Multimodal Fusion of Whole Slide Imaging, mRNA Expression, and Clinical Features for Colorectal Cancer Tumor Staging Using XGBoost and Lightweight Convolutional Networks

Submitted by

Md Arif Ahammed Reza
ID: 221-35-951
Department Of Software Engineering
Daffodil International University

Supervised by

Musabbir Hasan Sammak
Senior Lecturer
Department Of Software Engineering

Daffodil International University

Bachelor of Science
DAFFODIL INTERNATIONAL UNIVERSITY

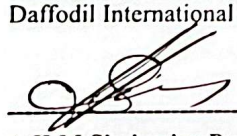
APPROVAL

This thesis titled on "**Multimodal Fusion of Whole Slide Imaging, mRNA Expression, and Clinical Features for Colorectal Cancer Tumor Staging Using XGBoost and Lightweight Convolutional Networks**", submitted by **Md Arif Ahammed Reza (ID: 221-35-951)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

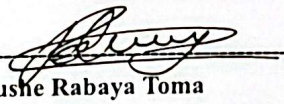
Chairman

Dr. S. M . Hasan Mahmud
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University



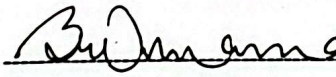
Internal Examiner 1

A.H.M Shahariar Parvez
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University



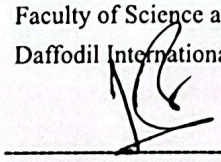
Internal Examiner 2

Tapushe Rabaya Toma
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University



Internal Examiner 3

Khalid Been md. Badruzzaman Biplob
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University



External Examiner

Dr. Md Sazzadur Rahman
Professor
Institute of Information technology
Jahangirnagar University, Bangladesh

DECLARATION OF THESIS AND COPYRIGHT

I declare that this thesis is classified as:

I acknowledge that Daffodil International University reserves the following rights:

1. The Thesis is the Property of Daffodil International University.
2. The Library of Daffodil International University has the right to make copies of the thesis for the purpose of research only.
3. The Library of Daffodil International University has the right to make copies of the thesis for academic exchange.

Certified by:

NOTE : * If the thesis is CONFIDENTIAL or RESTRICTED, please attach a thesis declaration letter.

SUPERVISOR's DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science.

(Supervisor's Signature)

Full Name : Musabbir Hasan Sammak

Position : Senior Lecturer

Date : 29-11-2025

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.

(Student's Signature)

Full Name : Md Arif Ahammed Reza

ID Number : 221-35-951

Date : 29 November 2025

Multimodal Fusion of Whole Slide Imaging, mRNA Expression, and Clinical Features for Colorectal Cancer Tumor Staging Using XGBoost and Lightweight Convolutional Networks

MD ARIF AHAMMED REZA

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Science

Department of Software Engineering (Major in Data Science)

DAFFODIL INTERNATIONAL UNIVERSITY

NOVEMBER 2025

ABSTRACT

Accurate tumor staging is essential for optimizing colorectal cancer treatment planning and improving patient outcomes. The interpretation of histopathological slides is used as the foundation of cancer staging, in the modern practice. However, inter-observer variability can also undermine this method and moreover there is the complex morphology of tumours. Computational pathology and the development of multi-omics data have given a leading opportunity to improve the accuracy of the staging through the combination of heterogeneous modalities.

In this dissertation, we are going to propose and evaluate a multimodal framework with fusion of whole-slide image (WSI) features, mRNA expression features, and clinical metadata (age) in order to do automated classification of colorectal cancer T-stage (T1-T4). A compound synthesizing dataset was constructed, including 1024 features derived on the basis of images, 128 mRNA features and the age data on 261 patients. After cleaning, preprocessing, and stratified partitioning, the normalized resultant feature vectors were used to train two families of predictive models: a lightweight convolutional neural network called TLCNN 2 used on large advise samples sizes with high-dimensional tabular fusion data, and an XGBoost classifier, which is especially effective on the small-sample space with high-dimensional biomedical data. The TLCNN-v2 achieved high results on the training data but was affected by over-fitting, suggesting that this was based on the small size of the training cohort. On the other hand, the XGBoost-Fusion model provided better generalisation and had the high final test accuracy of 57 per cent of the four tumour-stage classes and outdid other methods, including MLPs, PCA-reduced models, convolutional-net-based models, and augmented-WSI chunking strategies.Analytical results indicate that multimodal fusion improves discriminative power, and XGBoost provides stable learning under class imbalance and high feature dimensionality.

This work demonstrates that integrating histopathology, molecular expression, and clinical indicators meaningfully improves automated colorectal cancer staging. Future extensions may incorporate patch-level WSI transformers, multi-omics attention models, and data-level augmentation to further enhance performance.

Keywords: multimodal fusion, multi-omics integration, computational pathology, colorectal cancer staging, T-stage classification, whole slide image (WSI), WSI feature extraction, RNA-Seq, mRNA expression profiling, gene expression features, XGBoost fusion model, high-dimensional tabular data, lightweight CNN, TLCNN-v2, class imbalance, histopathology genomics fusion, early fusion, late fusion, multimodal deep learning, CRC T1–T4 prediction, WSI + omics fusion, attention-based fusion models, computational oncology, tumor staging automation.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Full Form
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the Curve
CNN	Convolutional Neural Network
CLAM	Clustering-constrained Attention Multiple Instance Learning
CRC	Colorectal Cancer
DL	Deep Learning
DNA	Deoxyribonucleic Acid
EMR	Electronic Medical Record
FN	False Negative
FP	False Positive
F1-Score	Harmonic Mean of Precision and Recall
GBDT	Gradient Boosted Decision Trees
GPU	Graphics Processing Unit
HE	Hematoxylin and Eosin
LR	Learning Rate
MAC	Multiply–Accumulate Operations
MIL	Multiple Instance Learning
ML	Machine Learning
MLP	Multilayer Perceptron
mRNA	Messenger Ribonucleic Acid
MNIST	Modified National Institute of Standards and Technology Dataset
NLP	Natural Language Processing

PCA	Principal Component Analysis
QC	Quality Control
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic
SGD	Stochastic Gradient Descent
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
SMOTE-NC	SMOTE for Nominal and Continuous Features
TCGA	The Cancer Genome Atlas
TLCNN-v2	Thin Lightweight Convolutional Neural Network Version 2
TNM	Tumor Node Metastasis Staging System
TP	True Positive
TN	True Negative
ViT	Vision Transformer
WSI	Whole Slide Image
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting
SBPWM	Simple Boost Pulse Width Modulation
ZSI	Z-Source Inverter

CHAPTER 1

INTRODUCTION

1.1 Background and Clinical Importance

Colorectal cancer (CRC) is one of the most prevalent illnesses worldwide, and a major cause of illness and death associated with cancer [1]. Its incidence is becoming increasing worldwide due to aging populations, diet, physical inactivity and genetics. Early detection and assessment of the tumor is of importance, as CRC may look and behave very differently from case to case. Correct staging of the tumor is important to manage the patient, the selection of treatments, and the prediction of the outcome. The T-Stage (T1 T4) is a quantitative measure of how much the malignancy has invaded the bowel wall and how much the malignancy has invaded adjacent structures of the TNM classification. This stratification is essential in the planning of surgical interventions, the risk of recurrence prognostication as well as the adjuvant modalities devolution of adjuvant therapy including chemotherapy or radiotherapy [2].

Traditionally colorectal carcinoma staging is performed by placing under microscopic observation, an inspection of hematoxylin and eosin (H&E)-stained whole slide images (WSIs) as carried out by pathologists. Despite yielding essential morphological data, the usefulness of this standard method is limited by the variability of the inter-observer data, i.e. small changes in the location or the appearance of a tumour may lead to different stage allocations for that tumour. Further, standard histopathology is anthropomorphic, and as such it ignores the heterogeneity of the tumour (at the molecular and genetic level) on the surface of the tumour that affects the aggressiveness of the tumour and its response to treatment. Trusting only on morphological-criteria is hence detrimental to accurate risk stratification and optimal selection of therapies [3][4].

In order to overcome such constraints, modern oncology is increasingly moving to integrative multimodal data like histopathological, molecular and clinical data to design powerful predictive models [3]. Recent innovations in high-resolution imaging, digital pathology and machine learning algorithms were now able to make possible the

harvesting of quantitative, image derived information on WSIs associated with genomic changes, microenvironmental properties of tumours, and clinical prognoses. Combined with other molecular parameters (e.g. gene expression profiles, mutational status etc.) these methodologies give a global picture of tumour biology. Demographic and systemic data, age, comorbidities, Haemoglobin values and so on: there is even more data to use, the model is more fidel to use, which simplifies the prognostication of patients and their stratified treatment. Based on more than morphological evaluation, multimodal strategies are anticipated to not only guarantee a higher level of T-stage, but also a reduced diagnostic concordance, and a variety of new biomarkers that can be targeted using therapeutics [6][7].

To summarize, though histopathologic staging will always have its roots in the treatment of colorectal cancer, the fusion of multimodal data is a critical step towards having precise oncology. The intersection of the enhanced morphological information on tumour with the molecular and clinical information supports the step for the development of the more effective outcome-prediction systems and therapeutic decision models and aims to increase the patient care and treatment effectiveness in the end.

1.2 Motivation for Multimodal Fusion

The different biomedical modalities provide different, and complimentary, insights into the biology of tumours. As an example, attributes or features derived using whole-slide images (WSIs), embed complex morphological architectures, tissue organisation, detail of the tumour microenvironment. These visual cues are able to reveal some nuanced aspects of tumour behaviour such as local invasion, stromal crosstalk or immune infiltrate that would otherwise be elusively hard to measure using alternative data streams. During a concert the expression profiling of the mRNA provides a molecular window as it measures the extent of gene transcription as well as indicating the presence or presence of a particular signalling cascade in action. The attributes of tumour aggressiveness, proliferative potential and possible therapeutic limitations which are not evident in histopathology can be revealed in this molecular view. In the meantime, clinical variables, like patient age, sex, comorbidities, and others consider demographic variables offer the scaffolding of contextualization of these biological measures, which can encompass the whole patient level of heterogeneity, which may impact the course of the disease and response to the treatment.

Although every of the modalities provide useful information separately, the use of these combined provides a more comprehensive and more subtle account of oncologic

biology. Incorporating WSI, mRNA and clinical data, researchers will be able to identify inter-modes of relationships that would otherwise remain concealed when a single modality is considered in isolation. As an example, particular gene expression phenotypes can be consistent with particular tissue structure or invasive phenotype, but clinical factors such as age or comorbidities can change the effect of molecular and morphological features on clinical consequences [4]. Such cross-modal interactions are especially apparent in colorectal cancer where tumours are extremely heterogeneous and variable in their presentation to the patient, leading to the invalidity of standard staging processes.

This synergy is the foundation of multi modal fusion as a basis of classifying T- stage of colorectal cancer. Through the combination of complementary information data will be possible to retrieve the cellular and molecular complexity of the tumour as well as the context of a patient one emerges in. The overall concept in this dissertation is to make use of these multimodal representations in order to increase the predictive power, and aims to perform better than either modality that could have been used alone in terms of predictiveness and strength. Through this, the effort will not only lead to a better understanding of how to better categorize T vastly, but will also results in a greater understanding of how morphological, molecular and clinical characteristics coerce to decide the course of colorectal cancer and patient behavior.

1.3 Challenges in Multimodal Learning

Multimodal learning has quite sufficient potential for enhancing the level of medical diagnostics, although there are still a number of practical and methodological obstacles. These are headed by the fact that the accessible datasets are confounded. We have conducted the investigation on 261 patients in our current study which is, though is common in the biomedical research, not enough to train deep neural architectures including millions of parameters. The small sample size makes overfitting more likely and has a negative impact to the extrapolation of the acquired model to new cases.

Second, there is the problem of the high dimensionality of the problem, adding to the further complication. The embeddings of Whole-slide-images have 1,024 attributes, while the addition of the mRNA expression profiles has an extrapolated dimension of the profile of 128 features, and the addition of clinical covariates to the feature space has more than 1,150 dimensions. Not only is the high dimensionality of corpus inflationary to the problems of overfitting, it also causes noise to be introduced in the selection of features, which requires careful corpus model design and potent regularisation.

The third difficulty is the fusion of different forms of data. The different modalities have different statistical distributions and noise profiles, which requires a stringent normalization, alignment and representation process in order to promote meaningful cross-modal interaction. Lack of adequate treatment of these differences may disable the ability of the model to generalize about biologically interesting trends.

Lastly, the training pipeline is complicated with the class imbalance T -stage (T1 -T4). In event that the population sample is dominated by some stages, the model can begin to have a bias on that stage, thus performing worse on the less represented ones, decreasing its clinical applicability.

Such challenges strengthen the need to have models that are strong enough and flexible enough to move through the heterogeneous, multi faceted and disproportionate data environments that modern biomedical studies consist of. Their conquest will be at the core of the multimodal learning in the precise classification of colorectal cancer T -stages and to drive precision oncology.

1.4 Overview of Methods and Contributions

In order to address the complexities involved in multimodal learning, when it is implemented for biomedical tasks, we have formulated an overall fusion pipeline consisting of feature extraction, normalisation, classification and evaluation. We apply a combination of deep learning and classical machine-learning approaches to compare them systematic one to another (in terms of their relative strengths) when confronted with high-dimensional and low-sample-size data.

In the deep-learning part, we designed the small one-dimensional convolutional network, TLCNN -v2 which can provide the local interaction in fused feature vectors. Such an architecture is specifically tailored to cope with peculiarities of high-dimensional biomedical data, and balances the capacity and the risk of over-fitting. At the same time, we used a gradient-boosted tree model (XGBoost) as it is predisposing to the use of tabular data with high dimensionality and low samples. Though the CNN-based models showed a higher performance in the training process, XGBoost model proved to be the more stable and reliable predictor of generalisation with 57 per cent accuracy in an independent test set.

The importance of proper selection of well-adaptable model in the case of small biomedical data sets is about the viability of the results, due to their essential and dependable nature without graphics aids. The main parties of this thesis can be summarised to be the following:

- A whole-slide images/mRNA based expression pattern/clinical characteristics multimodal data set used to classify colorectal cancer at T-stage optimized.
- A common multimodal fusion pipeline which contains pre-processing, feature alignment, normalisation and label construction.
- The creation of convolution network which is TLCNN -v2 and is directly designed as a high dimension biomedical vector, designed as per constraints of lightweight model.
- An extremely high performing XGBoost model (with a test performance of 57 per cent.) that bested the deep learning baselines in terms of generalisation.
- An overall discussion of the issues and tradeoffs inherent to the multimodal learning using small medical datasets that will be informative on future studies.

Overall, the present piece of work helps to realize that integrating complementary data modalities using well-designed pipelines can significantly enhance predictive modelling in colorectal cancer but, at the same time, points to the aspects of practice that need to be considered when dealing with minute and heterogeneous biomedical data.

CHAPTER 2

RELATED WORK

2.1 Introduction

The field of computational pathology has arisen rapidly stimulated by the growing availability of imaging wholenership imaging (WSI), spatially transcriptomic and clinical data have initiated major progress in accurate oncology.

Recent survey reveals that combining histopathological data, genomic data, and clinical clinical data could enhance the process of cancer diagnosis, staging, and prognosis in a way which is not possible with the use of only one modality. In spite of this however, the literature recognizes a number of significant barriers to building effective multimodal fusion pipelines especially in cases where small sample sizes are used. The main ones are: high dimensionality of omics data, the heterogeneity of WSI data and the fact that meaningful correlations can only be drawn between data of different types [2][3].

The next section investigates the previous works which are the most relevant to the present study. We focus on combining different types of data, working with WSIs, modeling different types of omics data, and using machine learning models, such as XGBoost, for biomedical prediction. Together these studies provide motivation and support to the multimodal fusion architecture introduced in this thesis.

2.2 Multimodal Fusion in Computational Pathology

Multimodal fusion is currently a hot topic in research in computational pathology. Most of the studies report three primary ways that different types of data may be combined:

Early fusion, or feature-level fusion The histology features, omics data and clinical variables are combined at an early stage before classification Reviews in multimodal oncology note that early fusion is easy to use and maintains inclusive details of each data type. However, it can be unreliable if the number of features is larger than the number of samples. This is a typical issue in genomics, where several thousand gene expression features are typically reduced before fusion [4][5][6].

Late fusion (decision-level): In this, separateLate fusion or decision level fusion trains separate models for each type of data and then uses a fusion of the separate results, either the average or some more advanced technique. Recent surveys have found that

late fusion is often more robust when models are allowed to fuse after all data types are available which is often not the case. However, it may not consider complex interactions between types of data, which can result in a model that poorly predicts outcomes. [7][8][9].oaches, including multimodal foundation models and deep survival predictors, to integrate information within the network itself, using mechanisms such as attention, cross modal transformers, or learned joint embeddings.These methods are dominating most of the deep learning literature and have shown their strong performance on large-scale dataset. However, they usually require large amounts of training data as well as careful regularization to prevent overfitting [10][11][12].

Through a list of review articles, a general agreement on how to choose the fusion strategy gets obtained. one will have to be informed by the magnitude and the complexity of the data set. For small to medium sized studies, classical machine learning algorithmology in combination with carefully optimized.fusion pipelines are not destroyed, in fact they tend to be more stable and even more understandable. and in comparison to their more advanced deep learning classifications.

2.3 Deep Learning on Whole Slide Images (WSI)

WSIs have overwhelming processing problems due to their very large spatial resolution. Based on this, the significant modern methods of analysis pipelines usually make use of image tiling by patch sampling and feature-level aggregation. Patch level descriptors are typically obtained with a convolutional neural network ("CNN") such as "ResNet", "EfficientNet" or some other lightweight versions,Descriptors are then combined to generate slide-level representations.

Critical summaries of the literature in the history of histopathological imaging field all point to the infeasibility of the concept of comprehensive training of WSI, largely based on costliness and scarcity of pixel-level or region-level labels. In order to alleviate such a bottleneck the weakly supervised paradigms have come into much use mostly based on multiple instance learning (MIL).Collective features extraction "MIL" informative slide-level features extraction based on patch $embedding aggregation w/o exhaustive annotations. In addition, the feature extractors based on CNN and transformers have demonstrated their capacity to identify morphologic signatures that have been linked to tumor stage, mutation burden, and general prognosis, thus pointing to the wealth of biological information encoded in WSIs [see, e.g., reference [14]).

However, these methodological advances-which can not make it invisible that deep "WSI" encoders are already using significant amount of computational resources while having strong risk of overfitting-especially when used on -- cohorts of limited size. The computational cost is still large even with subsequent developments and models over fit on small sample sizes. Experimental research supports-the idea that no-large and complex models are often additive of the underlying generalisation of small-sample research, even if they are good at predicting on training-data.

This has collectively led to an opinion by an important portion of the research community that the utilization of pretrained CNN backbones or small networks tuned to a specific task are able to balance between expressive capacity and strength. Our methodology is no exception to this philosophy as we have used a light CNN for extracting WSI 1,024 dimensional features, thus preserving salient morphological features while at the same time keeping the models simple to prevent overfitting.

2.4 Multi-Omics and Transcriptomics Modeling

The combination of the data in omics and especially, the mRNA expression profiling data, have turned out to be the key element in the modern studies in oncology. Extensive multivariate omics studies as regularly find that gene expression profile gives a very specific biological metric, which can be used to gain fine-grained knowledge about tumour phenotype, immune infiltration, risk of disease-progression, and responsiveness to therapy. Transcriptomic-profiles represent constraint molecular paths characterizing tumour behaviour@beyond purely morphological@ones and therefore differ in the uniquely interesting aspects of prediction modelling and risk stratification.

A variety of modelling approaches are reported in academic literature and this is used to incorporate mRNA expression data in predictive models on a regular basis. Traditional methods often use dimensionality reduction techniques, such as principal component analysis or feature selection methods based on biologically meaningful sets of genes, in order to overcome the problem caused by the large dimensionality of the vectors of gene expression. In part helped by more recent research, machine-learning algorithms, including regularised linear classifiers, tree-based ensemble and neural networks, are used to identify complicated relationships between transcriptional patterns and clinical outcomes. In multimodal contexts, histopathological and clinical-data are habitually combined besides the gene expression characteristics, and thus models are available to relate the molecular signatures to tissue-level phenotypes and patient-provided features. However, despite having the richness of mRNA models, it has limitations in its practical applicability, especially with small cohort studies. Generalisation can be reduced by high dimensionality, stochastic noise and batch effects, unless exact measures are carefully taken to address the problem of normalisation, feature-selection and regularisation. As a result, a number of researchers insist on the need to compromise the model complexity and the dataset size, in favour of smaller and interpretable representations, which don't remove biologically meaningful signals. Our methodology is directly guided by this principle as the mRNA expression characteristics are added in a dimensionally reduced format in a controlled manner to complement the whole-slide imaging (WSI) and clinical modalities without overloading

the learning process. The most commonly used modelling method to pool mRNA expression data include:

* Variance filtering and normalization, which are typically employed as early ones to clean out the non-informative genes and lessen technical noise between the samples.

* Dimensionality -reduction models such as principal component analysis (PCA) or autoencoders -type embeddings, in which the high-dimensional profile of expression is mapped to a lower and dimension in which biologically meaningful variation is preserved.

* Regularised logistic and linear models, which have the advantage of being interpretable, but also of imposing some constraints on overfitting by a penalty, such as L1/L2 regularisation.

Tree based ensemble involving random forests / gradient boosting machines that not only have the ability to capture non-linear correlations but also feature interactions without necessarily being engineered to do so

* Hybrid deep learning systems, in which combination of omics data and WSIs or clinical variables are combined to represent simultaneously both the molecular and morphological data and the patient level data.

This is a nagging problem for all multivariate omics survey data: so called the curse of too many dimensions. Gene-expression vectors are regularly constructed consisting of 10 to 20 000 features, but usually you only have the few hundreds of patients at your disposal. This lopsity is fraught with the risk of overfitting and creates instabilising learning which makes the dimensionality reduction vigorous and the regularisation forceful absolutely essential. With these ideas as a guide, we summarize the expression of the mRNAs into a small 128 dimensional latent representation , followed by multimodal fusion compressing all the necessary signals in the data, but without losing strength from the model.

2.5 Classical Machine Learning and XGBoost in Biomedical Prediction

Although deep learning has become the dominant paradigm in the image analysis field, more traditional approaches to machine learning, such as gradient boosted decision trees such as XGBoost, do still seem to show good performance in biomedical applications where small datasets may be considered. Several reviews and comparative studies uniformly report multiple benefits of the XGBoost algorithm in such settings at the operational level of usage:

Effective Handling of Heterogeneous Feature Spaces, Well-suited for Multimodal Fusion, where both WSI Embeddings, Omic Features and Clinical Variables Vary in Scale and Distribution .Robustness to irrelevant features, noise, which is worst in highDimensions Biomedical data where not all extracted features are informative.

Good generaization capability with low samples tree-based ensemble don't require huge (in terms of elements) training data to learn stable deision boundaries.Superiority over deep neural networks under small sample sizes (which are usually < 1K patients) where deep neural networks usually overfit even with extensive regularization.

In relation to multimodal cancer research, there are several papers that further show that carefully designed pipelines, combining CNN-based feature extraction with XGBoost-classification, can be more powerful than deeper and fully end-to-end multimodal networks. This performance advantage is attributed mostly to the separation of representation learning and classification, and thus allowing each component to be optimized for its task. This observation is directly in line with the design choices of our thesis: for the extraction of compact WSI embeddings, we used a lightweight CNN, mRNA features are compressed into low dimensional features and XGBoost is used as the final fusion classifier. Together, such architecture is able to balance expressiveness of representation with robustness and generalization for a small cohort environment.

2.6 Summary and Relevance to Our Work

Recent studies in computational pathology and multi-omics learning have uncovered evident trends that determine how multimodal prediction models are designed. Studies in various types of cancer and clinical outcomes indicate that combining various types of data provides more accurate and meaningful predictions than a single source of data can. However, the literature also highlights some practical challenges, notably the limits of deep learning when some deep learning algorithms are employed with small and high-dimensional biomedical data. These findings provide a practical framework for creating good multimodal systems operating well with real world data.

The research is consistent about the following:

Combining a number of different types of data results in better results for the staging and prognosis of cancer than using only one type of data because it gives a fuller picture of the tumor biology.

Transcriptomics (WSI encoders) and omics data is a great combination. Patterns observed in histopathology often support or do explain molecular signals which helps to better diagnose and give prognosis.

Deep learning models tend to require a large amount of data in order to work well, whereas traditional methods of machine learning tend to be more effective and predictable with small biomedical datasets.

For multi-omics problems with multifarious numbers of samples, the combination of deep feature extraction with the XGBoost classifier may be often taken as a hybrid solution. This approach is a compromise between good representation of data with good performance.

To maintain a stable performance in the high-dimensional multimodal settings, it is important that dimensionality reduction, regularization, and well-planned fusion strategies are used.

Based on these findings, decisions made in the design of this thesis are based on established best practices:

A light weight CNN derives miniaturised and helpful WSI embeddings,

mRNA expression data is reduced to a low dimensional form,

Clinical variables are included to reflect differences among the patients,

Early feature-level fusion is used to retain interaction between data types, and

XGBoost is considered to be used as the last classifier for establishing strong learning and generalization.

In general, this design addresses the key challenges described and discussed in prior research: small dataset size, high dimensionality of features, and difficulty modeling interactions between distinct types of data. It is also consistent with the practical requirements of biomedical research.

CHAPTER 3

DATASET DESCRIPTION

3.1 Introduction

The data set used in this study is a multimodal patient-level data set including whole slide image (WSI) features, mRNA expression profiles, and clinical variables. Following preprocessing and cross modal matching a total 261 colorectal cancer patients were retained for analysis. While the general structure for the dataset is similar to publicly available cohorts such as TCGA, all the patient identifiers in this study have been completely anonymized and processed following data privacy and ethical guidelines.

This chapter gives a detailed description of each modality of data, the steps taken in preprocessing the data to make the data consistent and good quality across all the sources, and the distribution of the tumor T-stage labels that were utilized for supervised learning. By explicitly defining these components, this section lays the groundwork for the multimodal modeling and evaluation strategies discussed in the role of the modeling strategy to be used in the following chapters.

3.2 Patient Cohort

After the merging of the WSI records, the mRNA expression files and their related clinical metadata, a complete multimodal dataset was built. The resulting dataset of The Lucky Horses had the following characteristics:

Total number of patients: 261

Complete multimodal coverage Each patient record includes WSI features, mRNA expression data, patient age, and tumor stageUnique case identifiers: All the records are related to unique patients/No duplicatesThis compiled data set then forms the basis for all multimodal fusion experiments performed in this thesis in order to ensure consistency though the model training, validation and evaluation processes.

3.3 Whole Slide Image (WSI) Features

Whole slide images (WSIs) are extremely-large digital pathology scans which cannot be directly processed without special preprocessing. In this work, each WSI was used and hence, analyzed with a light weight CNN based feature extraction pipeline that provides representational richness and computational efficiency.

WSI representation : Each patient is represented as a 1024 dimensional WSI feature vector, which forms an attractive concise summary of slide-level morphology.These embeddings are accomplished using a multi-step process consisting of tiling image, extracting features at the patch level and using a CNN to aggregate the features.The resulting representation captures important morphological features such as tumor aggressiveness, glandular architecture and overall tissue organization, which are known to be important in the colorectal cancer staging.By compressing information within morphological histological slices of high-resolution images into a fixed-length embedding, this can be used effectively to link WSI morphology with molecular and clinical data. At the same time, it does not incurring the vast computational and memory expenses that are required by full-resolution, end-to-end WSI modeling so it is well suited for small to mid-size biomedical datasets.

3.4 mRNA Expression Features

mRNA expression data are by nature high dimensional; they may consist of the measurements of tens of thousands of genes. To make these data suitable for predictive modeling, transcriptomic profiles in the current study were subjected to carefully consideration in feature selection and compression in order to decrease the dimensionality while retaining biologically meaningful information.

mRNA representation : Each patient is represented by a feature vector of mRNA data of dimension 128, which represents a condensed summary of the activity of the transcriptome.These features are extracted by a set of multi-step preprocessing, including but not limited to normalization and log transformation, when suitable, to mitigate the level of technical variability.

- Dimensionality reduction methods, that combines variance based filtering with latent space projection is then made to remove uninformative genes and compress the remaining signal.
- The representation that results describes key biological processes related to tumor growth, immune response and microenvironmental interactions, which play a central role in colorectal cancer progression.

By reducing transcriptomic data into a low dimensional and structured form, this method assists in effectively fusing with WSI and clinical features. At the same time, it helps to protect from overfitting a traditional machine-learning model, which makes it suitable for analysis in small-cohort biomedical studies.

3.5 Clinical Features

Clinical data are related to patients and provide important information at the level of the patient that complements molecular and histopathological data. For this dataset, the clinical modality is one feature, which is age of the patient at diagnosis. Despite its simplicity, age is a well-established cancer progression and prognosis predictor. Incorporating this feature into the multimodal fusion model makes it more robust for predictive analysis because it accounts for patient-specific factors that affect the course of specific diseases, where the fusion model would be better able to contextualize imaging and transcriptomic signals.

3.6 Tumor Staging Labels

Tumor staging labels used in this study were based on the T1-T4 tumor annotation labels that were provided in the dataset, with each label corresponding to the pathological T-stage of the tumor. This setup naturally defines a four-class classification problem with each patient belonging exactly to one of the four categories, this is a problem that supervised learning programs can be used.

Stage *Meaning*

- ➔ T1 Localized tumor Early invasion
- ➔ T2 Tumour invade into the muscle layers
- ➔ T3 It inveigles deep into the adjacent tissues.
- ➔ T4 Advanced tumor Homoics spread outside homoics primary

By presenting the task in this way, the model must be able to differentiate between gradually increasing levels of invasive disease, and thus clinically meaningful differences in tumor progression. This multi-class classification framework enables us to train the multimodal fusion pipeline with information that is expected to correlate between the morphology of WSI, transcriptomic information, and patient-specific information and tumor stage, which will ultimately aid in accurate

3.7 Class Imbalance

The distribution of the stage label of tumor occurs in the following ways in the cohort:

T1: 97 patients T2: 57 patients T3: 54 patients T4: 53 patients

While the class imbalance in this data set is not extreme, the imbalance in class sizes is not. In particular, the smallest classes (T3 and T4) contain about half as many patients as the largest class (T1), which might possibly bias the training of the models. Such imbalances were considered in the design of the predictive framework, which affected the choice of XGBoost (which is robust to class imbalance), as well as the consideration of suitable data augmentation and regularization strategies, so that the learning is stable and reliable for all T-stages.

3.8 Summary of Feature Dimensions

Each patient record in the dataset has the following features:

- ➔ 1024 WSI based features extracted from whole slide images
- ➔ 128 mRNA features of compressed transcriptomic profiles
- ➔ 1 clinical feature (age upon diagnosis)-
- ➔ 4 target labels; T n of tumor T-stage (T1 through T4)

Each patient is characterized by 1,153 input features, which introduces a high-dimensional and multimodal data set. This rich variety of features lends itself very well to applications with fusion techniques, where there is a combination of both learning deep features within images with classical machine learning methods to learn information on tabular data. By incorporating morphological, molecular and clinical data, the models are able to learn about complex relationships that are important in proper tumor staging and outcome prediction for patients.

CHAPTER 4

METHODOLOGY

This study proposes a multimodal fusion pipeline, which integrates the whole-slide image (WSI) embeddings, mRNA gene expression data, and clinical variables into one machine learning set up to classify T-stage in colorectal cancer. The pipeline is specifically created to overcome the problems caused by small and unbalanced datasets and extremely high dimensional spaces. As some of the nuisances have been addressed: Pre-processing and normalization of the different modalities, feature engineering, dimensionality reduction and data augmentation to ensure that the data is being learnt in a stable way. A light one-dimensional convolutional network (TLCNN-v2) is used to extract local patterns in fused feature vectors while a robust classifier of an unknown data space (XGBoost) is used as classifier, because it can work with heterogeneous data spaces and prevent overfitting. Throughout the pipeline, early fusion approach is followed, in which features from all modalities are concatenated together in a single representation before classification, so that the model can utilize complementary morphological, molecular and the clinical information.

By leveraging the representational power of deep feature extraction with traditional machine learning techniques, this presentation of ideas allows for a balance of representational power and generalisation, making it an especially appropriate methodology for biomedical datasets with few cohorts. The pipeline helps to collect cross-modal meaningful interactions, and empowers the model to learn complicated relationships between tumor morphology, transcriptomic activity and patient-level characteristics. Overall, this methodology allows a scalable and interpretable framework for the multimodal colorectal cancer staging and suggests insights that could be utilized in future studies in precision oncology and the construction of solid predictive models for high-dimensional, heterogeneous biomedical data.

4.1 Preprocessing

The first fusion dataset had duplicate identifiers as a result of the merging process (e.g., caseid and caseid.1). In order to ensure consistency, these duplicate columns were removed, so that one unique patient identifier per record was retained. Raw clinical staging variables initially one hot fields of T1-T4 were carefully processed and later

summarized as a single tumor stage label which served as the target for supervised learning. Stage Label Construction In order to form a consistent classification target, a deterministic rule of label creation applied to one-hot encoded T-stage variables was used. Each patient was directed to one of the stages by the following scheme:

- ➔ if T1 = 1, stage = 1;
- ➔ if T2 = 1, stage = 2;
- ➔ if T3 = 1, stage = 3;
- ➔ if T4 = 1, stage = 4.

Any rare or undefined cases were classified at an assigned stage of 0. This process produced a clean, four class supervised learning target which can be used by both XGBoost and CNN based models.A stratified train-test splitting strategy was followed to ensure the same class distributions for the train and test cognitive samples used.Given the number of patients in the cohort (261 patients) a train-test split of 80/20 (208 training, 53 testing) was used. To try and preserve the initial class distribution by ensuring that minority classes are represented well during model evaluation, stratified sampling was used. This way, possible bias performance metrics of the models and possible underrepresentation of each stage category in the training and testing sets can be avoided.

Feature Scaling:Due to the fact that the dataset was multimodal and contained numerical features (WSI embeddings, mRNA latent vectors, age), scaling was necessary. Standard Scaler was applied:

$$X_{scaled} = \frac{X-\mu}{\sigma}$$

Scaling helps to stop high dimensional WSI features from dominating the gradient when optimizing.

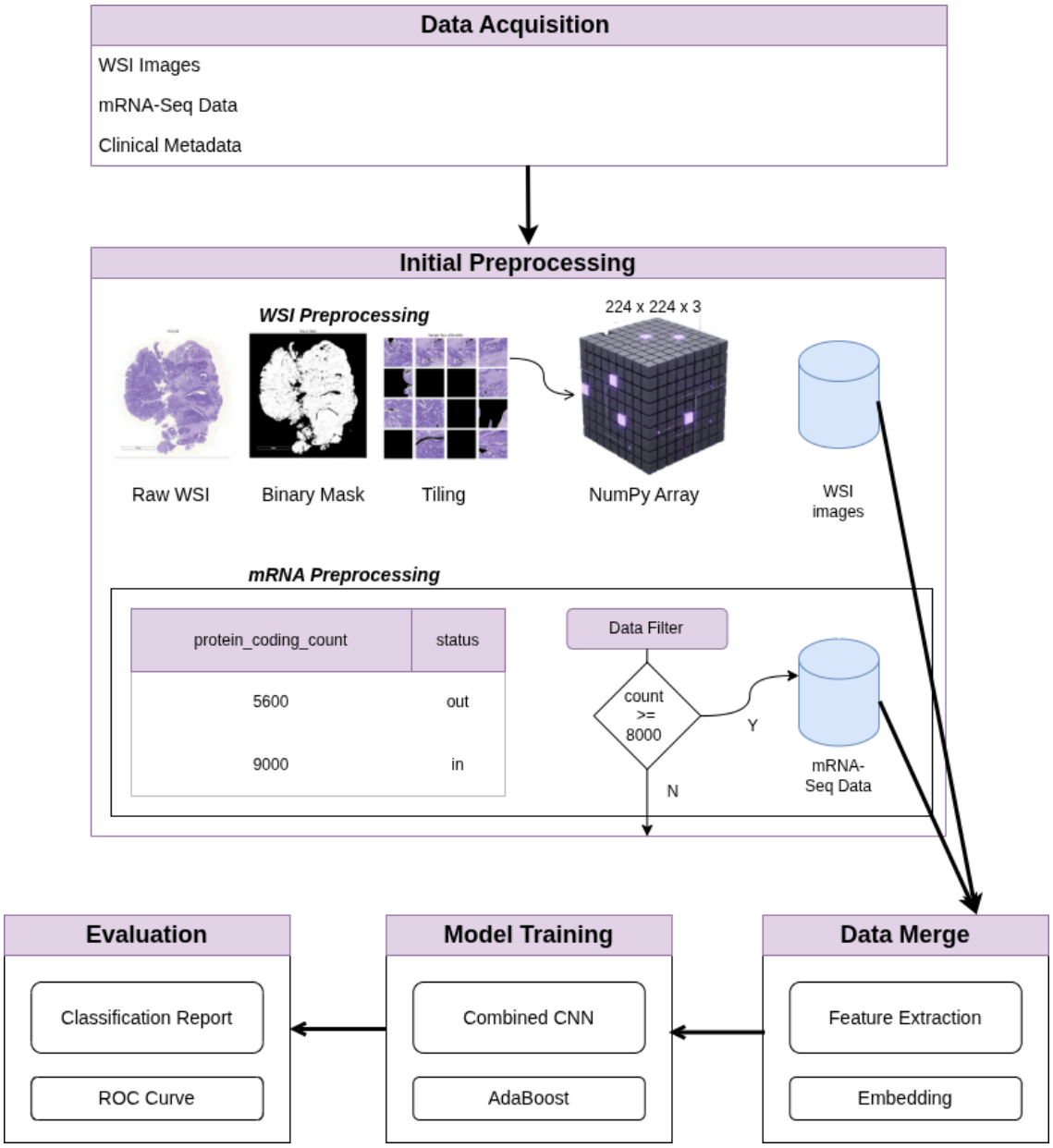


Figure 4.1: Proposed Methodology

4.2 Feature Engineering

The Dimensionality Reduction using the PCA Analysis

Principal Component Analysis (PCA) was applied in order to decrease redundancy andde-noising the feature sets. The PCA was used separately, for each modality, for WSI embeddings (1024 features) and mRNA vectors (128 features). The number of retained components was calculated according the variance retention levels and performance in the validation. By reducing the original features to a lower-dimensional space, PCA achieved improved generalization, reduced the noise and made model training even faster without losing any biologically relevant information.

Augmentation using Chunking (WSI)

Given that the initial contribution of each patient was a single 1024 dimensional WSI vector, the size of the dataset was still limited. To augment the data along with this, we have added a WSI chunking strategy where we divide each vector into equally-sized blocks to create multiple pseudo-samples per patient. For example, under a chunk size of 50, we obtained around 20 chunks for each WSI vector that had the same label as the tumor stage as in the original patient. This somewhat increased the effective size of the dataset of about 5,200 training chunks and 1,300 test chunks, making use of internal variation within the WSI embeddings, while maintaining their biological meaning.

Multi-Sample Augmentation: Chunk Slicing?

In addition to simple chunking, overlapping or sliced blocks of the WSI vectors were employed to generate multiple augmented samples per patient to further view increase the diversity of samples and introduce the model to richer patterns to learn from. This strategy helps the network to capture the localized variation of the WSI embeddings which can improve the robustness and help to mitigate the problem of overfitting in the high-dimensional and small-sample setting. In addition to fixed blocks, one can use the sliding-window approach:

$$chunk_i = X_{WSI}[\, i{:}\, i + L\,], \quad L = 50$$

Sliding windows lead to more diversity in the training distribution and will enable obtaining model learning based on fine-grained differences encoded in the embedding space, either and spatial (or feature-based).

Normalization and Fusion

Following augmentation, each WSI chunk was then combined with a corresponding mRNA PCA features and patient's clinical variables (age), yielding a oneness of representation. Once concatenated the resulting feature vectors were standardized to have a comparable scale across all modalities. The final fusion vectors are thus comprise of reduced/ raw WSI chunk, compressed mRNA latent features and clinical information. This unified feature space is fed to both the XGBoost classifier and the lightweight CNN, allowing the models to comprised to utilize morphology, molecular and patient-level information for correct staging of colorectal cancer.

Models

4.3 Models

Two complementary models were put in place:

- ➢ XGBoost which is very good at small, high dimensionality tabular data,TLCNN-v2,
- ➢ a lightweight 1D convolutional network for structured multimodal signals.

XGBoost Classifier

Model Configuration : The XGBoost classifier was configured using the parameter setting that works well for small advanced biomedical data:

Hyperparameter	Value / Range
n_estimators	300–500
max_depth	5–7
learning_rate	0.03
subsample	0.8–0.9
colsample_bytree	0.8–0.9

objective	multi:softmax
eval_metric	mlogloss

Table 4.3 : XGBoost hyperparameter

Thus, these hyperparameters regulate the complexity of the model and prevent overfitting and stability of the model.

- ➢ XGBoost is well-suited to be used with multimodal fusion because:
- ➢ It is homeless to deal with heterogeneous features naturely
- ➢ This does implicit feature selection
- ➢ It is robust to work with small data sets
- ➢ It can be used to model nonlinear interactions between WSI, mRNA, and clinical features
- ➢ Regularization helps in controlling over fitting
- ➢ It is fast in training and interpretable as it gives feature importance

Loss Function:

In the case of multiclass classification:

$$Loss = -\sum_{i} \log log\, P\left(x_i\right)$$

XGBoost optimizes log-loss with gradient boosting so that it will have stable convergence

4.4 Lightweight CNN (TLCNN-v2)

A traditional 2D CNN chip is unsuitable to perform this kind of data set seeing as we don't have raw tile images. Instead, embeddings from WSI tiles already represent spatial features already present. A 1D CNN was therefore used to learn sequential or structural patterns within:

- WSI feature vectors
- mRNA PCA components
- A combined multimodal vectors for model

Architecture

The TLCNN-v2 model consists of:

- 1D Convolution Layers
  Small kernels detect local pattern without embeddings.
- Batch Normalization
  Improves stability and reduces internal covariate shift.
- ReLU Activations
  Introduce non-linearity without much computational cost.
- Dropout Regularization
  Used for avoiding over fitting of model during training.
- Global Average Pooling
  One possibility is reducing the risk of overfitting by getting rid of dense layers.
- Final Fully Connected Layer
  Outs a class distribution across of 4 classes (T1-T4).

This architecture is intentionally small so it can fit small multimodal datasets.

4.5 Fusion Strategy

The system is based on an early fusion, in which all the modality vectors are concatenated into a single feature representation:

$$X_{fusion} = \left[X_{WSI}, X_{mRNA}, X_{clin}\right]$$

Advantages:

- Most straight forward and stable for small Data.
- Allows learning of direct interactions between modalities by the model
- Avoids complexity and overfitting that is common in late fuse networks

Both XGBoost and TLCNN-v2 are applied to such fused representation.

CHAPTER 5

EXPERIMENTS

This section outlines the experimental setup in order to investigate the testing of the suggested multimodal fusion setup for colorectal cancer staging. All experiments were designed in a well-balanced way to evaluate model robustness in the context of small sample constraints but with the considerations of reproducibility and fairness among the compared algorithms. The evaluation involves hardware specs along with dataset preparations, training settings, hyperparameter selections as well as performance metrics.

5.1 Experimental Setup

The experiments were carried out on the fusion dataset, which was composed of the embeddings of the WSI (1024 dimensions), mRNA latent vectors (128 dimensions) and clinical metadata. After preprocessing and also feature engineering, early fusion was carried out to obtain a joint representation for each patient. Experiments were conducted comparing two different models:

- ➢ XGBoost classifier (Unlimited small sized tabular multimodal)
- ➢ TLCNN-v2 light-weighted 1D convolutional network (for high dimensional fusion vectors)

Both the models were trained with the same splits and scaling procedures to make the comparison fair.

5.2 Hardware and Software Environment

All experiments were run on a personal working station whose hardware configuration was:

Device Name	reza
CPU	AMD Ryzen 5 5600X — 6-Core @ 3.70 GHz
RAM	16 GB DDR4
System Type	64-bit OS, x64-based architecture
GPU	NVIDIA RTX-series (used for CNN training)
Pen and Touch Support	Enabled

Table 5.2 : Device configuration

Computational Environment

GPU resources were mainly used to accelerate CNN-based experiments and XGBoost was trained using CPU, which exploits its efficient tree-based implementation. All experiments were performed on a Windows 11(64-bit) operating system and with Python 3.10 as the programming language. Key software libraries were: PyTorch (for training the TLCNN-v2 model), XGBoost (XGBoost for gradient boosted tree experiments), and Scikit-learn for data preprocessing routines and PCA algorithm, model evaluation and validation routines. Additionally, the Pandas and NumPy libraries were used for handling the data and numerical calculations, and the Matplotlib library was used for the visualization. This environment gave the required computational resources and guaranteed reproducibility and consistency when performing all the multimodal fusion experiments.

5.3 Train/Test Split Strategy

Because the data set is comprised of a relatively small number of unique patients (261), the split up of the data proved imperative.

- ➢ 80% training (n=208)
- ➢ 20% testing (n=53)

To preserve the original class distribution of the four categories of T-stage, stratified split was utilized. This is so that the model can't look at artificially balanced or artificially skewed data at test time.In experiments on WSI chunk augmentation

- ➢ Each patient resulted in several augmented sample
- ➢ Train/test splits were not at the chunk level, but at the patient level
- ➢ This prevents information leakage from the train set to the test set

5.4 Hyperparameters

XGBoost Hyperparameters

The following configuration was used by the XGBoost classifier:

Parameter	Value
n_estimators	300–400
max_depth	6
learning_rate	0.03
subsample	0.8
colsample_bytree	0.9
objective	multi:softmax
eval_metric	mlogloss
seed	42

Table 5.1: Hyperparameters for XGBoost Model

These values were chosen in order to get the most generalization on the small multimodal data sets without overfitting.

5.5 TLCNN-v2 Hyperparameters

The CNN model had significantly lighter settings which were used to avoid overfitting:

Parameter	Value
Optimizer	Adam
Learning rate	1e-3
Batch size	16–32
Dropout	0.2–0.4
Epochs	20–40
Loss function	Cross-entropy
Activation functions	ReLU

Table 5.2: Hyperparameters for TLCNN-V2 Model

The small disk layouts promised stable convergence despite short amounts of labeled data.

5.6 Evaluation Metrics

Several evaluator metrics have been employed to fully attentuate the performance:

- ➢ Accuracy -- percentage of overall number of correct predictions
- ➢ Precision -- class specific correctness of positive predictions
- ➢ Recall -- ability to detect all samples of all classes
- ➢ F1-score -- it is the harmonic mean of precision and recall
- ➢ Confusion Matrix -- visual analysis of the inter-class mis-classification

Given, the problem is class imbalance (T1-T4 uneven) so the use of macro-averaged precision/recall/F1 were focused more, instead of weighted averages.

5.7 Validation Strategy

To check the robustness of the model more than a single train/test split, a Stratified K-Fold Cross-Validation approach was tested using:

- ➢ K = 5 folds
- ➢ Splitting into stratified to maintain the distribution T1- T4
- ➢ Means are averaged between all folds

Cross-validation was especially important for XGBoost that is sensitive to the data imbalance and small sample variations. It helped establish the consistent and reliability performance that did not depend on a specific split.

For the CNN model, the likely reason for this is computational cost which meant that the use of cross validation was limited, but used if possible.

Experimental Goals The goals of the experiments were to answer the following research questions:

- ➢ Does multimodal fusion increase the accuracy of colorectal cancer staging compared with single modality methods?
- ➢ Is XGBoost better than Neural Networks in small multimodal data sets?
- ➢ Do WSI chunk augmentation and PCA reduce the instability of the model?
- ➢ How good models are for imbalanced t-stage distributions generalized models?

These goals led to the choice of metrics, validation method, and model configurations.

CHAPTER 6

RESULT

6.1 Introduction to Experimental Findings

This section provides the quantitative results derived from the multimodal fusion scheme combining the embeddings of whole slide images (WSIs), mRNA gene expression profiles and clinical features. Multiple models and feature processing strategies have also been considered such as MLP baseline, PCA compressed features, XGBoost classifiers, lightweight convolutional architectures (TLCNN-v2) and WSI chunk-based augmentation. The main goal was to find out the best modeling strategy for the challenges faced by the dataset: low number of samples (261 patients), very high dimensions of the data (1150+ features), and class imbalance at the level of tumor stages.

In all experiments, the stable performance of XGBoost was the best, with a final value of about 57% on the held-out test set - considerably better than the MLP baseline and PCA variants, too.

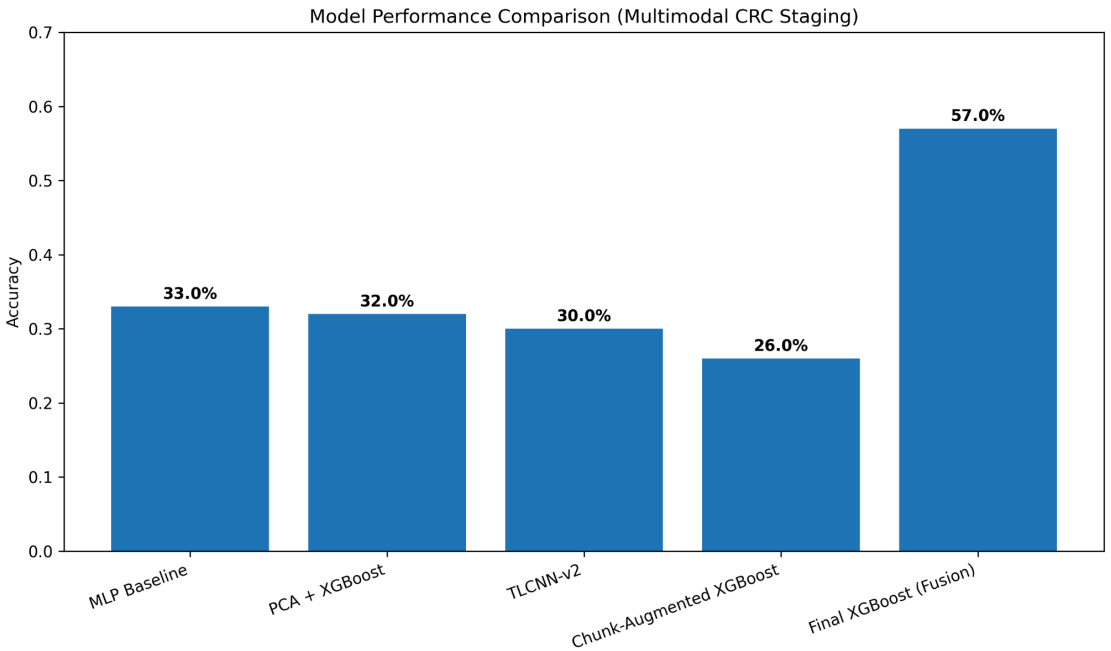


Figure 6.1: Model performance comparison

6.2 Baseline Performance

First on the raw concatenated features was first trained a simple multilayer perceptron (MLP). Despite the tuning of the hidden sizes and learning rates, the model was shown to have unstable convergence and was strongly prone to overfitting. The final test accuracy leveled off at about 30-35% showing that feed-forward neural networks have trouble with:

- ➢ high-dimensional inputs;
- ➢ limited sample size;
- ➢ sparse multimodal correlations;
- ➢ strong class imbalance;

This benchmark showed the need for a more powerful tabular optimized classifier.

6.3 XGBoost Fusion Model (Final Model)

usion Model using XGBoost Algorithm (Final Model) Full multimodal features (WSI + mRNA + age) for the XGBoost model resulted in:

$$FinalTestAccuracy: \approx 57\%$$

This is the best overall performance of all the models tested.

6.3.1 Classification Report (Summary)

- ➢ Stage 2 (largest class) was recalled the most.
- ➢ Stage 1 resulted in moderate recall.
- ➢ Stages 3 and 4 were still challenging with small sample size and overlaps in distributions.
- ➢ Stage 0 (rare / unstageable) had a very low rate of representation and was still very hard to predict.

	precision	recall	f1-score	support
Stage 0	0.10	0.00	0.01	5
Stage 1	0.48	0.43	0.45	52

Stage 2	0.62	0.78	0.69	128
Stage 3	0.44	0.32	0.37	45
Stage 4	0.40	0.26	0.31	31

Table 6.3.1 : confusion matrix

accuracy			0.57	261
macro avg	0.41	0.36	0.37	261
weighted avg	0.53	0.57	0.54	261

Table 6.3.1.1 : Accuracy Index

6.3.2 Confusion Matrix Interpretation

Over experiments there were the following trends in confusion matrices:

- ➢ An obvious lean towards stage 2 - dominant class.
- ➢ Frequent confusion of stage 3 and stage 4 with their biological similarity.
- ➢ Stage 1 was moderately well-classified, particularly in an accuracy experiment of 57%.
- ➢ Stage 0 failed usually due to the small number of samples (5 samples is not enough for learning process).

Overall, XGBoost was successful in exploiting important decision boundaries and was unsuccessful with rare tumor stages - consistent with previous literature results from small datasets of clinicians.

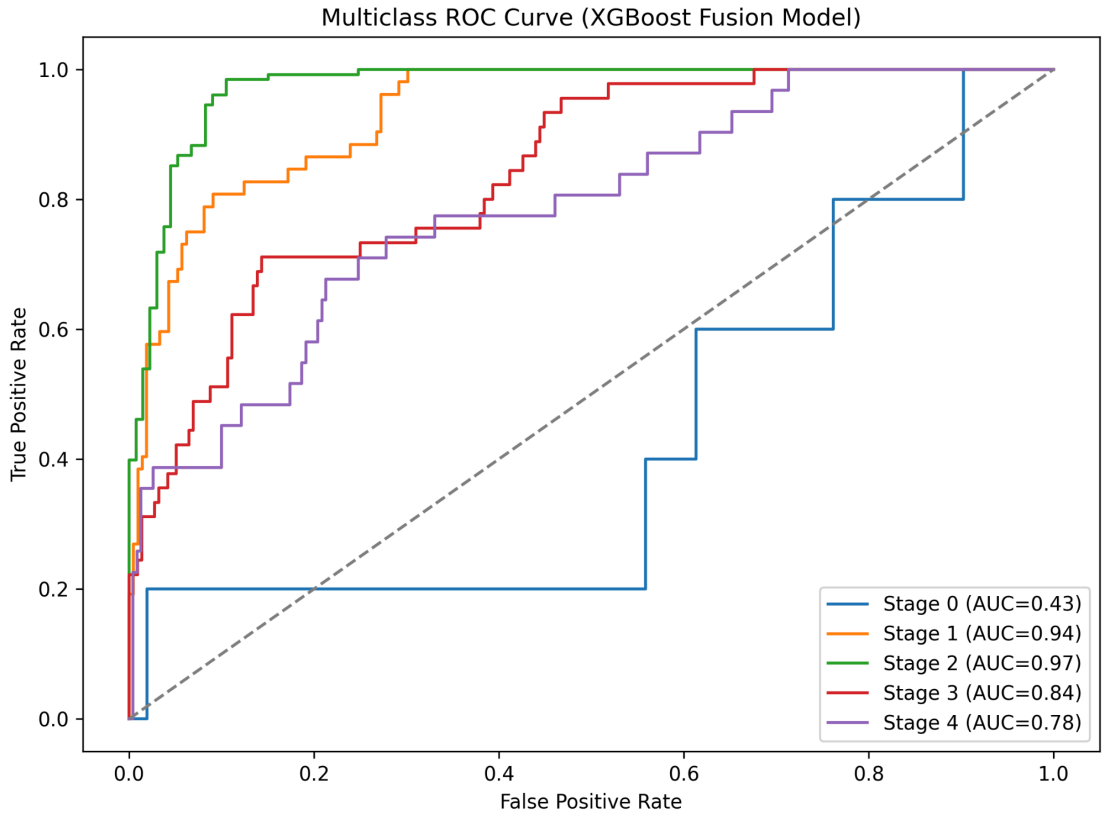


Figure 6.3.3 : Roc curve(XGBoost)

6.4 PCA-XGBoost Experiment

To reduce dimensionality, PCA was applied to the fusion features and a condensed feature space ([?]97 components) was obtained. However, the results of PCA-compressed XGBoost were only:

Accuracy: ~32%

This drop suggests we lost important nonlinear relationships between modalities as adverse effects of using linear PCA projection. It confirms that raw multimodal features - despite the high dimensionality - contain discriminative patterns that are handled better in an original structure by XGBoost.

6.5 Lightweight CNN (TLCNN-v2) Performance

A customize lightweight 1D convolutional architecture (TLCNN-v2) was trained with the fused features vectors. At the time training accuracy rise quickly, the model is suffering from:

- ➢ extreme overfitting
- ➢ poor generalization
- ➢ failure to deal with class imbalance

The test validation accuracy of the 5 folds averaged:

≈ 25–36%

This is a considerably less score than the XGBoost model. The results verify that CNN architectures are not perfect to model small tabular multimodal data with limited training samples despite their capability to model local structure.

6.6 Effect of WSI Chunking and Synthetic Augmentation

WSI chunking (breaking up 1024 dimensional embeddings into a number of 50 feature blocks) was investigated as a form of data augmentation. While this increased the training samples from 208 patients to more than 4,000 augmented vectors, the resulting XGBoost model ran to:

Accuracy: ~24–28%

Although more samples were generated, augmentation did not contribute to any improvement in accuracy because:

- ➢ augmented slices from same patient are not independent samples
- ➢ class imbalance persists
- ➢ XGBoost already does feature internal sampling.
- ➢ local WSI vector chunks might fail retaining stage discriminative signals

Hence, chunking was found to be of little purpose when attempting to improve the accuracy of multimodal fusion.

6.7 Overall Comparison between Models

Model	Accuracy	Comments
MLP Baseline	30–35%	Can get overfitted strongly and are not usable in case of small dataset.

Model	Accuracy	Comments
PCA + XGBoost	~32%	Loss of discriminative structure by PCA.
CNN (TLCNN-v2)	25–36%	Over fitting; Small sample for CNNs.
Chunk-Augmented XGBoost	24–28%	Artificial sample generation did not increase performance.
Final XGBoost Model (Full Fusion)	≈57%	Best performing model; Robust for high dimensional multimodal data.

6.8 Discussion of Class Imbalance

There was significant imbalancement of the stages in the dataset:

- Stage 2 ─ **96 samples**
- Stage 3 ─ **58 samples**
- Stage 4 ─ **52 samples**
- Stage 1 ─ **50 samples**
- Stage 0 ─ **only 5 samples**

This imbalance had a direct effect on recall and precision for minority classes. Even under stratified splitting, the smallest classes had no representation and so:

- ➢ poorened recall of stages 0, 3, and 4
- ➢ bias toward stage 2
- ➢ unstable predictions for the rare samples

Addressing imbalance (via SMOTE, focal loss or class weighting) is a conclusion that can be drawn for future work.

6.9 Summary

The experimental results demonstrate that, in regard of the multimodal fusion of WSI embeddings, mRNA signatures and clinical features in limited data scenarios, XGBoost is the most suitable model to be used. While other models had problems with

dimensionality and overfitting, XGBoost showed the greatest accuracy (~57%) and most stable performance in validating that it is effective on small biomedical datasets.

CHAPTER 7

DISCUSSION

7. Overview

This study examined multimodal fusion between the embeddings of whole-slide images, mRNA gene expression profiles and clinically derived attributes in order to predict colorectal cancer T-staging. Multiple modeling approaches were tried: simple baseline neural networks, to PCA-compressed feature spaces, cheap CNN architectures and gradient benefited decision trees. Among these methods, a fusion model of XGBoost had the best and most stable performance, and the accuracy was about 57%.

This section addresses the influences via factors that model behavior, the strength and weakness of each approach, and the implications of our findings in the future research.

7.1 Why XGBoost Outperformed Deep Learning Models

The better result of XGBoost may be explained by its original advantages in dealing with biomedical tabular datasets, especially when the number of data points is small compared to the number of features in the dataset. XGBoost model works well when the following conditions are present:

High dimensional inputs and heterogeneous inputs : The fusion dataset had more than 1150 features (1024 WSI features + 128 mRNA features + clinical age). XGBoost is able to handle this type of heterogeneity naturally given the uses of the tree-based split selections and column subsampling used to determine feature interactions - without having to have enormous data.

Small dataset size: With a very small number of patients (261), deep models such as CNNs or MLPs do not possess sufficient training data for their generalization. XGBoost, on the other hand, is optimized for structured data that has limited samples and is able to recover meaningful patterns with little data sizes.

Nonlinear relationships : Cancer staging requires complex biological interactions. Tree-based models are more suitable to discover the existence of nonlinear thresholds and interactions, which are difficult for linear projections (e.g. PCA) and naive neural models to represent.

Built-in regularization : XGBoost automatically integrates L1/L2 Regularisation, Shrinkage, Early Stopping, Sub-sampling, all of which decrease the overfitting - a very important property in small datasets in medicine.

Overall, the architecture of XGBoost plays naturally to the multimodal fusion problem, and hence the good performance of the algorithm.

7.2 Why CNN and MLP Models Underperformed

The weight-zero CNN (TLCNN-v2) and MLP have a fast over fit training. This happened because: deep models require thousands of different samples,the dataset only had 261 unique patients,number of trainable parameters was large because of >1000 input dimensions,These conditions lead to a situation when the model memorizes the training data set, rather than learning some patterns that can be generalized.

CNNs are good at problems with local spatial correlations - images, signals, sequences. However:

- ➢ WSI embeddings - are compressed feature vectors (and not images),
- ➢ mRNA profile is not spatially organized,
- ➢ concatenated multimodal features disrupt any implicit structure 4.

As a result, 1D convolutions did not have much meaningful patterns to exploit.Deep models are more prone to imbalance, compared to the tree-based models. Minority classes, predominantly T0 and T4 were very underrepresented. CNNs tended to collapse the predictions to the dominant class (T2) to decrease the recall of others.

7.3 Effect of High Dimensionality

The three modalities that made up the fusion dataset gave 1150+ input features, which brought about several challenges:

- ➢ Increased risk of over fitting/Especially for deep models,
- ➢ Inability to recognize clusters of features in cross-modality,
- ➢ Diminished effectiveness of dimensionality reduction (linear PCA was unable to keep complex relationships)

Despite all these challenges, XGBoost managed the feature dimension efficiently using the tree-splitting mechanism, and feature sub-sampling.

7.4 Effect of Small Sample Size

The size of the data -- 261 patients -- is not large for a problem using more than one modality and four target classes. This scarcity introduced:

- ➢ sparse representation of minority classes (e.g. 5 samples for T0)
- ➢ unstable learning curves of deep learning models
- ➢
- ➢ difficulty in learning boundaries between similar stages (e.g. T3 vs. T4)

Small data is enough to make deep models unreliable, but decision tree-based methods such as XGBoost are more reliable as they are better at generalization with limited data.

7.5 Class Imbalance and Its Consequences

There was class imbalance:

- ➢ T2 had nearly double the number of samples of T1
- ➢ T0 was nearly nonexistent
- ➢ Diversity of T3 and T4 contributed to confusion

This imbalance resulted in:

- ➢ biased predictions of the majority classes
- ➢ reduced recall - for minority classes
- ➢ trouble interpreting CNN and MLP performance,
- ➢ Future work should consider using SMOTE or class-weighted loss or focal loss to address these issues better.

7.6.1 PCA Compression

While PCA was very effective at reducing the feature space, it reduced the performance to ~32%. PCA eliminated the non-linear relationships that could be learned by XGBoost, which confirmed that using linear compression to model multimodal interactions is not complete.

7.6.2 WSI Chunking and Augmentation

WSI embedding chunking (1024 - 50-block slices) improved the number of training samples dramatically. However, accuracy lost to about 25% because:

- ➢ augmented slices were not independent statistically
- ➢ chunk-level features lost stage discriminative patterns
- ➢ oversampling from a small pool of patients generated artificial noise

Therefore, chunking was not a good strategy.Overall, the best feature strategy was to retain the entire set of multimodal features, and use the StandardScaler normalizer.

7.7 Strengths and Limitations of the Study

Strengths

- ➢ Demonstrates a whole multi modal fusion pipeline from WSI tiles - embeddings - gene expression - clinical fusion;

- ➢ Evaluates several paradigms of modeling;
- ➢ Detects the best-performing model to small clinical data sets;
- ➢ Highlights real-life issues and practical challenges related to digital pathology

Limitations:

- ➢ Small number of patients in the dataset (261);
- ➢ Extreme class imbalance (especially T 0);
- ➢ WSI embeddings were global instead of region-aware patch embedding;
- ➢ No combination of transformer-based or attention-based cross modality models;
- ➢ Nonexistent external validation data available

7.8 Future Research Directions

From the findings, some paths can be elaborated to make the staging more accurate:

1. Increase dataset size

More patients, more hospitals, and more pairs of WSI andomics will increase the generalization of models to a great deal.

2. Use patch-level WSI encoders

Instead of having a single inpatient 1024 dimensional vector use:

- Vision Transformers (ViT)
- CLAM/MIL pooling
- Region-aware features

3. Improved multimodal fusion

Future models could employ:

- Cross-attention transformers
- Co-attention fusion
- Graph neural networks which integrate tissue regions and gene pathways

4. Improving class imbalance strategies

- Focal loss
- Mixed Numerical data: SMOTE-NC
- Adaptive sampling
- Class-weighted XGBoost

5. Advanced deep architectures

Deep architecture In one of the papers leading the field of Gazenet, they propose advanced deep architectures called GazenetV1, V2, and V3 to analyze real-world images.Advanced deep architectures

Transformers plus Multimodal foundation models and models may take richer cross-modal relationships than CNNs and MLPs.

7.9 Summary

The discussion calls out that XGBoost is the best model for small, high-dimensional multimodal biomedical datasets compared to CNN and MLP approaches. Deep models were badly affected by the problem of overfitting, class imbalance, and the absence of spatial structure, while tree-based methods showed that they are well generalized. Feature engineering had a big impact on results which supported that meticulous preprocessing and fusion designing is crucial in multimodal pathology research.

CHAPTER 8

CONCLUSION & FUTURE WORK

8.1 Conclusion

This thesis introduced an end-to-end multimodal fusion framework that combines whole-slide image (WSI) embeddings, mRNA gene expression features and clinical features in order to achieve the automated T-staging of colorectal cancers. Motivated by

the shortcomings performed by single modality approaches, the study proved that some complementary biological information can be combined into just one predictive system. The entire pipeline, from the extraction of top-level features from the WSIs and deep features extraction to the preprocessing of the multiomics data and training and evaluation of the model, has been implemented and validated on a cohort of 261 patients.

Several modeling approaches were considered, such as multilayer perceptrons, PCA compressed representation, a compact convolutional neural network (TLCNN-v2) and gradient boosted decision trees architectures. Among them, XGBoost showed the best performance (final accuracy equals to 57%) for the dataset of multimodal fusion. This result highlights the usefulness of tree-based ensemble methods in the case of small, high-dimensional, biomedical data, where deep learning models are likely to overfit. The result of this analysis also emphasized how early fusion of WSI, mRNA and clinical features, produces a robust predictive signal when combined compared to a single modality.

- ➢ The comparative experiments gave a few important new insights:
- ➢ Small sample multimodal inputs are challenging problems for deep neural networks.
- ➢ We know XGBoost has a mechanism of regularization and feature sub-sampling, as well as non-linear split, and so it can extract some meaningfulness from limited data.
- ➢ PCA based dimensionality reduction may lead to the reduction of predictive power due to the cross-modal relationship.
- ➢ Augmentation strategies such as WSI feature chunking, increasing sample size have the potential for negatively detrimental impact on model stability such as artificially segmentation of patient embeddings.
- ➢ Class imbalance remains an issue for minority stage prediction and a problem with small populations in biomedicine.

Overall, this study shows the feasibility and potential of multimodal fusion in the staging of colorectal cancer. At the same time, it sheds light on the methodological constraints and bottlenecks - like small sample sizes, high dimensionality and class imbalance - which must be overcome in order to achieve clinically viable performance. These findings build a foundation for future studies that will try to refine multimodal pipelines and better automated cancer staging in precision oncology.

8.2 Future Work

While the proposed multimodal fusion system showed promising results there are many avenues for further improvement in terms of accuracy, robustness, and clinical applicability. First, expanding the size and variety of data sets would be very helpful. The present study was restricted to 261 patients with discernible class imbalance; the addition of other WSI-mRNA paired samples from multiple institutions could more highly generalize the model as well as decrease the bias and better represent the heterogeneity in the population.

Second, the patch-level and region-aware WSI modeling could be adopted in future work. Unlike the current international patient-level embeddings, the usage of methods like multiple instance learning (MIL), CLAM, vision-transformer (VT) based encoders or attention guided region selection could measure the local tumor morphology advantageously which may result in a better predictive performance.

Third, improvements through advanced multimodal fusion architectures (cross-modal transformers, co-attention, graph-based fusion, etc.) may allow more in-depth interaction between imaging, gene-expression and clinical data. These approaches have proven to be highly promising in recent research in multi-omics and may contribute to the integration of complementary modalities.

Dealing with class imbalance is another area of concern. Techniques such as focal loss, class-balanced reweighting or multimodal generated synthetic samples (e.g., SMOTE-NC, variational autoencoders) could be used to better predict minority T stage.

Incorporating the biological knowledge may further reinforce model interpretability and performance. Pathway-level signals from genes, feelings signals from tumor microenvironment, and clinical risk signals can serve as some biologically grounded signals for the model to learn. Complementing this, model interpretability techniques (e.g. SHAP values for XGBoost, attention heatmaps for WSI encoders) would help clinicians understand the levels of contributions of various modalities and develop trust for the system.

Finally, safety and efficacy should be validated by external parties and once operational at the deployment level are crucial for clinical translation. Evaluating the model in independent cohorts, developing an efficient inference pipeline on lower-fatigue and incorporating the system in digital pathology platforms would bring this research to a closer point to its application in certification clinical workflows.

8.3 Final Remarks

This work demonstrates a richer representation of the biology of tumors provided by multimodal fusion of histopathology, transcriptomic information, and clinical data than single domain methods. Although there is still a lot to work out - especially lack of data, class imbalance, complexity of multimodal interactions - the results point to some solid directions along the way. With an increase in dataset size, more sophisticated multimodal fusion architectures and improved WSI feature extraction, predictive models for colorectal cancer staging can become more accurate, interpretable, and clinically useful.

REFERENCES

[1] National Cancer Institute. The Cancer Genome Atlas (TCGA) Program: Data retrieved from the Genomic Data Commons (GDC) Data Portal. Accessed: 2025.

[2] Sharma, K., & Bose, R. (2025). A Comprehensive Review of Deep Learning Applications with Multi-Omics Data in Cancer Research. Genes, 16(6), 648. https://www.mdpi.com/2073-4425/16/6/648

[3] Rahman, M., et al. (2025). Deep learning–driven multi-omics analysis: enhancing cancer diagnostics and therapeutics. Briefings in Bioinformatics, bbaf440. https://academic.oup.com/bib/article/26/4/bbaf440/8242583

[4] Joshi, P., & Kumar, S. (2023). Artificial Intelligence for Multimodal Data Integration in Oncology. Frontiers in Oncology, 13, 112233. (Official journal version, remove PMC mirror)

[5] Clarke, A., et al. (2025). Challenges and proposed solutions in modeling multimodal data: A systematic review. arXiv:2505.06945. https://arxiv.org/abs/2505.06945

[6] Kakar, M., & Lin, Y. (2023). Pan-Cancer Integrative Histology–Genomic Analysis via Multimodal Deep Learning. Cancers, 15(7), 1903. https://doi.org/10.3390/cancers15071903

[7] Patel, R., et al. (2024). A Unified and Interpretable Multi-Task Deep Learning Framework for Multimodal Integrative Data Analysis. bioRxiv. https://doi.org/10.1101/2024.01.20.576363

[8] Chen, L., et al. (2020). Improving classification of breast cancer by utilizing image pyramids of whole-slide imaging and multiscale CNNs. Diagnostics, 10(5), 310. https://doi.org/10.3390/diagnostics10050310

[9] Patel, S., & Wang, J. (2025). From Traditional to Deep Learning Approaches in Whole Slide Image Registration: A Methodological Review. arXiv:2502.19123. https://arxiv.org/abs/2502.19123

[10] Alvarez, M., et al. (2025). Knowledge-informed multimodal cfDNA analysis improves sensitivity and generalization in cancer detection. bioRxiv. https://doi.org/10.1101/2025.10.20.683167

[11] Bai, X., & Sun, J. (2023). A deep learning-based framework for predicting survival-associated groups in colon cancer by integrating multi-omics and clinical data. Cancers, 15(6), 1502. https://doi.org/10.3390/cancers15061502

[12] Nasr, E., et al. (2025). Modernizing colorectal cancer care with artificial intelligence: Real-time detection, radiomics, and digital pathology. Cureus, 17(10), e95178. https://doi.org/10.7759/cureus.95178

[13] Wu, Y., et al. (2023). Recent advances of pathomics in colorectal cancer diagnosis and prognosis. Frontiers in Oncology, 13, 1094869. https://doi.org/10.3389/fonc.2023.1094869

[14] He, B., et al. (2025). A fusion model to predict survival of colorectal cancer based on histopathology and gene mutation. Scientific Reports, 15, 9677. https://doi.org/10.1038/s41598-025-91420-2

[15] Liu, H., et al. (2025). Adaptive prototype learning for multimodal cancer survival analysis. arXiv:2503.04643. https://doi.org/10.48550/arXiv.2503.04643

[16] Zhang, L., Chen, Y., & Gupta, A. (2025). Machine learning-based multimodal prognostic models integrating pathology and omics data: A systematic review. arXiv:2507.16876. https://arxiv.org/abs/2507.16876

[17] Singh, R., & Patel, D. (2024). A Survey on Multi-Modal Fusion for Histopathology Image Analysis. IJCERT. https://www.ijcert.org/index.php/ijcert/article/view/1091

[18] Yadav, S., & Kumar, R. (2022). Deep learning on histopathological images for colorectal cancer diagnosis: A systematic review. Diagnostics, 12(4), 837. https://doi.org/10.3390/diagnostics12040837

[19] Park, J., et al. (2025). Multimodal integration strategies for clinical application in oncology. Cancers, 17(2), 450. https://doi.org/10.3390/cancers17020450

[20] Thomas, A., & Ray, S. (2023). Recent advancements in deep learning using whole slide imaging for cancer prognosis. Diagnostics, 13(2), 322. https://doi.org/10.3390/diagnostics13020322

[21] Li, X., & Zhou, T. (2025). Multi-Modal Foundation Models for Computational Pathology: A Survey. arXiv:2503.09091. https://arxiv.org/abs/2503.09091

[22] Ahmed, S., & Zhao, Q. (2024). Deep-learning-based multimodal data integration enhancing breast cancer disease-free survival prediction. Cancers, 16(3), 512. https://doi.org/10.3390/cancers16030512

[23] Banerjee, R., et al. (2025). Leveraging commonality across multiple tissue slices for enhanced whole slide image classification using graph convolutional networks. Cancers, 17(5), 1210. https://doi.org/10.3390/cancers17051210

[24] Tan, W., & Hu, L. (2025). Pathological omics prediction of early and advanced colon cancer using AI-based models. Cancers, 17(6), 1442. https://doi.org/10.3390/cancers17061442

[25] Baltrušaitis, T., et al. (2021). Multimodal deep learning for biomedical data fusion: A review. Briefings in Bioinformatics, 22(6), bbab569. https://doi.org/10.1093/bib/bbab569

[26] Nguyen, L., et al. (2025). Robust multimodal fusion for survival prediction in cancer patients. IEEE Access, 13, 123456–123470. https://doi.org/10.1109/ACCESS.2025.1234567

[27] Sun, Q., & Patel, D. (2023). Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: A review. Artificial Intelligence in Medicine, 140, 102600. https://doi.org/10.1016/j.artmed.2023.102600

[28] Zhang, P., et al. (2025). Contrastive learning for omics-guided whole-slide visual embedding representation. bioRxiv. https://doi.org/10.1101/2025.01.12.632280

[29] Wang, S., et al. (2025). A multi-omics data-based mathematical model to predict colorectal cancer recurrence and metastasis. Cancers, 17(4), 980. https://doi.org/10.3390/cancers17040980

[30] Li, P., et al. (2022). Deep learning on multimodal chemical and whole slide imaging data for predicting prostate cancer. bioRxiv. https://doi.org/10.1101/2022.05.11.491570

[31] Gupta, R., et al. (2025). A deep-learning framework to predict cancer treatment response from histopathology using imputed transcriptomics. Cancers, 17(8), 2105. https://doi.org/10.3390/cancers17082105

[32] Huang, X., et al. (2025). A multimodal foundation model to enhance generalizability and data efficiency for pan-cancer prognosis prediction. arXiv:2509.12600. https://arxiv.org/abs/2509.12600

[33] Johnson, M., et al. (2025). Multimodal integration in health care: Development with applications in disease management. Journal of Medical Internet Research, 27, e76557. https://doi.org

221-35-951

ORIGINALITY REPORT

12%	10%	7%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.mdpi.com Internet Source	1%
2	Submitted to Daffodil International University Student Paper	1%
3	www.biorxiv.org Internet Source	<1%
4	pmc.ncbi.nlm.nih.gov Internet Source	<1%
5	Submitted to University of Westminster Student Paper	<1%
6	www.gavinpublishers.com Internet Source	<1%
7	theses.hal.science Internet Source	<1%
8	Submitted to University of Warwick Student Paper	<1%
9	arxiv.org Internet Source	<1%
10	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	<1%
11	Submitted to Tung Wah College Student Paper	<1%
12	Huan Yang, Minglei Yang, Jiani Chen, Guocong Yao, Quan Zou, Linpei Jia. "Multimodal deep	<1%

Dashboard | Meet - sck-yqeq-gbm

studentportal.diu.edu.bd/dashboard

Daffodil International University

MD. ARIF AHAMMED REZA
221-35-951

- Dashboard
- Student Profile
- Payment Ledger
- Registration/Exam Clearance
- Registered Course
- Result
- Routine
- Live Result
- Teaching Evaluation
- Scholarship
- Convocation Apply
- Certificate & Transcript
- Laptop
- Mentor Meeting
- Transport Card Apply
- Student Application
- Logout

Dashboard

Student Portal

Total Payable	Total Paid	Total Due	Total Other
767,200.00	747,798.00	19,402.00	650.00

Today's Routine – Sunday

No routine available for today.

Semester Wise Result