2024

Flexible multimodal learning for whole slide image analysis

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Thesis

FLEXIBLE MULTIMODAL LEARNING FOR WHOLE SLIDE IMAGE ANALYSIS

by

HARSH SHARMA

B.Tech., Indian Institute of Technology, Guwahati, 2019

Submitted in partial fulfillment of the

requirements for the degree of

Master of Science

2024

Approved by

First Reader

Vijaya B Kolachalama
Associate Professor of Medicine & Computer Science

Second Reader

Margrit Betke
Professor of Computer Science

Third Reader

Bryan Plummer
Assistant Professor of Computer Science

Acknowledgments

I would like to express my deepest gratitude to my guide, Prof. Vijaya Kolachalama, for his unwavering support, guidance, and mentorship throughout my thesis journey. His invaluable insights and encouragement have been instrumental in shaping my research and helping me grow as a scholar. I am also incredibly grateful to Yi Zheng, whose assistance and support were crucial whenever I needed it. His expertise and willingness to help have been invaluable, and I am truly thankful for his contributions to my work. I extend my heartfelt thanks to my lab mates, who have made this experience not only enlightening but also enjoyable. Their camaraderie, collaboration, and support have been essential to my success, and I am grateful for the friendships we have formed.

I would also like to thank my dear friends, Anwesha Saha and Nilesh Pandey, for their constant encouragement and support. Their belief in me and their unwavering friendship have been a source of strength throughout this journey.

Most importantly, I am eternally grateful to my parents for their unconditional love, support, and sacrifices. Their faith in me and their endless encouragement have been the driving force behind my achievements, and I am forever indebted to them.

Finally, I would like to express my sincere appreciation to Boston University for providing me with the opportunity to pursue my thesis. The resources, facilities, and academic environment have been exceptional, and I am honored to have been a part of this esteemed institution.

Harsh Sharma

CS Department

FLEXIBLE MULTIMODAL LEARNING FOR WHOLE SLIDE IMAGE ANALYSIS

HARSH SHARMA

ABSTRACT

Whole slide imaging (WSI) has revolutionized digital pathology, enabling the digitization of entire tissue samples on glass slides into high-resolution images, providing a detailed view of tissue morphology essential for analysis and interpretation. However, practical challenges arise due to the varying availability and cost of different slide staining techniques. Hematoxylin and Eosin (H&E) staining is commonly performed and cost-effective, but it offers limited information compared to more expensive and less frequently conducted immunohistochemical (IHC) staining, which provides specific molecular insights critical for diagnosing conditions like Alzheimer's disease (AD) and chronic traumatic encephalopathy (CTE).

Addressing the need for a comprehensive analysis in the absence of complete modality sets, this research introduces "MultiStainKD ," a novel framework designed to handle multimodal WSI data effectively, even when certain modalities, such as IHC, are missing. MultiStainKD employs a combination of computational architectural design and training strategies, optimizing the use of available data to enhance diagnostic accuracy. It utilizes a blend of convolutional and transformer-based components to capture the nuanced interplay of local and global tissue features, thus ensuring robust performance without the necessity for costly feature generation processes traditionally used to compensate for missing modalities.

Our extensive testing shows MultiStainKD 's superior capability in managing missing modalities and setting new benchmarks in AD and CTE prediction accuracy.

By leveraging the inherent complementary information within accessible stains and slides, MultiStainKD demonstrates its practical value and effectiveness, aligning with the goal of maintaining diagnostic quality despite the unavailability of comprehensive staining modalities. This advancement signifies an important step forward in multi-modal learning for digital pathology, potentially enhancing diagnostic accuracy and supporting patient care when full modality data is not available.

Contents

List of Tables

List of Figures

List of Abbreviations

ABeta	.....	Amyloid Beta Whole Slide
ABMIL	.....	Attention-Based Multiple Instance Learning
AD	.....	Alzheimer's Disease
Adam	.....	Adaptive moment estimation
AT8	.....	Tau Stain Whole Slide
AUC	.....	Area Under the Curve
BCE	.....	Binary Cross-Entropy
Biel	.....	Bielschowsky Silver Stain Whole Slide
CLAM	.....	CLuster Attention Multiple instance learning
CLS	.....	Class Token
CNN	.....	Convolutional Neural Network
CTE	.....	Chronic Traumatic Encephalopathy
CV	.....	Computer Vision
DL	.....	Deep Learning
DS	.....	Digital Staining
DTFD	.....	Double Tier Feature Distillation
FC	.....	Fully Connected(Layer)
FCN	.....	Fully Connected Network
FN	.....	False Negative
FP	.....	False Positive
GCN	.....	Graph Convolution Networks
GNN	.....	Graph Neural Network
GTP	.....	Graph Transformer for Pathology
HSV	.....	Hue, Saturation, Value Image
LHE	.....	Hematoxylin & Eosin Stain Whole Slide
MIL	.....	Multiple Instance Learning
MLP	.....	Multilayer Perceptron
MRI	.....	Magnetic Resonance Imaging
NLP	.....	Natural Language Processing
Pdropout	.....	Progressive Dropout
PPEG	.....	Pyramid Position Encoding Generator
ReLU	.....	Rectified Linear Unit
ResNet	.....	Residual Network
RGB	.....	Red, Green, Blue Image

RMSProp	.....	Root Mean Square Propagation
ROC	.....	Receiver Operating Curve
SGD	.....	Stochastic Gradient Descent
TN	.....	True Negative
TP	.....	True Positive
TransMIL	.....	Transformer based Multiple Instance Learning
ViT	.....	Vision Transformer
WSI	.....	Whole Slide Image

Chapter 1

Introduction

Whole slide images (WSIs) have transformed digital pathology by enabling the digitization of entire tissue samples into high-resolution images. WSIs, open up new avenues for research and clinical applications, enabling the development of digital diagnostic tools and facilitating collaboration among pathologists. However, the sheer size and complexity of WSIs present several challenges in terms of storage, processing, and analysis. The traditional Computer vision algorithms struggle to handle the massive amount of data contained in these images, making it important to develop novel approaches to handle this type of data.

Digital staining plays a crucial role in the field of histology (the study of tissues), enabling the visualisation and identification of specific cellular and tissue components. By applying various staining techniques to tissue samples, pathologists highlight and differentiate different structures or cellular structures present in the WSI, making it easier to detect and diagnose pathological conditions. In the context of neurodegenerative disorders like Alzheimer's disease (AD) and chronic traumatic encephalopathy (CTE), it is particularly important to identify the important pathological features associated with the conditions.

Hematoxylin and eosin, also known as H&E, is one of the stains that is most frequently used in histology. It emphasizes the tissue's general morphology, which consists of hematoxylin staining the cell nuclei in blue and eosin staining the cytoplasm and extracellular matrix in pink. H&E staining provides a foundation for the

initial assessment of tissue structure and can reveal abnormalities in cell morphology and organisation.

Bielschowsky Silver stain is used for the purpose of visualising nerve fibres, axons, and neurofibrils within the central nervous system. This stain is very helpful in determining whether or not *senile plaques*, which are a typical feature of Alzheimer's disease, are present in the brain. Providing a full evaluation of the pathological changes that have occurred in the brain. Pathologists often combine this staining with other stains to conduct a thorough evaluation of the brain's pathological changes.

The Amyloid Beta (ABeta) stain is yet another important stain that becomes relevant while discussing Alzheimer's disease. Amyloid-beta plaques are extracellular deposits of the amyloid-beta protein that begin to build in the brains of people who have Alzheimer's disease (AD). The stain specifically targets these plaques. Using ABeta stain, it is possible to visualise and quantify these plaques, which provides vital information regarding the amount and distribution of this clinical characteristic.

Pathologists frequently use the Tau (AT8) stain to determine the presence of neurofibrillary tangles, another important characteristic of Alzheimer's disease. One of the antibodies known as AT8 is able to identify *hyperphosphorylated tau* protein, which is responsible for the formation of neurofibrillary tangles, which are intracellular aggregates. Patients with Alzheimer's disease are more likely to experience cognitive impairment due to the presence of these tangles, which disrupt neuronal activity. The AT8 stain is used to detect and evaluate the severity of tau pathology in the brain, as well as its distribution throughout the brain.

Digital staining techniques have greatly enhanced the ability to study WSIs and diagnose neurodegenerative disorders like Alzheimer's disease and CTE. By leveraging the different types of WSIs for a single patient, we can analyse and predict with much better accuracy the pathology of the patient.

In recent years, computational methods for analysing WSIs in the context of neurodegenerative disorders have advanced. These methods aim to automate the detection of pathological features, such as amyloid plaques and neurofibrillary tangles, to aid in the diagnosis and research of conditions like Alzheimer's disease (AD) and Chronic Traumatic Encephalopathy (CTE).

One of the most widely used approaches for WSI analysis is based on Multiple instance learning (MIL). MIL methods have shown a lot of success in various computer vision tasks, including image classification and object detection. In the context of WSIs, MILs are typically employed by taking patches from the image as different instances; the features are extracted from each patch using a CNN based module that can be pretrained on the whole slide image or just using the original WSI. We then aggregate these features to generate predictions at the slide level.

Patch-based CNN methods suffer from a loss of spatial context and relationships between patches. By focusing on individual patches in isolation, these methods fail to capture the spatial dependencies present in the tissue. To address this issue, researchers have explored the use of graph neural networks (GNNs) for WSI analysis. GNNs can model the spatial relationships between patches by representing the WSI as a graph, where nodes correspond to patches and edges represent their spatial connections. This enables the incorporation of contextual information as well as the acquisition of more meaningful representations of tissue morphology.

The use of transformer-based architectures is another promising trend in WSI analysis. Transformers, which have revolutionised natural language processing (NLP), have shown great potential in Computer Vision (CV) tasks. They can capture long-range dependencies and learn global context. Methods like TransMIL (Transformer-based Multiple Instance Learning) (Shao et al., 2021) have been proposed to leverage the power of transformers for WSI classification. However, the computational cost

associated with applying transformers to the entire WSI remains a significant challenge, and efficient methods for handling the large-scale nature of these images are still an active area of research.

Several limitations and challenges persist in WSI analysis, despite its progress. Handling multimodal data is one major challenge. Real-world scenarios use multiple types of data, such as different stains or imaging modalities, to diagnose the patient. Integrating and leveraging the complementary information from these multimodal sources can provide a more comprehensive understanding of the disease pathology. However, current methods often struggle to effectively combine and analyse multimodal data, leading to suboptimal performance and limited interpretability.

Another significant challenge when dealing with models that can incorporate multiple modalities is the absence of certain modalities of data. In clinical settings, it is common to have incomplete or missing WSIs for some patients due to various reasons, such as tissue availability, quality issues or the cost of creating the data. Handling missing data is crucial for developing robust and generalizable models that can operate in real-world conditions. While some recent works have attempted to address this issue (Chen et al., 2021) for survival prediction using multimodal data with missing modalities, there is still a need for more efficient and effective approaches to deal with missing data in WSI analysis.

My thesis's main contributions are as follows:

- A novel deep learning architecture that can effectively handle missing slides and make accurate diagnostic predictions based on the available information.

- Techniques for extracting discriminative features from WSIs that are informative for the diagnostic task and robust to variations in staining quality and tissue preparation.

- Methods for integrating multimodal information from multiple slides to improve diagnostic accuracy by leveraging the complementary nature of different stains.
- Extensive evaluation of the proposed method on a diverse patient dataset, demonstrating its robustness, generalisation ability, and potential for clinical application in diagnosing AD and CTE.
- Beating the current state of the art in multimodal methods for whole slide imaging

In this thesis, the problems of missing data and combining different types of information in WSIs for diagnosing AD and CTE are looked at. The goal is to make digital pathology better and help pathologists make correct and dependable diagnostic decisions in the face of real life scenarios of data availability.

Chapter 2

Literature Review

2.1 Deep Learning on Whole Slide Image

In recent years, the analysis of whole slide images (WSIs) has widely adopted deep learning techniques. However, the unique characteristics of WSIs pose significant challenges compared to natural images, requiring special treatment and innovative approaches.

One of the primary challenges in working with WSIs is their enormous size. At high magnifications, such as 20×, a single WSI can contain an average of 15,000 patches of size 256×256 pixels. This large number of patches far exceeds the typical sequence length used in natural language processing tasks, where word embeddings have a maximum sequence length of around 2048 [(Vaswani et al., 2023)]. Due to the high space complexity of WSI bags, it is no longer possible to directly use popular deep learning architectures like Vision Transformers (vision-transformer), which have been successful in natural image analysis.

To address this challenge, researchers have explored various approaches to handling WSIs effectively. One common approach is to convert the WSI into a bag of patches and treat the problem as a multiple instance learning (MIL) task. Each WSI in MIL functions as a bag, treating the patches within it as instances. Attention-based MIL like ABMIL [(Ilse et al., 2018)] and DTFD MIL [(Zhang et al., 2022b)] can be used with this formulation. Other MIL-specific algorithms can also be used [(Campanella et al., 2019),(Feng and Zhou, 2017),(He et al., 2012),(Hou et al., 2016),(Li

et al., 2021), (Lu et al., 2021)]. Another approach used more recently is sampling the patches used more recently by SparseConvMIL [(Lerousseau et al., 2021)]. Sampling the image's patches significantly reduces the number of patches, making it easier to manage deep learning models. Sampling, however, comes at the cost of losing fine-grained details and potentially important diagnostic information. Another approach to handling WSIs is to preserve their structural information by employing graph neural network (GNN) based methods like FS-GCN-MIL(Zhao et al., 2020), MIL-GNN (Tu et al., 2019), and GTP(Zheng et al., 2022). GNNs can capture the spatial relationships between patches and model the inherent structure of the WSI. By representing the WSI as a graph, where patches are nodes and their spatial connections are edges, GNNs can learn meaningful representations that incorporate both local and global context. However, these methods are not trivial at effectively combining information from multiple WSIs, limiting their ability to capture complex relationships and dependencies across different slides.

Despite these challenges, deep learning has shown promising results in various tasks related to WSI analysis, such as tumour detection, grading and diagnosis. Researchers have proposed novel architectures and training strategies specifically tailored to the unique characteristics of WSIs. We'd build on some of these methods to extract information from the Whole Slides and fuse it with information from other modalities.

2.2 MultiModal Learning

Multi-modal learning has gained significant attention in recent years due to its ability to leverage information from multiple data sources and modalities. In the field of biomedical research, multi-modal learning has shown promising results by combining different imaging techniques, such as Magnetic Resonance Imaging (MRI), to improve

diagnostic accuracy and patient outcomes.

Fully Convolutional Networks (FCNs) [(Long et al., 2015), (Ronneberger et al., 2015)] have been widely used in multi-modal learning for medical image segmentation and feature extraction. FCNs have demonstrated excellent performance in capturing local patterns and semantic information, making them a popular choice for processing biomedical data. However, the inductive bias of convolution, which focuses on local dependencies, makes it difficult for FCNs to explicitly model long-range dependencies. This limitation can be particularly relevant in the context of whole slide images, where capturing global context and relationships between distant regions is crucial. One notable work is HyperDenseNet, proposed by Dolz et al. [(Dolz et al., 2019)]. HyperDenseNet builds dual deep networks for different modalities of MRI and links features across these streams to capture complementary information. Tseng et al. [(Tseng et al., 2017)] take a step further by designing a novel encoder-decoder structure to capture and fuse low-level and high-level features from multiple modalities. They then fuse the results from each branch to generate the final output.

Multi-modal learning has explored transformer-based architectures to capture more complex relationships and long-range dependencies. Vision Transformers (ViT) [(Dosovitskiy et al., 2021)] have shown promising results in capturing global context and modelling long-range dependencies. However, directly applying transformers to whole slide images is computationally challenging due to the large number of patches and the quadratic complexity of self-attention mechanisms.

(Chen et al., 2021)] propose a framework for whole slide images using co-attention to fuse information from two input streams. Co-attention mechanisms allow the model to attend to relevant features from different modalities and learn cross-modal interactions. However, the computational complexity of co-attention can increase significantly with the number of modalities, making it challenging to scale to higher-

dimensional multi-modal data.

Despite their progress, many existing methods in multi-modal learning assume that a complete set of data from all modalities is available. However, in real-world scenarios, particularly in medical applications, the full set of data might not always be accessible. This poses a challenge for multi-modal learning methods that rely on the completeness assumption.

As multi-modal learning continues to evolve, developing methods that can effectively handle incomplete data and scale to higher-dimensional multi-modal inputs remains an important research direction. MultiStainKD represents a step forward in this direction, providing a framework that can work with multiple modalities efficiently and handle scenarios where the complete set of data may not be available. By addressing these challenges, MultiStainKD aims to contribute to the advancement of multi-modal learning in biomedical research and beyond.

2.3 Multimodal Learning with missing Information

Dealing with missing modalities is a critical challenge in multi-modal learning, particularly in healthcare settings where incomplete data is a common occurrence. Various approaches have been proposed to address this issue, each with its own strengths and limitations.

One notable work in this area is M3 Care [(Zhang et al., 2022a)], which focuses on handling missing information in healthcare settings by imputing data from other patients. While this approach has shown promise in certain domains, it may not be feasible in the context of whole slide images. Imputing slide-level information by looking at other patients is a highly complex task due to the unique characteristics and variability of histopathological data. Synthesising such high-dimensional and detailed information accurately is a significant challenge.

Generation-based methods, such as those proposed in [(Ma et al., 2021),(Vasco et al., 2022)], have gained popularity in handling missing data. These methods aim to generate missing modalities or features based on the available data. However, the computational cost associated with these approaches can be significant, particularly when dealing with large-scale and high-dimensional data such as whole slide images. The computational complexity increases exponentially with the number of missing modalities, making it challenging to apply these methods in practice.

More recently, mmFormer[(Zhang et al., 2022c), for example, generated features from scratch using a convolutional encoder-decoder architecture. [(Dorent et al., 2019),(Havaei et al., 2016)] takes a step further by introducing a multimodal variational auto-encoder to benefit incomplete multimodal segmentation by generating missing modalities. These approaches demonstrate the potential of generative models for addressing missing data challenges.

However, our proposed technique, MultiStainKD , takes a different approach to handling missing modalities. Instead of relying on generating missing features, MultiStainKD derives the necessary information from the available modalities. By leveraging the complementary information present in the existing stains and slides, MultiStainKD can effectively capture the relevant patterns and relationships needed for accurate prediction.

The key advantage of MultiStainKD 's approach is that it eliminates the need for computationally expensive feature generation processes. MultiStainKD can efficiently handle missing data by directly utilizing the available modalities, eliminating the added complexity and computational burden of synthesizing high-dimensional features.

The architecture of MultiStainKD adapts adaptively to the available modalities, enabling it to make accurate predictions even in the absence of certain stains or

slides. MultiStainKD can capture both local and global dependencies by using both convolutional and transformer-based components. This lets it make good use of the information in all the available modes.

This approach not only reduces the computational cost but also enhances the model's robustness and practicality in real-world scenarios where incomplete data is prevalent. MultiStainKD can easily apply to a wide range of datasets and settings, even when the complete set of modalities is not available, by eliminating the reliance on feature generation.

Moreover, MultiStainKD 's ability to derive information from the available modalities allows it to capture the inherent relationships and dependencies between different stains and slides. By leveraging the complementary nature of the modalities, the model can potentially uncover insights and patterns that methods that solely rely on feature generation may miss.

Chapter 3

Problem Formulation

Building upon the introduction, where we discussed the importance of multi-stain whole slide image (WSI) analysis in the context of diagnosing neurodegenerative disorders, we now formally define the problem statement and objectives of this thesis.

Given a set of N patients, each patient may have a different number of available WSIs, specifically ABeta, AT8, Biel, and LHE stains. The goal is to predict whether each patient has a positive or negative diagnosis for Alzheimer's disease (AD) and Chronic Traumatic Encephalopathy (CTE), which can co-occur in the same patient. The problem can be formulated as follows:

$$\hat{y}_i = f(X_i^{ABeta}, X_i^{AT8}, X_i^{Biel}, X_i^{LHE}), i \in 1 \dots N \tag{3.1}$$

where $\hat{y}_i$ is the predicted label for the i-th patient, and f is a function that takes as input the available WSIs for the i-th patient. The input slide X_i^{slide} can be represented as:

where,

$$X_i^{slide} \in \begin{cases} \phi, & \text{if: slide missing} \\ \mathbb{R}^{w \times h}, & \text{otherwise} \end{cases} \tag{3.2}$$

Here, $\mathbb{R}^{w \times h}$ represents the input slide dimension, where w and h are the width and height of the slide, respectively. In cases where a particular slide is missing for a patient, we mask the corresponding input.

The main challenges to this problem include:

- Handling missing slides: The proposed method should be able to effectively handle cases where one or more slides are missing for a patient, while still making accurate predictions based on the available information.

- Extracting relevant features: The model should be capable of extracting discriminative features from each available slide that are informative for the diagnostic task while being robust to variations in staining quality and tissue preparation.

- Integrating multimodal information: To improve diagnostic accuracy, the method should effectively combine the information from multiple slides, leveraging the complementary nature of different stains.

- Scalability and efficiency: Given the large size of WSIs and the potential for a high volume of patient cases, the proposed approach should be computationally efficient and scalable to handle real-world workloads.

To address these challenges, we propose a novel deep learning-based approach that incorporates techniques for handling missing data, extracting relevant features from WSIs, and integrating multimodal information. We will extensively evaluate the proposed method on a diverse dataset of patients to assess its robustness, generalization ability, and potential for clinical application in diagnosing AD and CTE.

By addressing the challenges associated with missing data and multimodal information integration in the context of diagnosing AD and CTE using WSIs, this thesis aims to contribute to the advancement of digital pathology and assist pathologists in making accurate and reliable diagnostic decisions.

Chapter 4

Data Characteristics

The distribution of the whole slide images used in this study is shown in Table 4.1. The dataset includes a total of 712 WSIs, with each type of stain available in varying quantities. AT8 staining is the most common; it is present in 98% of patients. The ABeta staining technique is present in 42% patients. The Biel staining method is present in 36% of patients. Finally, LHE staining is present in 45% of patients.

Slide Type	Count	% of patients
ABeta	135	42.2
AT8	316	98.7
Biel	116	36.2
LHE	145	45.3
Total	712	-

Table 4.1: Distribution of all the whole slide images

Each slide type contributes unique information to the understanding of Alzheimer's disease pathology.

- **LHE** slides give a general view of the structure and appearance of tissues, helping to find important areas and possible abnormalities.
- **Biel** slides are important for showing how nerve fibers and *neurofibrillary tangles* are spread out in the brain.
- **ABeta** slides emphasize amyloid-beta plaques, another key feature of Alzheimer's disease pathology. The existence and spread of these plaques can give us information about how much the disease has developed.

- **AT8** slides target neurofibrillary tangles. The presence and distribution of tau tangles can help determine the disease's progression and understand its effects on brain cell activity.

By combining the information from these different slide types, in our method, we learn to classify the patient in the presence of at least one type of data.

Chapter 5

Multistain prediction, MultiStainKD

MultiStainKD is a supervised model that takes in inputs from multiple Whole slide images corresponding to each individual patient. Fig. 5·1 defines the general pipeline. Features were generated through a Resnet50 based architecture. The section on Architecture discusses architecture. The section on Loss Function discusses how we modified the loss function to account for data imbalance. The section on Cleaning and preprocessing discusses different ways we cleaned the data and extracted features from relevant information on the whole slide.

5.1 Architecture

The architecture of MultiStainKD is a Late Fusion Technique. Figure 5·1 describes the architecture. We first pass the model through individual Modality Encoders and also through a shared encoder. We intend to replace any missing information with features aggregated from other modalities. Then pass each feature vector to their own classifiers. The next sections detail the different components of the architecture.

5.1.1 Self Attention Block

The self attention mechanism, inspired by the TransMIL algorithm (Shao et al., 2021), plays a crucial role in identifying the important patches within each slide. The process begins by initialising a CLS token and concatenating it with the initial features. We then apply linear attention to the features and concatenate the result with the original

features. We add position embeddings (Algorithm 4) to the concatenated features and then perform another linear attention operation. We again concatenate the resulting features with the previous features and then pass them through a layer normalization step. Finally, the CLS token is separated from the modified features and the output features are returned. Refer to Algorithm 1 for a detailed description of the steps involved in the self attention block.

Algorithm 1 Self Attention Block

Input: Features X
Output: Modified Feature X'
$X \leftarrow X_i$ {Initial features}
CLS $\leftarrow$ Initialize CLS token
$X \leftarrow [X, \text{CLS}]$ {Add CLS token to features}
$X_{SA1} \leftarrow \text{LinearAttention}(X)$
$X \leftarrow [X, X_{SA1}]$, Concatenate
$X \leftarrow \text{PositionEmbedding}(X)$
$X_{SA2} \leftarrow \text{LinearAttention}(X)$
$X \leftarrow [X, X_{SA2}]$, Concatenate
$X \leftarrow \text{LayerNorm}(X)$
$\text{CLS}, X' \leftarrow X[:, 0], X[:, 1:]$ {Output features}
return X'

This self attention block allows the model to capture the relationships between different patches within a slide and emphasise the most informative ones. By iteratively applying linear attention and incorporating position embeddings, the model can effectively learn and assign importance to different regions of the slide. The CLS token inclusion allows the model to have a global representation of the slide, which can be useful for downstream tasks. Figure 5·3 provides a visual representation of the self attention mechanism. Overall, the self attention mechanism enhances the model's ability to focus on the most relevant information within each slide.

5.1.2 Keyless Attention

The keyless attention mechanism, as depicted in Figure 5·4, is inspired by [(Long et al., 2018)]. It enables the model to adaptively assign importance to different features based on their relevance to the task at hand. The process consists of two main steps. In Step 1, the input features are passed through a multi-layer perceptron (MLP) block, which applies a series of transformations to learn complex feature representations. The output of the MLP block serves as the input for Step 2, where a softmax operation is applied to generate attention weights. The model then multiplies these attention weights element-wise with the original input features, enabling it to prioritize the most relevant features and suppress the less significant ones. Finally, a fully connected (FC) layer processes the weighted features to obtain the final feature representation.

The keyless attention mechanism offers several advantages. Firstly, it eliminates the need for explicit key-value pairs, simplifying the attention computation. Second, the model can effectively capture the most informative features and make a refined representation for further processing by learning attention weights through the MLP and softmax operations. This adaptive weighting scheme allows the model to dynamically adjust its attention based on the specific characteristics of the input data.

Moreover, keyless attention is lightweight and requires significantly fewer parameters compared to traditional attention mechanisms. This makes it particularly suitable for scenarios with limited training data, as it is less prone to overfitting. The reduced parameter count also contributes to improved computational efficiency and faster inference times.

By focusing on the most relevant features and suppressing noise, keyless attention can improve the discriminative power of the learned representations and enhance the overall performance of the model.

5.2 Loss function

Our dataset exhibited significant class imbalance, with the CTE positive class accounting for 60% of the data and the AD positive class accounting for just 25% of all cases. To address this imbalance, we employed weighted Binary Crossentropy as the loss function. We also want our model to be agnostic about missing modalities, for that we use a Knowledge Distillation Loss(KDLoss)

5.2.1 Weighted Binary Cross Entropy

The weighted Binary Crossentropy loss is defined as follows:

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N}\left[w^{+}y_i\log(\hat{y}_i) + w^{-}(1-y_i)\log(1-\hat{y}_i)\right] \tag{5.1}$$

where N represents the total number of observations, y_i denotes the actual label ($y_i \in \{0, 1\}$), and $\hat{y}_i$ denotes the predicted class. The weights w_i^+ and w_i^- are allocated to positive and negative samples, correspondingly, and are dynamically computed based on the class distribution within the dataset.

To calculate the class weights, we first determine the total number of samples N and the count of positive class samples C^+. The count of negative class samples C^- is then derived as $C^- = N - C^+$. The weights are inversely proportioned to their class counts to balance the scale between the frequently and infrequently occurring classes. This process is mathematically encapsulated in Algorithm 2.

5.2.2 KDLoss

For our knowledge distillation branch, we aim to ensure that the shared encoder produces similar outputs when faced with missing information. To achieve this, we employ an L1 loss between the outputs of the shared encoder and the modality-specific encoder. The L1 loss is a commonly used distance metric that measures the absolute

Algorithm 2 Calculate class weights w^+ and w^-

Require: Total number of samples N, number of positive samples C_+
Ensure: Weights w^+ and w^- for positive and negative classes, respectively
$C_- \leftarrow N - C_+$ {Calculate the number of negative samples}
$w^+ \leftarrow \frac{N}{C_+}$ {Calculate the weight for positive samples}
$w^- \leftarrow \frac{N}{C_-}$ {Calculate the weight for negative samples}
$all\_weights \leftarrow \sqrt{(w^+)^2 + (w^-)^2}$ {Compute the norm of the weights}
$w^+ \leftarrow \frac{w^+}{all\_weights}$ {Normalize the weight for positive samples}
$w^- \leftarrow \frac{w^-}{all\_weights}$ {Normalize the weight for negative samples}

difference between two vectors or matrices. By minimizing the L1 loss, we encourage the shared encoder to learn representations that are consistent with those learned by the modality-specific encoder.

The knowledge distillation loss (KDLoss) is computed as follows:

$$\text{KDLoss} = \sum_{i}^{m} \text{L1 Loss}(X\text{specific}, X_{\text{shared}}), \text{where modality} \neq \phi \tag{5.2}$$

Here, m represents the number of modalities, and X_{specific} and X_{shared} denote the outputs of the modality-specific encoder and the shared encoder, respectively. The L1 loss is calculated between the corresponding outputs of the two encoders for each available modality.

It is important to note that the KDLoss is only computed for modalities that are present and not missing. This selective application of the loss allows the model to focus on learning from the available information and avoids penalizing the shared encoder for missing modalities.

By minimizing the KDLoss, the shared encoder is encouraged to capture the essential information from the available modalities and produce representations that are consistent with those learned by the modality-specific encoders. This knowledge distillation approach helps the shared encoder to effectively handle missing information and maintain robust performance even when certain modalities are absent.

The KDLoss serves as a regularization term in the overall training objective, complementing other loss functions such as the classification loss or reconstruction loss. By jointly optimizing the KDLoss and other relevant losses, the model learns to leverage the available modalities while being resilient to missing information.

Using both these the loss for our model becomes:

$$\text{Loss} = \text{Weighted BCE} + \lambda \times \text{L1 Loss} \tag{5.3}$$

λ is a hyperparameter

5.3 Data Cleaning and Preparation

This section outlines the steps taken to clean and prepare the data for further analysis, including background separation, patching and feature generation. This section is inspired from CLAM(Lu et al., 2021)

5.3.1 Background Separation

In order to properly analyze the image and also reduce the computational burden from non informative regions, it is necessary to separate the background from the foreground. In order to accomplish this, the image is converted to the HSV (Hue, Saturation, Value) colour space. Next, we proceed to manipulate the Saturation component. In order to identify the foreground, a combination of OTSU Thresholding, Median blurring, and Morphological closing techniques are employed on a downsampled version of the image. Next, the foreground mask identified from this process is upsampled to match the dimensions of the original image. Utilizing a downsampled version of the image allows us to work with more manageable input sizes. We can overlook minor boundary errors that may arise from downsampling. This process guarantees that we exclusively handle image sections that contain valuable informa-

tion.

5.3.2 Patching

Once the foreground and background have been successfully separated, the image is then divided into smaller patches measuring 512 x 512 pixels each. Through this process, approximately 40,000 patches are generated per whole slide image, depending on the size of tissue present in the slide. By breaking down the image data into manageable segments, patching enables more efficient processing and analysis.

5.4 Masked Training

We use random slide masking to force the model to learn from fewer slides while promoting cross-attention. Our main goal in this research is to enable the model to learn to predict even in the absence of certain slides. We aim to improve the model's generalizability. Due the nature of data collection and the focus on acquiring certain whole slides , some of the slide are more abundantly present the dataset. The most available slide is present in 90% of the cases, while the least available slide is present only in 35% of the cases

To address this issue, we introduce a dynamic slide masking process that simulates the absence of certain slides during training. Every slide has a presence indicator, and each one is assigned a masking ratio to show how likely it is to be masked. We conduct a grid search to find the best masking ratios, and then generate a random mask for each training iteration using these ratios. We make sure that the model receives at least one slide to make predictions.

The model is trained using randomly masked slides, which helps it learn from a wide range of situations that resemble real-life scenarios where slide availability can differ. The algorithm used for slide masking is described in Algorithm 3.

By incorporating this random slide masking process, we address the challenge of data variability and promote the model's generalisability. The model learns to make accurate predictions with incomplete information, as in a real world scenario. This approach tackles the issue of data imbalance and prepares the model to handle situations where certain slides are missing, ultimately leading to more reliable and accurate predictions in practical applications.

Algorithm 3 Dynamic Slide Masking

Require: $slidePresenceIndicator$: Tensor indicating whether a slide's data is present (1) or absent (0)
Require: $slideMaskingRatios$: Tensor of individual masking ratios for each slide
Ensure: $finalSlideMask$: Tensor indicating the presence of slides after dynamic masking

Initialize $randomMask$ tensor with the same shape as $slidePresenceIndicator$
for each index in $slidePresenceIndicator$ **do**
  Generate a random number $randomValue$ between 0 and 1
  if $randomValue$ $slideMaskingRatios$[index] **then**
    Set $randomMask$[index] $\leftarrow$ 1
  else
    Set $randomMask$[index] $\leftarrow$ 0
  end if
end for
$finalSlideMask \leftarrow slidePresenceIndicator * randomMask$ {Apply dynamic masking}
if $sum(finalSlideMask) == 0$ **then**
  Randomly select an index i from where $slidePresenceIndicator$ is 1
  Set $finalSlideMask[i] \leftarrow 1$ {Ensure at least one slide remains unmasked}
end if
Return: $finalSlideMask$

Algorithm 4 PPEG(Pyramid Position Encoding Generator)

Input: A bag of feature embeddings H_S^l after correlation modelling, where $H_S^l \in \mathbb{R}^{(N+1)\times d}$.

Output: The feature embeddings H_S^P after conditional position encoding and local information fusion, where $H_S^P \in \mathbb{R}^{(N+1)\times d}$.

1: **Split:** H_S^l is divided into patch tokens H_f and class token H_C;
$H_f, H_C \leftarrow Split(H_S^l)$, where $H_f \in \mathbb{R}^{N\times d}, H_C \in \mathbb{R}^{1\times d}$.

2: **Spatial Restore:** patch tokens H_f are reshaped to H_S^f in the 2-D image space;
$H_S^f \leftarrow Restore(H_f)$, where $H_S^f \in \mathbb{R}^{\sqrt{N}\times\sqrt{N}\times d}$.

3: **Group Convolution:** using a set of group convolutions with kernel k and $\frac{k-1}{2}$ zero paddings($k = 3, 5, 7$) to obtain H_t^f, $t = 1, 2, 3$;
$H_t^f \leftarrow Conv(H_S^f)$, where $H_t^f \in \mathbb{R}^{\sqrt{N}\times\sqrt{N}\times d}$, $t = 1, 2, 3$.

4: **Fusion:** H_S^f and the H_t^f, $t = 1, 2, 3$ obtained from the convolution block processing are added together to obtain H_S^F;
$H_S^F \leftarrow H_S^f + H_1^f + H_2^f + H_3^f$, where $H_S^F \in \mathbb{R}^{\sqrt{N}\times\sqrt{N}\times d}$.

5: **Flatten:** H_S^F are flattened into sequence H_{se};
$H_{se} \leftarrow Flatten(H_S^F)$, where $H_{se} \in \mathbb{R}^{N\times d}$.

6: **Concat:** connect H_{se} and class token H_C to obtain H_S^P;
$H_S^P \leftarrow Concat(H_{se}, H_C)$, where $H_S^P \in \mathbb{R}^{(N+1)\times d}$.

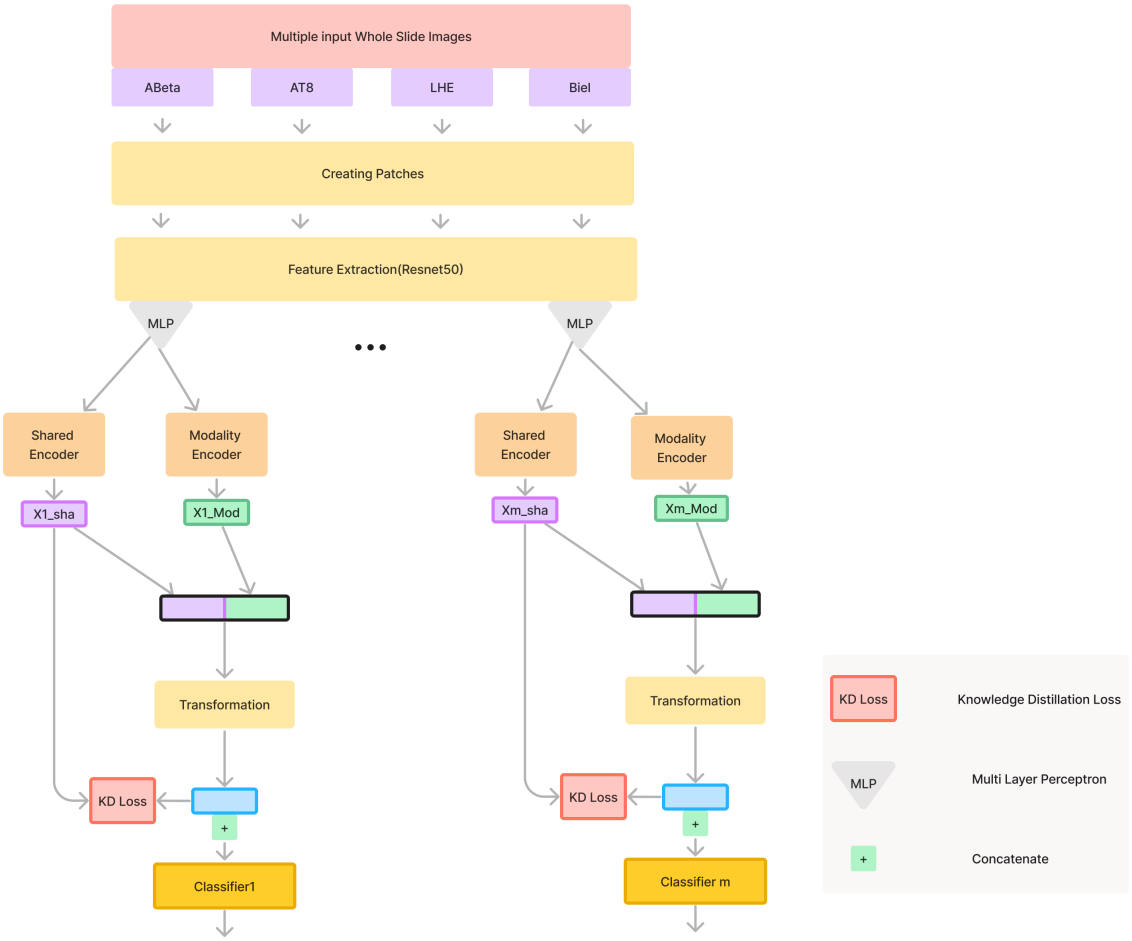

Figure 5·1: Figure presents the general pipeline of the proposed MultiStainKD approach. Multiple input whole slide images, such as ABeta, AT8, LHE, and Biel, undergo patch creation. The extracted patches from each slide are processed independently using separate self attention blocks. The self attention blocks capture local feature relationships within each slide. We also calculate calculate the embeddings through a shared encoder. We then concatenate the attended features from all slides and pass them through a multi-layer perceptron (MLP) to learn cross-slide representations. Finally the results are aggregated to calculate the logits for the patient.

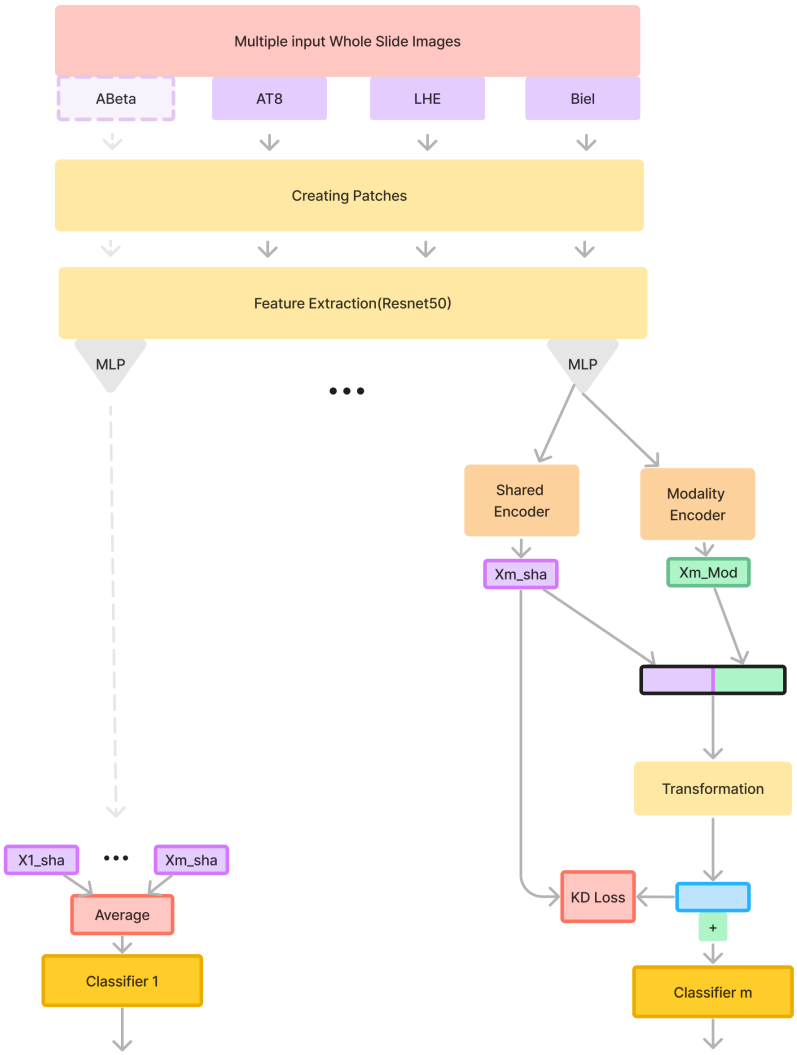

Figure 5·2: Figure Presents the general pipeline when a modality is missing. In this example, the ABeta modality is missing; the rest of the pipeline works as it is; since no features are generated from ABeta, there would be no contribution from the feature.

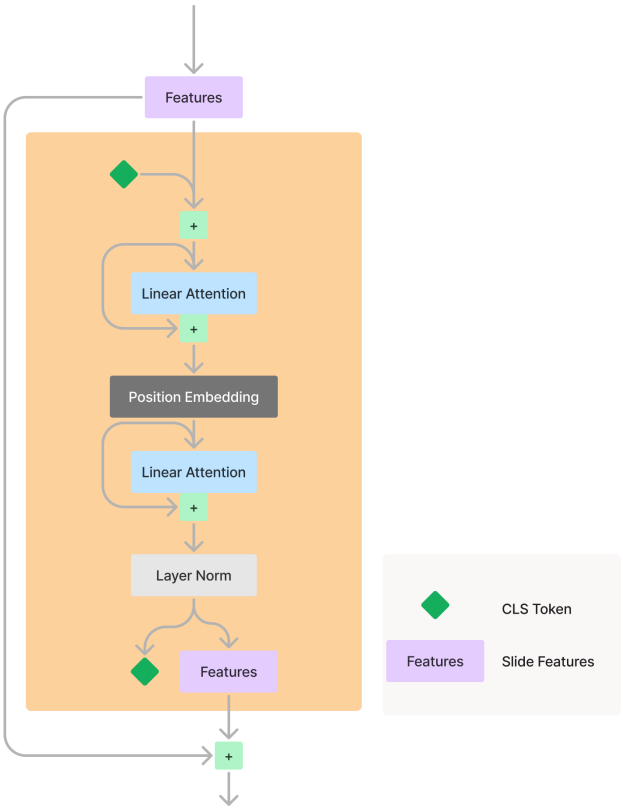

Figure 5·3: Figure Illustrates the self attention mechanism employed in the proposed method. The input features X undergo a series of operations to capture both local dependencies and global context. First, a linear attention layer identifies important relationships among the features. Followed by the addition of position embeddings to incorporate spatial information. Another linear attention layer further refines the features based on positional context. The resulting features are added to the previous ones and normalized using layer normalization. A CLS token is added to capture global slide-level information. The final output features X' and the CLS token serve as rich representations for downstream tasks.

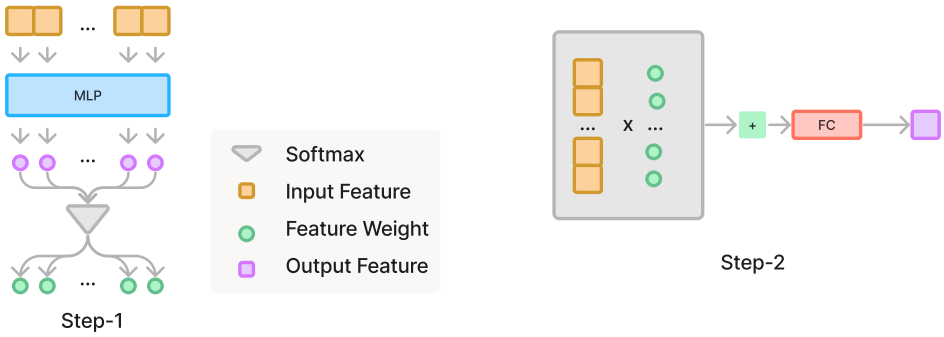

Figure 5·4: Keyless Attention Block Step 1: Input features are passed through an MLP block for feature transformation. Step 2: The Softmax operation generates attention weights, which are multiplied with input features. The weighted features are then processed by a fully connected (FC) layer to obtain the final feature representation.

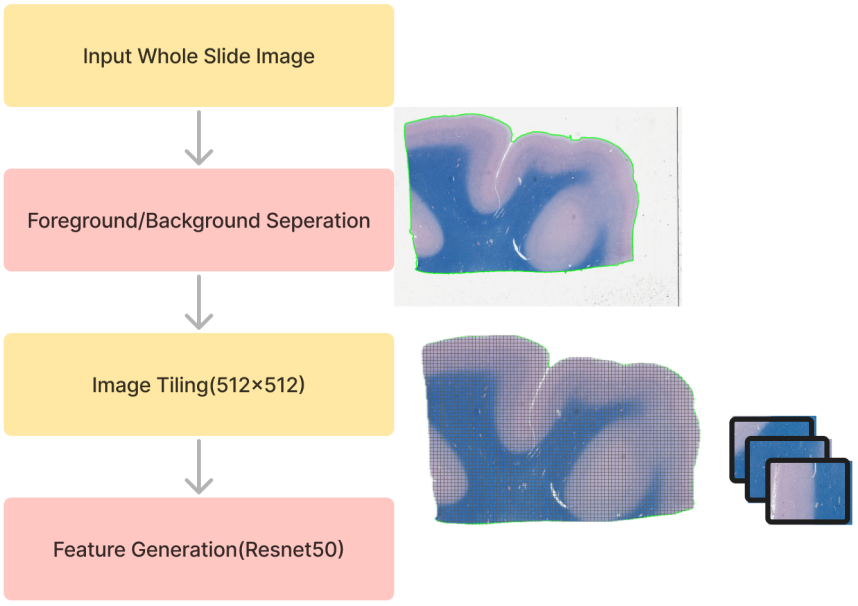

Figure 5·5: Figure describes the Preprocessing pipeline: Remove the background, Tile the image. Generate Features for each patch using ResNet based model

Chapter 6

Experiments

To demonstrate the robustness of the proposed model, we conducted two experiments designed to evaluate its performance under different conditions of missing data. In the first experiment, we focused on cases where the AT8 stained slide was consistently present; similarly, in the second experiment, we ensured that the ABeta stained slide was consistently available. In both experiments, at least two slides were available for all patients.

The experimental setup involved training the model using the available patient data while employing a slide masking technique. This approach allowed the model to learn to make predictions even when some information was missing. Randomly masking out certain slides during training forced the model to adapt and effectively leverage the available data.

To assess the model's robustness, we employed the trained model in two testing scenarios. In the first scenario, we removed the AT8 stained slide from the test cases, while in the second scenario, we removed the ABeta stained slide. We then evaluated the model's performance under these conditions of missing data.

Remarkably, the model demonstrated comparable scores between the two testing scenarios. Despite the absence of either the AT8 or ABeta stained slide, the model was able to maintain its predictive accuracy. This finding suggests that the model is robust to missing data and can effectively utilize the available information from the remaining slides to make accurate predictions.

These experimental results provide valuable insights into the model's capability to handle real-world situations where certain slides may be unavailable or missing. The model's robustness, which ensures reliable results even in the presence of incomplete data, is a crucial feature.

In the following sections, we will delve into the details of each experiment, present the quantitative results, and discuss the implications of these findings on the practical application of the proposed model.

6.1 Evaluation metric

To evaluate the performance of our proposed model, we employed the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) metric. The AUC quantifies the model's overall performance, providing a single scalar value that represents the probability of the model ranking a randomly chosen positive instance higher than a randomly chosen negative instance.

The choice of ROC AUC as our evaluation metric is particularly suitable for our case due to several reasons. Firstly, our dataset has an imbalanced distribution of classes, where the number of samples in each class is not equal. ROC AUC is robust to class imbalance and provides a fair assessment of the model's performance regardless of the class distribution. Secondly, ROC AUC is threshold-independent, meaning it evaluates the model's performance across all possible classification thresholds. This is advantageous as it eliminates the need to manually select an optimal threshold and allows for a comprehensive assessment of the model's discriminative power.

Moreover, ROC AUC has a clear probabilistic interpretation, making it intuitive to understand and compare different models. An AUC of 0.5 indicates a random classifier, while an AUC of 1.0 represents a perfect classifier. By using ROC AUC, we can effectively measure the model's ability to distinguish between different classes

and make informed decisions based on the probability scores.

To ensure the reliability and generalizability of our results, we employed a 10-fold cross-validation strategy. We divide the dataset into 10 folds in this approach. We train and evaluate the model 10 times, using 9 folds for training and the remaining fold for testing each time. We compute the performance metrics, including ROC AUC, for each fold and average the final results across all 10 folds. This cross-validation technique helps to mitigate overfitting and provides a more robust estimate of the model's performance on unseen data.

By using ROC AUC as our evaluation metric and performing 10-fold cross validation, we aim to thoroughly assess the performance of our proposed model and ensure its effectiveness in handling the challenging task at hand.

We compare our results with (Garcia et al., 2019) and methods commonly used for fusion of features as baselines.

For all the experiments, we extract features from Resnet50 and pass the extracted features to each of the models while keeping the other structure the same. I tried the best of my capabilities to keep the replicate the results using our dataset. All the experiments follow the same batch size and learning rate to keep them consistent across the board.

Chapter 7

Results and Discussion

7.1 Experiment 1: Evaluating Model Robustness with AT8

The first experiment focused on evaluating the model's robustness when the AT8 stained slide was consistently present during training and testing. AT8, an antibody that specifically targets hyperphosphorylated tau protein, is a key pathological hallmark of Alzheimer's disease (AD) and chronic traumatic encephalopathy (CTE). With AT8 stained slides available for approximately 95% of the cases in our dataset, it was a suitable candidate for this experiment.

The primary objective was to assess the model's ability to learn and make accurate predictions even when the AT8 slide was missing during testing. By training the model with AT8 slides always present and then evaluating its performance in scenarios where AT8 was intentionally removed, we aimed to investigate the model's capability to leverage information from the remaining slides.

Table 7.1 presents the ROC AUC scores achieved by the model in different scenarios when predicting AD. Similarly, Table 7.2 presents the ROC AUC scores achieved by the model when predicting CTE. When at least two slides were present, including AT8, the model obtained an ROC AUC score of 0.846 for the AD/non-AD classification and 0.858 for the CTE/non-CTE classification. These scores showcase the model's efficacy in leveraging the AT8 stained slide's information alongside other accessible slides. Interestingly, when the AT8 slide was purposely removed during testing, the model still achieved competitive performance. For AD/non-AD classifi-

cation, the ROC AUC score decreased to 0.740 which is 1.5% better than the next best method.

Similarly, when predicting CTE/non-CTE, our model achieved the highest AUC score of 0.864 even when the AT8 modality was missing during testing. This suggests that the model was able to effectively leverage information from the remaining slides to make accurate predictions, highlighting its robustness in handling missing modalities.

These results underscore the importance of the model's ability to adapt and make reliable predictions even in the absence of a key modality like AT8. The model's performance in these scenarios demonstrates its potential to support clinical decision-making, particularly in cases where certain stained slides may not be available.

Architecture	AD/non-AD	
	With AT8	Missing AT8
Unimodality		
UniModality(AT8)	0.877	-
Simple Additive methods		
AdditiveFusion	0.842	**0.725**
TransformerFusion	0.788	0.718
Current SOTA		
(Garcia et al., 2019)(smoothing=10)	0.741	0.610
Our Method		
MultiStainKD (Our method)	**0.846**	0.740

Table 7.1: Predicting AD: ROC AUC scores when at least 2 slides present and AT8 always present. Marked in Red are the Best scores and in Bold are the second best

7.2 Experiment 2: Evaluating Model Robustness with ABeta

The second experiment focused on evaluating the model's robustness when the ABeta stained slide was consistently present during training and testing. ABeta, short for amyloid-beta, is a protein that forms plaques in the brains of individuals with Alzheimer's disease (AD). Although ABeta is not as prevalent as AT8 in our dataset,

Architecture	CTE/non-CTE	
	With AT8	Missing AT8
Unimodality		
UniModality(AT8)	0.837	-
Simple Additive methods		
AdditiveFusion	0.858	0.847
TransformerFusion	**0.859**	0.865
Current SOTA		
(Garcia et al., 2019)(smoothing=10)	0.825	0.820
Our Method		
MultiStainKD (Our method)	0.863	**0.864**

Table 7.2: Predicting CTE: ROC AUC scores when at least 2 slides present and AT8 always present. Marked in Red are the Best scores and in Bold are the second best

being present in only 40% of the cases, it still plays a significant role in the pathology of AD.

The objective was to assess the model's ability to learn and make accurate predictions when ABeta slides were available during training but intentionally removed during testing. By evaluating the model's performance in the absence of ABeta slides, we aimed to investigate its capability to leverage information from the remaining slides and maintain robust performance.

Table 7.3 presents the ROC AUC scores achieved by the model in different scenarios when predicting CTE. Similarly, Table 7.4 presents the ROC AUC scores when predicting AD. When at least two slides were present, including ABeta, the model obtained an ROC AUC score of 0.888 for CTE/non-CTE classification, which is slightly less than the best method. Remarkably, when the ABeta slide was intentionally removed during testing, the model exhibited robust performance. For CTE/non-CTE classification, the model maintained a high ROC AUC score of 0.799, outperforming the next best method by 1

When predicting AD/non-AD, the best model achieved an AUC score of 0.83, while our model closely followed with a score of 0.826. Notably, our model performed

the best when the ABeta modality was hidden, achieving a score of 0.833.

These results highlight the model's robustness and adaptability in handling missing modalities, particularly in the case of ABeta. Despite the intentional removal of ABeta slides during testing, the model demonstrated strong performance, maintaining high ROC AUC scores for both CTE/non-CTE and AD/non-AD classification tasks.

The model's ability to leverage information from the remaining slides and compensate for the absence of ABeta is a testament to its robustness and potential for real-world application. In clinical settings where certain stained slides may not be available, our model's performance suggests that it can still provide reliable predictions and support decision-making processes.

Architecture	CTE/non-CTE	
	With ABeta	Missing ABeta
Unimodality		
UniModality(ABeta)	0.832	-
Simple Additive Methods		
AdditiveFusion	0.888	0.760
TransformerFusion	0.862	0.763
Simple Additive Methods		
(Garcia et al., 2019)(smoothing=10)	0.870	**0.790**
Our Method		
MultiStainKD (Our method)	**0.872**	0.799

Table 7.3: Predicting CTE: ROC AUC scores when at least 2 slides present and AT8 always present

Architecture	AD/non-AD	
	With ABeta	Missing ABeta
Unimodality		
UniModality(ABeta)	0.846	-
Simple Additive Methods		
AdditiveFusion	0.830	**0.795**
TransformerFusion	0.727	0.769
Simple Additive Methods		
(Garcia et al., 2019)(smoothing=10)	**0.831**	0.678
Our Method		
MultiStainKD (Our method)	0.826	0.833

Table 7.4: Predicting AD: ROC AUC scores when at least 2 slides present and AT8 always present

7.3 Ablation Studies

7.3.1 Ablation study on weight of Knowledge Distillation

We investigate the impact of the weight parameter λ on the performance of our model. The weight parameter λ is used to balance the contribution of the knowledge distillation loss (KDLoss) in the overall training objective. To conduct this ablation study, we train our model with different values of λ and evaluate its performance on two scenarios: when the AT8 stain is available (With AT8) and when the AT8 stain is missing (Missing AT8). Table 7.5 presents the results of our ablation study. We observe that setting λ to 0.2 yields the best performance when the AT8 stain is available, achieving an accuracy of 0.846. This suggests that incorporating the KDLoss with a moderate weight helps the model to better leverage the information from the available modalities and improves its overall performance. When λ is set to 1, giving equal weight to the KDLoss and other loss components, we observe a significant drop in performance both with and without the AT8 stain, indicating that overemphasizing the knowledge distillation process can be detrimental to the model's learning and generalization abilities.

Weight(λ)	With AT8	Missing AT8
0	0.793	0.730
0.2	0.846	0.740
0.5	0.792	0.651
1	0.662	0.600

Table 7.5: Comparison of Lambda used to combine Loss

Chapter 8

Future Works

While our research has made strides in addressing the challenge of missing modalities in digital pathology, there is still ample room for improvement and further exploration.

One key area for future work is enhancing the robustness of our model to missing modalities. Although MultiStainKD has demonstrated its ability to handle incomplete data effectively, developing even more resilient and adaptable methods can further improve its performance and practicality in real-world scenarios. Investigating advanced techniques for data imputation, such as matrix completion or collaborative filtering, could help mitigate the impact of missing modalities and provide more accurate predictions.

Exploring generative models is another promising direction for future research. In recent years, there has been a surge of interest in using autoencoder-based models to generate features for missing modalities. These models learn to reconstruct the missing data by capturing the underlying patterns and relationships within the available modalities. By incorporating generative models into our framework, we can potentially enhance its capability to handle missing modalities and improve overall performance. Variational autoencoders (VAEs) and generative adversarial networks (GANs) are two popular architectures that have shown promising results in feature generation and could be explored in the context of digital pathology.

Moreover, exploring the integration of domain knowledge and expert guidance into our model could greatly enhance its interpretability and clinical relevance. Col-

laborating with pathologists and incorporating their expertise into the development and evaluation process can help ensure that our model captures meaningful and diagnostically relevant features. By involving domain experts, we can also validate the generated features and assess their biological plausibility, increasing the trustworthiness and acceptability of our model in clinical settings.

Bibliography

Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., and Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309.

Chen, R. J., Lu, M. Y., Weng, W.-H., Chen, T. Y., Williamson, D. F., Manz, T., Shady, M., and Mahmood, F. (2021). Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3995–4005.

Dolz, J., Gopinath, K., Yuan, J., Lombaert, H., Desrosiers, C., and Ayed, I. B. (2019). Hyperdense-net: A hyper-densely connected cnn for multi-modal image segmentation. https://arxiv.org/abs/1804.02967

Dorent, R., Joutard, S., Modat, M., Ourselin, S., and Vercauteren, T. (2019). *Hetero-Modal Variational Encoder-Decoder for Joint Modality Completion and Segmentation*, page 74–82. Springer International Publishing.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. https://arxiv.org/abs/2010.11929

Feng, J. and Zhou, Z.-H. (2017). Deep miml network. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 1884–1890. AAAI Press.

Garcia, N. C., Bargal, S. A., Ablavsky, V., Morerio, P., Murino, V., and Sclaroff, S. (2019). Dmcl: Distillation multiple choice learning for multimodal action recognition. https://arxiv.org/abs/1912.10982

Havaei, M., Guizard, N., Chapados, N., and Bengio, Y. (2016). Hemis: Hetero-modal image segmentation. https://arxiv.org/abs/1607.05194

He, L., Long, L. R., Antani, S., and Thoma, G. R. (2012). Histology image analysis for carcinoma detection and grading. *Computer Methods and Programs in Biomedicine*, 107(3):538–556.

Hou, L., Samaras, D., Kurc, T. M., Gao, Y., Davis, J. E., and Saltz, J. H. (2016). Patch-based convolutional neural network for whole slide tissue image classifica-tion. https://arxiv.org/abs/1504.07947

Ilse, M., Tomczak, J. M., and Welling, M. (2018). Attention-based deep multiple instance learning. https://arxiv.org/abs/1802.04712

Lerousseau, M., Vakalopoulou, M., Deutsch, E., and Paragios, N. (2021). Sparseconvmil: Sparse convolutional context-aware multiple instance learning for whole slide image classification. In Atzori, M., Burlutskiy, N., Ciompi, F., Li, Z., Minhas, F., Müller, H., Peng, T., Rajpoot, N., Torben-Nielsen, B., van der Laak, J., Veta, M., Yuan, Y., and Zlobec, I., editors, *Proceedings of the MICCAI Workshop on Computational Pathology*, volume 156 of *Proceedings of Machine Learning Research*, pages 129–139. PMLR.

Li, B., Li, Y., and Eliceiri, K. W. (2021). Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. https://arxiv.org/abs/1411.4038

Long, X., Gan, C., Melo, G., Liu, X., Li, Y., Li, F., and Wen, S. (2018). Multimodal keyless attention fusion for video classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Lu, M. Y., Williamson, D. F. K., Chen, T. Y., et al. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5:555–570.

Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., and Peng, X. (2021). Smil: Multimodal learning with severely missing modality. https://arxiv.org/abs/2103.05677

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. https://arxiv.org/abs/1505.04597

Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al. (2021). Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34:2136–2147.

Tseng, K.-L., Lin, Y.-L., Hsu, W., and Huang, C.-Y. (2017). Joint sequence learning and cross-modality convolution for 3d biomedical segmentation.

Tu, M., Huang, J., He, X., and Zhou, B. (2019). Multiple instance learning with graph neural networks. https://arxiv.org/abs/1906.04881

Vasco, M., Yin, H., Melo, F. S., and Paiva, A. (2022). Leveraging hierarchy in multimodal generative models for effective cross-modality inference. *Neural Networks : The Official Journal of the International Neural Network Society*, 146:238–255.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need. arxiv.org/abs/1706.03762

Zhang, C., Chu, X., Ma, L., Zhu, Y., Wang, Y., Wang, J., and Zhao, J. (2022a). M3care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 2418–2428, New York, NY, USA. Association for Computing Machinery.

Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S. E., and Zheng, Y. (2022b). Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. https://arxiv.org/abs/2203.12081

Zhang, Y., He, N., Yang, J., Li, Y., Wei, D., Huang, Y., Zhang, Y., He, Z., and Zheng, Y. (2022c). mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. arxiv.org/abs/2206.02425v2

Zhao, Y., Yang, F., Fang, Y., Liu, H., Zhou, N., Zhang, J., Sun, J., Yang, S., Menze, B., Fan, X., and Yao, J. (2020). Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4836–4845.

Zheng, Y., Gindra, R. H., Green, E. J., Burks, E. J., Betke, M., Beane, J. E., and Kolachalama, V. B. (2022). A graph-transformer for whole slide image classification. *IEEE Transactions on Medical Imaging*, 41(11):3003–3015.

Curriculum Vitae