, Oct. 2022. [Online]. Available: https://eurjmedres.biomedcentral.com/articles/10.1186/s40001-022-00835-4

[9] R. Forghani, P. Savadjiev, A. Chatterjee, N. Muthukrishnan, C. Reinhold, and B. Forghani, "Radiomics and Artificial Intelligence for Biomarker and Prediction Model Development in Oncology," *Computational and Structural Biotechnology Journal*, vol. 17, pp. 995–1008, 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2001037019301382

[10] R. Forghani, A. Chatterjee, C. Reinhold, A. Pérez-Lara, G. Romero-Sanchez, Y. Ueno, M. Bayat, J. W. M. Alexander, L. Kadi, J. Chankowsky, J. Seuntjens, and B. Forghani, "Head and neck squamous cell carcinoma: prediction of cervical lymph node metastasis by dual-energy CT texture analysis with machine learning," *European Radiology*, vol. 29, no. 11, pp. 6172–6181, Nov. 2019. [Online]. Available: http://link.springer.com/10.1007/s00330-019-06159-y

[11] W. Wu, C. Parmar, P. Grossmann, J. Quackenbush, P. Lambin, J. Bussink, R. Mak, and H. J. W. L. Aerts, "Exploratory Study to Identify Radiomics Classifiers for Lung Cancer Histology," *Frontiers in Oncology*, vol. 6, Mar. 2016. [Online]. Available: http://journal.frontiersin.org/Article/10.3389/fonc.2016.00071/abstract

[12] M. Mahootiha, H. A. Qadir, J. Bergsland, and I. Balasingham, "Multimodal deep learning for personalized renal cell carcinoma prognosis: Integrating CT imaging and clinical data," *Computer Methods and Programs in Biomedicine*, vol. 244, p. 107978, Feb. 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260723006442

[13] H.-J. Yoon, A. Ramanathan, F. Alamudun, and G. Tourassi, "Deep radiogenomics for predicting clinical phenotypes in invasive breast cancer," in *14th International Workshop on Breast Imaging (IWBI 2018)*, E. A. Krupinski, Ed. Atlanta, United States: SPIE, Jul. 2018, p. 75. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10718/2318508/Deep-radiogenomics-for-predicting-clinical-phenotypes-in-invasive-breast-cancer/10.1117/12.2318508.full

[14] Z. Ning, W. Pan, Y. Chen, Q. Xiao, X. Zhang, J. Luo, J. Wang, and Y. Zhang, "Integrative analysis of cross-modal features for the prognosis prediction of clear cell renal cell carcinoma," *Bioinformatics (Oxford, England)*, vol. 36, no. 9, pp. 2888–2895, May 2020.

[15] M. Vallières, E. Kay-Rivest, L. J. Perrin, X. Liem, C. Furstoss, H. J. W. L. Aerts, N. Khaouam, P. F. Nguyen-Tan, C.-S. Wang, K. Sultanem, J. Seuntjens, and I. El Naqa, "Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer," *Scientific Reports*, vol. 7, no. 1, p. 10117, Aug. 2017. [Online]. Available: https://www.nature.com/articles/s41598-017-10371-5

[16] S. Pai, D. Bontempi, I. Hadzic, V. Prudente, M. Sokač, T. L. Chaunzwa, S. Bernatz, A. Hosny, R. H. Mak, N. J. Birkbak, and H. J. W. L. Aerts, "Foundation model for cancer imaging biomarkers," *Nature Machine Intelligence*, vol. 6, no. 3, pp. 354–367, Mar. 2024. [Online]. Available: https://www.nature.com/articles/s42256-024-00807-9

[17] J. Lipkova, R. J. Chen, B. Chen, M. Y. Lu, M. Barbieri, D. Shao, A. J. Vaidya, C. Chen, L. Zhuang, D. F. Williamson, M. Shaban, T. Y. Chen, and F. Mahmood, "Artificial intelligence for multimodal data integration in oncology," *Cancer Cell*, vol. 40, no. 10, pp. 1095–1110, Oct. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S153561082200441X

[18] V. Guarrasi, F. Aksu, C. M. Caruso, F. Di Feola, A. Rofena, F. Ruffini, and P. Soda, "A systematic review of intermediate fusion in multimodal deep learning for biomedical applications," *Image and Vision Computing*, vol. 158, p. 105509, May 2025. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0262885625000976

[19] S. Bakr, O. Gevaert, S. Echegaray, K. Ayers, M. Zhou, M. Shafiq, H. Zheng, J. A. Benson, W. Zhang, A. N. C. Leung, M. Kadoch, C. D. Hoang, J. Shrager, A. Quon, D. L. Rubin, S. K. Plevritis, and S. Napel, "A radiogenomic dataset of non-small cell lung cancer," *Scientific Data*, vol. 5, no. 1, p. 180202, Oct. 2018. [Online]. Available: https://www.nature.com/articles/sdata2018202

[20] C. Stringer and M. Pachitariu, "Cellpose3: one-click image restoration for improved cellular segmentation," *Nature Methods*, vol. 22, no. 3, pp. 592–599, Mar. 2025. [Online]. Available: https://www.nature.com/articles/s41592-025-02595-5

[21] S. Pai and I. Hadzic, "AIM-Harvard/foundation-cancer-image-biomarker: v0.0.1," Jan. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.10535536