Multi-Omics Predictive Model for Cancer Stage Classification

Angela Xu, Pratibha Pradeep, Saathvik Chandupatla

Abstract

Breast cancer remains one of the most prevalent and deadly forms of cancer worldwide, significantly impacting public health and patient outcomes. Accurate tumor stage classification is vital for guiding personalized treatment and improving prognosis. Our project aims to classify tumor stages by integrating multi-modal data—pathology imaging and genomic biomarkers—using a convolutional neural network (CNN) for image analysis, a random forest classifier for genomic data, and an AdaBoost ensemble to combine predictions. Due to dataset limitations, we shifted our focus from survival prediction to stage classification. Our results demonstrate the potential of multi-omics approaches in achieving nearly perfect classification accuracy and highlight the importance of collaborative methodologies in advancing cancer diagnostics and treatment planning.

1 Introduction

1.1 Background

Breast cancer is a leading cause of cancer-related mortality among women, with approximately 13% of women in the United States affected during their lifetime. Tumor staging is a critical component in oncology, influencing both treatment strategies and prognostic outcomes. Accurate staging facilitates tailored interventions, ranging from surgery to chemotherapy and radiotherapy. However, traditional diagnostic methods often rely on singular data modalities, such as clinical imaging or genomic analyses. These methods may overlook nuanced interdependencies between biological and morphological features, leading to suboptimal staging accuracy and delayed treatment.

Modern advancements in machine learning (ML) and artificial intelligence (AI) offer new opportunities for improving breast cancer staging by integrating diverse datasets. This integration leverages the complementary strengths of imaging and genomic data to provide more comprehensive diagnostic insights. Despite the promise of multi-modal approaches, achieving robust performance remains challenging due to the inherent heterogeneity of the data and the technical complexity of model integration.

1.2 Objective

Our project seeks to address these challenges by developing a multi-modal machine learning pipeline capable of accurately classifying breast cancer tumor stages. By integrating genomic biomarkers and pathology imaging, this study demonstrates the potential of hybrid methodologies in cancer diagnostics. Specifically, we employ a convolutional neural network (CNN) to analyze imaging data, a random forest classifier to process genomic data, and an AdaBoost ensemble to unify their predictions. This approach leverages the unique strengths of each model to improve classification accuracy and interpretability. Initially, our project aimed to predict patient survivability using combined genomic and imaging data. However, data limitations—specifically, the scarcity of overlapping datasets with both genomic and imaging modalities—necessitated a shift in focus. Tumor stage classification became our revised objective, offering both practical relevance and alignment with the available data. Accurate tumor staging has significant clinical implications, as it is crucial for personalized treatment planning and prognosis. By automating the integration and analysis of multi-modal data, our pipeline serves as a foundation for advancing precision oncology. Our work seeks to enhance diagnostic precision by leveraging multimodal data, thereby ensuring a more comprehensive understanding of tumor characteristics. We hope to improve clinical workflows and reduce the burden on clinicians by providing interpretable predictions for cancer staging. Lastly, we hope it bridges the gap between research and clinical practice.

2 Data Processing

2.1 Data Sources

We obtained pathology imaging data from the Cancer Imaging Archive, which included high-resolution tumor images capturing critical morphological details. Genomic data were gathered from the BRCA dataset and included RNAseq, miRNA, methylation, and mutational profiles. Although the original dataset contained 108 individuals, only 33 patients had both imaging and genomic data, which presented a significant limitation. To address this, we employed rigorous preprocessing and validation techniques.

2.2 Preprocessing

To prepare the pathology images for analysis, we converted them to grayscale, resized them to 224x224 dimensions, and categorized them into three stages (0, 1, 2). These steps ensured compatibility with the CNN while preserving diagnostic features such as cellular density and morphological irregularities. We used multi-threading to expedite these operations, allowing us to efficiently manage the complex directory structure. For the genomic data, we mapped clinical tumor stages to numeric labels and normalized the RNAseq, miRNA, and methylation features to comparable scales. This preprocessing helped reduce noise and align the genomic data for integration with the imaging dataset. Categorical encoding was applied to ensure consistency across classifiers.

3 Methodology

3.1 Convolutional Neural Network (CNN)

The CNN was selected for its ability to capture spatial and morphological features from pathology images. These features include patterns such as irregular nuclei, variations in cellular density, and tissue structures, all of which are crucial for distinguishing tumor stages. Our CNN architecture comprises three convolutional layers with ReLU activation functions, followed by max-pooling to reduce spatial dimensions. Fully connected layers consolidate these features, and a softmax output layer performs three-class classification. Dropout layers (with rates of 0.5 and 0.3) were included to mitigate overfitting, ensuring robust generalization during training. The CNN is trained on a 50/50 train-test split, achieving a validation accuracy of 85.7%. Figure 1 provides a visual overview of the CNN architecture, illustrating the sequence of layers and their respective parameters.

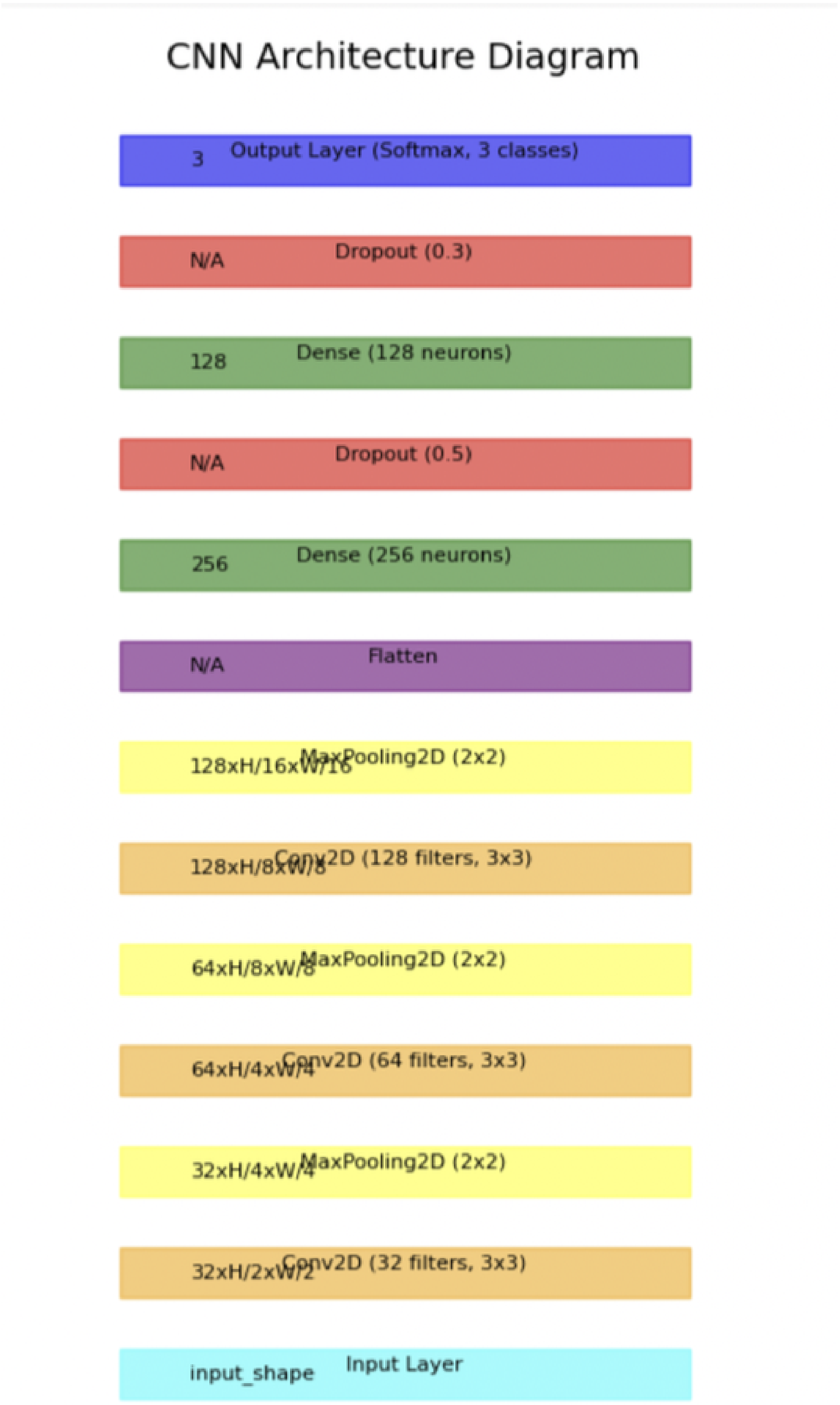
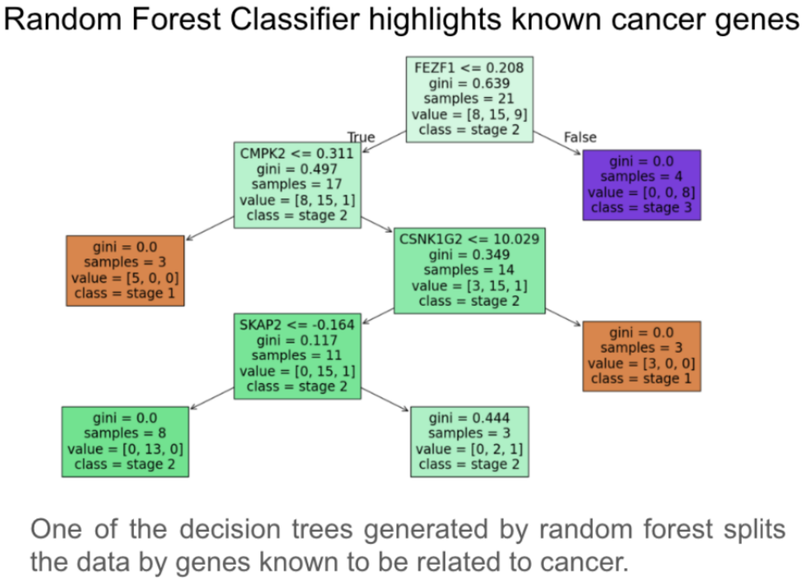


Figure 1: CNN Architecture Diagram.

3.2 Random Forest Classifier

We employed a random forest classifier to analyze genomic data, leveraging its capability to handle high-dimensional datasets and its inherent feature selection process. Genomic features, such as RNAseq, miRNA, and methylation profiles, were used to train the classifier. Using leave-one-out cross-validation, the random forest achieved an average accuracy of 50%. While limited in standalone accuracy, it provided valuable insights into the genomic contributions to tumor staging.



Random Forest Classifier highlights known cancer genes

One of the decision trees generated by random forest splits the data by genes known to be related to cancer.

Figure 2: Random Forest Decision Tree.

3.3 AdaBoost Ensemble

To integrate predictions from the CNN and random forest, we utilized an AdaBoost ensemble. AdaBoost dynamically adjusts the weights of base classifiers, prioritizing those with higher accuracy. This approach combines the spatial feature extraction capabilities of the CNN with the biological insights from the random forest, achieving nearly 100% accuracy in tumor stage classification. The ensemble's performance underscores the efficacy of multi-modal integration in addressing the limitations of individual classifiers.

4 Results

The CNN demonstrated robust performance, achieving an accuracy of 85.7% in classifying tumor stages. Its ability to extract morphological features from pathology images, such as irregular nuclei and variations in tissue structure, proved instrumental in distinguishing between stages. The confusion matrix visualizations (Figure 2) highlights the model's precision in early-stage classifications and minor misclassifications in borderline cases.

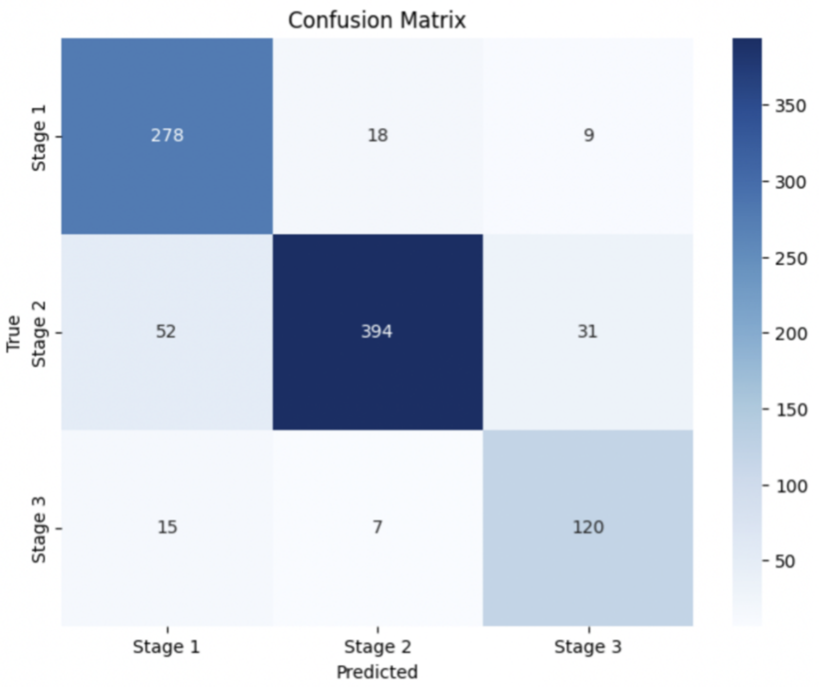


Figure 3: Confusion Matrix for CNN Stage Classification Results.

The random forest classifier provided biologically relevant but less accurate results, with an accuracy of approximately 50%. Its contribution to the ensemble model was supplementary, enhancing the robustness of predictions in edge cases where genomic data offered unique insights. The AdaBoost ensemble achieved nearly perfect accuracy, validating the effectiveness of multi-modal integration. This improvement underscores the potential of hybrid models to overcome the limitations of individual classifiers, paving the way for more reliable diagnostic tools in oncology.

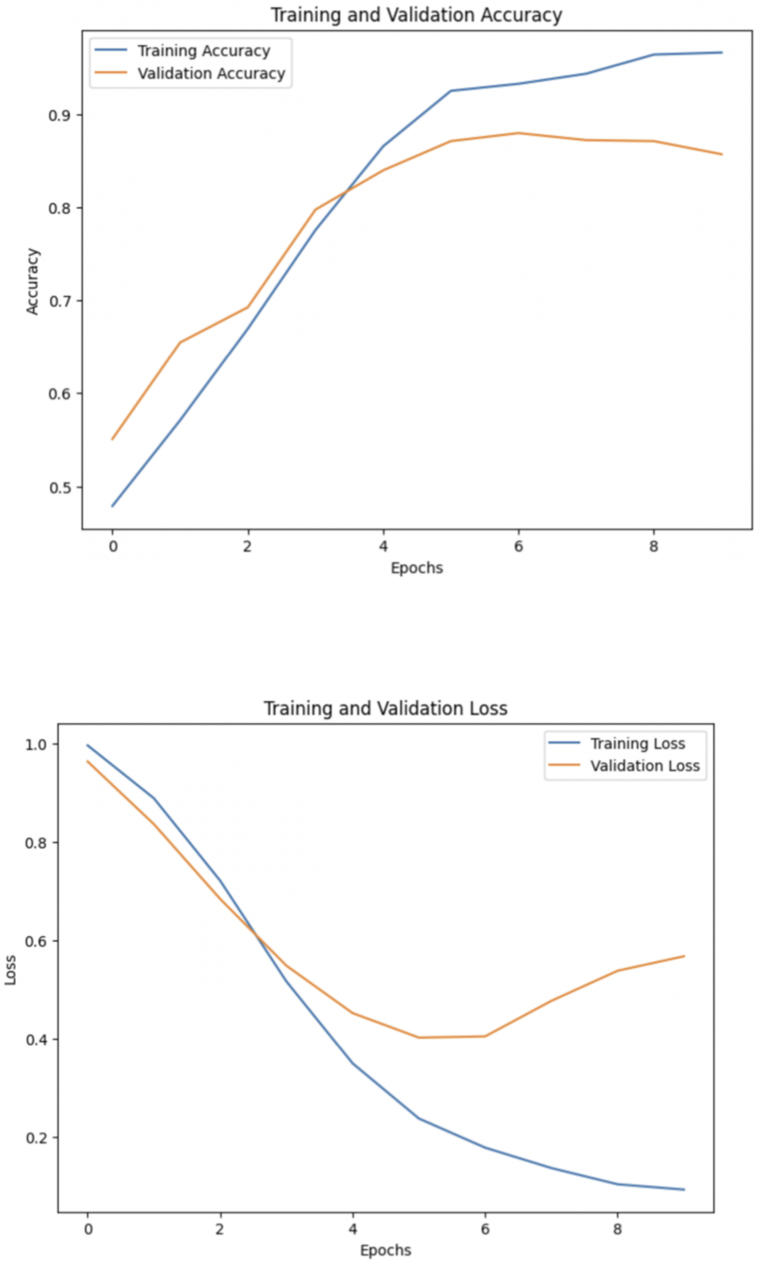


Figure 4: Training and Validation Accuracy and Loss Curves.

5 Conclusion

5.1 Challenges and Limitations

The primary challenge we faced was the limited dataset, with only 33 patients having both imaging and genomic data. This scarcity restricted the generalizability of our models and necessitated rigorous validation techniques. Additionally, integrating disparate data modalities required extensive preprocessing, including normalization, dimensionality reduction, and alignment of labels.

Overfitting was another challenge during CNN training, particularly given the small dataset. To address this, we employed dropout regularization and monitored validation accuracy to ensure balanced performance. For the random forest classifier, the high dimensionality of genomic features compounded by the limited sample size posed challenges in achieving reliable predictions. Expression data contains a lot of features (genes) most of which are noise and possibly are not related to breast cancer, which can confuse the model. Finally, to limit random forest overfitting, leaves were limited to a minimum of 3 samples.

5.2 Implications and Impact

Our project underscores the potential of combining imaging and genomic data for cancer diagnostics. By leveraging the strengths of deep learning and traditional machine learning, we demonstrated how multi-modal approaches can significantly enhance tumor stage classification accuracy. Accurate staging is crucial for guiding treatment decisions, such as determining the need for aggressive therapies or surgical interventions.

Clinically, this framework offers a pathway to integrate advanced analytics into diagnostic workflows, potentially improving patient outcomes through tailored treatments. The adaptability of our methodology also highlights its potential for application to other cancers or diseases where multi-omics data are available. Moreover, this project sets the foundation for further exploration of hybrid models in medical research.

5.3 Future Directions

To enhance robustness, future work could focus on expanding the dataset or employing data augmentation techniques. Exploring additional omics datasets, such as proteomics, and validating the framework in clinical settings are also promising directions.

Another promising direction involves refining the ensemble model by incorporating more sophisticated weighting mechanisms or experimenting with hybrid architectures, such as transformer-CNN models, to capture long-range dependencies in genomic data. Additionally, integrating unsupervised pre training on larger datasets may further enhance feature extraction capabilities. Validating the framework in clinical settings is a critical next step. By collaborating with oncologists and radiologists, we aim to assess the diagnostic utility of our models in real-world scenarios. Developing automated pipelines for preprocessing, training, and inference would also facilitate seamless integration into clinical workflows.

Lastly, the adaptability of our framework to other diseases beyond breast cancer should be explored. Expanding this methodology to other cancers or conditions with multi-dimensional datasets could establish it as a versatile tool for precision medicine.

6 References

1. Breast Cancer Facts and Statistics. Retrieved from BreastCancer.org.

2. Kang HYJ, Ko M, Ryu KS. Prediction model for survival of younger patients with breast cancer using the breast cancer public staging database. *Sci Rep.* 2024 Oct 28;14(1):25723. doi: 10.1038/s41598-024-76331-y. PMID: 39468113; PMCID: PMC11519337.

3. Poirion OB, Jing Z, Chaudhary K, Huang S, Garmire LX. DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. *Genome Med.* 2021 Jul 14;13(1):112. doi: 10.1186/s13073-021-00930-x. PMID: 34261540; PMCID: PMC8281595.

4. Lingle, W., Erickson, B. J., Zuley, M. L., Jarosz, R., Bonaccio, E., Filippini, J., Net, J. M., Levi, L., Morris, E. A., Figler, G. G., Elnajjar, P., Kirk, S., Lee, Y., Giger, M., & Gruszauskas, N. (2016). The Cancer Genome Atlas Breast Invasive Carcinoma Collection (TCGA-BRCA) (Version 3) [Data set]. *The Cancer Imaging Archive.*

5. Vasaikar SV, Straub P, Wang J, Zhang B. LinkedOmics: analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res.* 2018 Jan 4;46(D1):D956-D963. doi: 10.1093/nar/gkx1090. PMID: 29136207; PMCID: PMC5753188.

6. LinkedOmics TCGA-BRCA Data Download.

7. Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, David Madigan. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics, Ann. Appl. Stat.* 9(3), 1350-1371, (September 2015).