DEEP LEARNING FOR SUBTYPING CLASSICAL HODGKIN LYMPHOMA ON HISTOPATHOLOGY IMAGES: A COMPREHENSIVE LYMPHOSCOPE

by

Hicran Aldemir

B.S., in Biomedical Engineering, Ankara University, 2020

Submitted to the Institute of Biomedical Engineering
in partial fulfillment of the requirements
for the degree of
Master of Science
in
Biomedical Engineering

Boğaziçi University
2024

DEEP LEARNING FOR SUBTYPING CLASSICAL HODGKIN LYMPHOMA ON HISTOPATHOLOGY IMAGES: A COMPREHENSIVE LYMPHOSCOPE

APPROVED BY:

Doç. Dr. Mehmet Turan .....
(Thesis Advisor)

Doç. Dr. Derya Demir .....

Prof. Dr. Ahmet Ademoğlu .....

DATE OF APPROVAL: 6 August 2024

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to the DeepMIA Lab at Boğaziçi University, in the Institute of Biomedical Engineering, for providing me with an exceptional research environment and a platform to connect with experts in the field. I am especially grateful to Doç. Dr. Mehmet Turan for fostering such an exceptional environment. His guidance, support, and patience throughout my Master's program have been instrumental to my success.

I am deeply grateful to Doç. Dr. Derya Demir for generously giving me her time and meticulously guiding me throughout the project. I would like to thank the pathologists involved in this study for their valuable support and the committee for their valuable time and comments. Completing this study would not have been possible without the unwavering support they gave me throughout my research.

I would like to thank my dear Uğurcan Akyüz, who has been an incredible mentor throughout my Master's program, and I am grateful for his tireless encouragement and dedication.

I would like to thank the Scientific and Technological Research Council of Turkey (TUBITAK) for funding my Master's studies with the 2210 Scholarship Program and the TUBITAK 2232 International Outstanding Researcher Fellowship. In addition, I am grateful for the TRUBA (Turkish National e-Science e-Infrastructure) Cluster and the data storage services provided by TUBITAK Ulakbim, which have supported my research endeavors immensely.

Finally and most importantly, I would like to thank my family for their encouragement and understanding throughout this challenging but fulfilling journey.

ACADEMIC ETHICS AND INTEGRITY STATEMENT

I, Hicran Aldemir, hereby certify that I am aware of the Academic Ethics and Integrity Policy issued by the Council of Higher Education (YÖK), and I fully acknowledge all the consequences due to its violation by plagiarism or any other way.

Name :

Signature:

Date:

ABSTRACT

DEEP LEARNING FOR SUBTYPING CLASSICAL HODGKIN LYMPHOMA ON HISTOPATHOLOGY IMAGES: A COMPREHENSIVE LYMPHOSCOPE

Classical Hodgkin lymphoma (CHL) is an uncommon form of lymphoma that mainly affects young adults and adolescents. Accurate morphologic assessment of CHL is critical for precise subtyping and effective treatment planning. Despite its clinical importance, there is a notable research gap in the automated subtyping of CHL, in contrast to the extensive developments in more common cancers. To address this gap, we curated a comprehensive dataset of 1247 whole-slide images (WSI) for four CHL subtypes from three medical centers from Türkiye, ensuring a reliable basis for our research. We developed a deep-learning pipeline for Classical Hodgkin lymphoma subtyping. Our weakly supervised model, Instance and Embedding Fused Multiple Instance Learning (IEF-MIL), utilizes a multiscale dual-stream network, outperforming state-of-the-art MIL models in the existing literature. Furthermore, we have incorporated state-of-the-art self-supervised learning foundation models trained on hundreds of histopathology whole slide images into our weakly supervised pipeline through transfer learning. We have demonstrated the generalizability and limitations of these models on our out-of-distribution dataset. Additionally, our model enhances interpretability with heatmaps at three magnification levels, providing deeper insights into its predictions.

Keywords: Classical Hodgkin Lymphoma, Multiple Instance Learning, Self-Supervised Learning, Histopathology Foundation Models, Computational Pathology.

ÖZET

KLASİK HODGKİN LENFOMANIN HİSTOPATOLOJİ GÖRÜNTÜLERİ KULLANILARAK DERİN ÖGRENME İLE ALT TİPLEMESİ: KAPSAMLI BİR LENFOSKOP

Klasik Hodgkin lenfoma (KHL), genellikle genç yetişkinleri ve ergenleri etkileyen nadir bir lenfoma türüdür. KHL'nin doğru morfolojik değerlendirilmesi, doğru alt tiplerin belirlenmesi ve etkili tedavi planlaması için kritik öneme sahiptir. Klinik önemine rağmen, daha yaygın kanserlerin aksine, KHL'nin otomatik alt tiplerinin belirlenmesinde belirgin bir araştırma boşluğu bulunmaktadır. Bu açığı kapatmak amacıyla, Türkiyedeki üç tıp merkezinden dört KHL alt tipine ait 1247 tam slayt görüntüsünü (WSI) içeren kapsamlı bir veri seti oluşturduk ve araştırmamız için güvenilir bir temel sağladık. Klasik Hodgkin lenfoma alt tiplemesini yapmak için bir derin öğrenme metodu geliştirdik. Zayıf denetimli modelimiz, Instance and Embedding Fused Multiple Instance Learning (IEF-MIL), çoklu ölçekli çift akışlı bir model kullanarak mevcut literatürdeki en iyi çoklu örnekle öğrenme modellerinden daha iyi performans göstermiştir. Ayrıca, yüzlerce histopatoloji tam slayt görüntüleri üzerinde eğitilmiş kendi kendine denetimli öğrenme kaynak modellerini, zayıf denetimli metodumuza transfer öğrenme yoluyla entegre ettik. Bu modellerin genellenebilirliğini ve sınırlamalarını dağılım dışı bir veri seti olan kendi veri setimiz üzerinde gösterdik. Ayrıca, modelimiz, üç büyütme seviyesinde ısı haritaları oluşturarak tahminleri hakkında daha detaylı açıklamalar sunmaktadır.

Anahtar Sözcükler: Klasik Hodgkin Lenfoma, Çoklu Örnekle Öğrenme, Kendi Kendine Denetimli Öğrenme, Histopatoloji Kaynak Modelleri, Dijital Patoloji.

TABLE OF CONTENTS

ACKNOWLEDGMENTS . . . iii
ACADEMIC ETHICS AND INTEGRITY STATEMENT . . . iv
ABSTRACT . . . v
ÖZET . . . vi
LIST OF FIGURES . . . ix
LIST OF TABLES . . . xi
LIST OF ABBREVIATIONS . . . xii
1. INTRODUCTION . . . 1
1.1 Motivation . . . 1
1.2 Objectives and Contributions . . . 4
2. LITERATURE ANALYSIS . . . 5
2.1 Weakly Supervised Learning and Multiple Instance Learning (MIL) . . . 5
2.1.1 MIL Aggregation . . . 6
2.2 Feature Extraction Methods . . . 7
2.2.1 Convolutional Neural Networks . . . 8
2.2.2 Self-Supervised Learning (SSL) . . . 8
3. MATERIALS AND METHODS . . . 12
3.1 Datasets . . . 12
3.2 Whole Slide Image Preprocessing . . . 14
3.3 Model Architecture . . . 17
3.3.1 Self-Supervised Vision Transformer as a Feature Extractor . . . 17
3.3.2 MIL Aggregator . . . 19
3.4 Training Details and Hyperparameters . . . 21
3.5 Evaluation Metrics . . . 22
3.6 System Specifications . . . 23
4. RESULTS . . . 24
4.1 Classification Results on Test Set . . . 24
4.2 Ablation Study: The Performance of Multiple-Scales . . . 27
4.3 Interpretability with Heatmaps . . . 28

5. DISCUSSION 31
6. CONCLUSION 34
REFERENCES 35

LIST OF FIGURES

Figure 3.1 **Workflow Overview. A.** Lymph nodes are surgically removed, fixed in formalin, embedded in paraffin (FFPE), sliced, and stained with Hematoxylin and Eosin (H&E). Stained slices are scanned to create digitized whole slide images (WSIs). **B.** The study uses three datasets from Türkiye: EUH, BUH, and IUH, with patient and slide counts noted. WSIs are scanned at 40x magnification. The EUH and BUH datasets are combined for training and validation with 5-fold cross-validation, while the IUH dataset is used as an independent cohort. **C.** Tissue is segmented from WSIs into patches across three magnification levels. These patches are used for training, validation, and testing in a Multiple Instance Learning (MIL) network, which produces slide-level predictions and heatmaps for each magnification. 13

Figure 3.2 **Overview of the Proposed MIL Pipeline. A.** Preprocessing step showing multi-scale feature extraction from patches at 5x, 10x, and 20x magnifications, with embeddings concatenated for model training, validation, and testing. The CONCH model [1] is used for feature extraction with transfer learning. **B.** Architecture of the IEF-MIL model, featuring instance-based pooling (mean pooling of embeddings) and embedding-based pooling (global attention pooling). The final slide score and prediction are computed by averaging the mean and bag scores. 18

Figure 4.1 Class-based performance evaluation of the IEF-MIL model trained with CONCH features. **A.** Precision-recall curves and the area under the precision-recall curve (AUC-PR) provided for each class. **B.** Confusion matrix depicting the raw results of model predictions. 26

Figure 4.2 Comparison of F1 scores across different magnification combinations of feature vectors from six histopathology foundation models trained on the IEF-MIL model. 28

Figure 4.3 **Heatmap Visualization of Correctly Classified Subtypes.** The figure shows three magnification heatmaps for NSCHL, MCCHL, LRCHL, and LDCHL subtypes. Each heatmap displays top-scoring patches, with black squares indicating ROIs. Red areas highlight high-attention regions critical for classification, while blue areas show low attention. The heatmaps illustrate the model's focus and its alignment with subtype-specific features. 29

LIST OF TABLES

Table 3.1 Magnifications and pyramid levels for .ndpi, .mrxs, and .svs WSI file formats. Level 0 is 40x magnification. Magnification for each pyramid scale is calculated as Magnification = $\frac{40x}{\text{Scale}}$. Values for lower pyramid levels with extensive decimal places are excluded. 15

Table 3.2 Overview of feature extractor models used in this study. The asterisk (*) denotes the number of patches instead of WSIs. 16

Table 4.1 Comparison of AUC (bottom) and F1 (top) scores (mean ± SD) for the IEF-MIL and other MIL models on a test set using features from six backbones. The top-performing MIL model for each backbone is underlined. 25

LIST OF ABBREVIATIONS

H&E	Hematoxylin and Eosin
HRS	Hodgkin and Reed/Sternberg
AUC	Area Under the Curve
CNN	Convolution Neural Network
CL	Contrastive Learning
CHL	Classical Hodgkin Lymphoma
DL	Deep Learning
FN	False Negative
FP	False Positive
MIL	Multiple Instance Learning
MCCHL	Mixed Cellularity Classical Hodgkin Lymphoma
NSCHL	Nodular Sclerosis Classical Hodgkin Lymphoma
LDCHL	Lymphocyte-Depleted Classical Hodgkin Lymphoma
LRCHL	Lymphocyte-Rich Classical Hodgkin Lymphoma
ROC	Receiver Operating Characteristic Curve
SSL	Self-Supervised Learning
TN	True Negative
TP	True Positive
ViT	Vision Transformer
WSI	Whole Slide Image

1. INTRODUCTION

1.1 Motivation

Lymphoma is a type of cancer that originates from lymphocytes, the lymphatic system cells that fight infections. These cells are mainly found in the lymph nodes, spleen, lungs, liver, and bone marrow but can also form in all other organs of the body [2]. In healthy people, there are two types of lymphocytes: T-cells and B-cells. In lymphoma, these cells become malignant, proliferate abnormally, and exhibit some changes at the cellular level. Lymphoma is categorized into Hodgkin lymphoma (HL) and non-Hodgkin lymphoma (NHL). NHL is among the top 15 most prevalent cancers, accounting for 3% of cases or 544,352 individuals worldwide [3]. In contrast, HL, a rarer form, represents only 0.4% of global cancer cases, with 83,087 individuals diagnosed in 2020. It contributed to 0.2% of all cancer-related deaths, with 23,376 fatalities that year. HL is further categorized into two main subtypes: Classical HL (CHL), making up about 85% of cases, and Nodular Lymphocyte-Predominant HL (NLPHL) [4].

Classical Hodgkin lymphoma, the primary focus of this research, is the most common type of HL, and some of its subtypes might be considered a rare cancer. These subtypes, listed from the most to least common, are Nodular Sclerosis Classical Hodgkin Lymphoma (NSCHL), Mixed Cellularity Classical Hodgkin Lymphoma (MC-CHL), Lymphocyte-Rich Classical Hodgkin Lymphoma (LRCHL), and Lymphocyte-Depleted Classical Hodgkin Lymphoma (LDCHL) [5]. CHL predominantly affects adolescents and young adults, with a lower incidence in older age groups [6].

In the initial diagnosis of Classical Hodgkin lymphoma, a complete biopsy is required, in which a lymph node is surgically removed for examination. Core needle biopsies are available but are less suitable for HL diagnosis due to their limitations. They may not capture enough tissue to visualize the essential overall cellular architecture, such as the critical fibrotic bands for identifying NSCHL. In addition, because

malignant cells are sparse in HL compared to other lymphomas, core needle biopsies are more likely to miss these cancerous cells. Therefore, a complete biopsy is more effective and reliable for the diagnosis of CHL and its subtyping [2].

Diagnosis of Classical Hodgkin lymphoma is based on examining specific cancer cells, namely multi-nucleate Reed-Sternberg cells and mononuclear Hodgkin cells, in an abundant tumor microenvironment [4]. CHL primarily originates from germinal-center B cells. These cells transform, losing their typical characteristics, morphing into distinctly larger cells than healthy lymphocytes, and exhibiting nuclear variations. HRS cells exhibit an abnormal B-cell expression program. This means that while these cells originate from B cells, they lose the typical B-cell markers and functions. The defective expression program results in HRS cells showing atypical or reduced levels of B-cell-specific proteins and features, contributing to the pathogenesis of CHL and complicating the diagnosis and characterization of the disease. Although Hodgkin and Reed-Sternberg (HRS) cells are rare within the lymph node cell population, they, along with their surrounding cellular context with T and B lymphocytes, eosinophils, neutrophils, histiocytes, plasma cells, fibroblasts, and collagen fibers, play a crucial role in characterizing the disease [7, 8]. Pathologists use microscopic analysis at various magnifications to examine different tissue structures and cellular composition in tissue sections. In addition, they carefully search for HRS cells to obtain an accurate diagnosis of the disease.

The subtypes of CHL have different histologic features. Analyzing these features at different magnifications is crucial for an accurate diagnosis. MCCHL shows a rich tumor microenvironment under the microscope. At higher magnification, the HRS cells are scattered throughout the tissue and surrounded by a reactive infiltrate of immune cells such as eosinophils, neutrophils, plasma cells, and histiocytes. In contrast, NSCHL is characterized by collagen bands surrounding at least one nodule with HRS cells, which become visible at lower magnification. LRCHL shows scattered HRS cells predominantly of small lymphocytes in a nodular or diffuse background. Finally, LDCHL is characterized by a microenvironment with many HRS cells and a lack of normal background lymphocytes.

Current challenges in diagnosing CHL subtypes exist despite the histological clues mentioned above. While features like collagen bands, lymphocyte count variations, and mixed inflammatory infiltrates offer clues to specific subtypes, these characteristics can be subtle and often overlap. This, coupled with the possibility of biopsies exhibiting mixed subtype features, contributes to significant inter-rater variability among pathologists, particularly for less common subtypes (LRCHL and LDCHL). The limited experience with these rare subtypes in small institutions further complicates accurate subtyping due to a lack of experience with their specific diagnostic nuances.

Ultimately, automating the CHL diagnostic workflow with whole slide images (WSI) stained with Hematoxylin and Eosin (H&E) can help eliminate inter-rater variability arising from pathologists' experience level and the complexity of subtyping, ensuring consistent diagnosis. Several deep-learning methods have been proposed for automatically diagnosing lymphoma types based on H&E-stained slide images. While some of them included two lymphoma types [9, 10, 11, 12] as binary classification, others performed an extended study of lymphoma subtypes [13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23]. In the literature, the most extensive study of lymphoma was carried out by [13], which focused on classifying eight types of lymphoma, treating Classical Hodgkin lymphoma as a single category. Due to the challenge of having a limited number of samples for each type, they introduced an interpretable machine-learning approach named LymphoML. This method leverages the morphological features from whole slide images for lymphoma subtyping through feature engineering. Additionally, the works of [16, 20] introduced CNN-based deep learning methods aimed at the binary classification of Diffuse large B-cell lymphoma (DLBCL) versus non-DLBCL, with the latter group encompassing CHL samples. Lastly, Hashimoto et al. [23] designed a CNN-based multiple instance learning framework for identifying six lymphoma subtypes, including the two predominant subtypes of CHL: MCCHL and NSCHL.

Although these studies included Classical Hodgkin lymphoma as a single type [13] or as one side of a binary classification with other lymphoma types [16, 20] or included two subtypes of CHL in a multiclass classification task [23], none of them addressed the classification of CHL subtypes. This gap in the existing literature forms

one of the sides of the current research.

1.2 Objectives and Contributions

The main contributions of the thesis to the current literature are:

1. We developed an advanced deep-learning pipeline for classifying Classical Hodgkin lymphoma (CHL) subtypes. This pipeline demonstrates superior performance compared to state-of-the-art weakly supervised models, filling a critical gap in the literature on CHL subtyping.

2. Our research provides an in-depth evaluation of the latest Self-Supervised Learning (SSL) foundation models specifically adapted for histopathology. We focused on assessing six foundation models, each trained on extensive datasets that include a wide range of tissues and cancer types.

3. We compiled a dataset of 1,247 whole slide images of Classical Hodgkin lymphoma, sourced from three medical centers in Türkiye. These images were carefully selected, ensuring the quality and relevance of the dataset. This comprehensive dataset is set to be a valuable resource for advancing deep learning and histopathology research.

2. LITERATURE ANALYSIS

2.1 Weakly Supervised Learning and Multiple Instance Learning (MIL)

In deep learning, "weakly supervised learning" refers to scenarios where the training data is partially labeled. In histopathology, a single label is usually assigned to an entire WSI, labeling only at the slide or patient level. Due to the enormous size of these slides, it is impractical to use them entirely in deep learning frameworks. Therefore, WSIs are divided into smaller, manageable patches for analysis. However, manually labeling the numerous patches is time-consuming and tedious due to their sheer quantity. A common practice is to assign the slide or patient label to each patch to train supervised models. This approach can introduce noise in the patch-label pairs, as not all tissue areas in a WSI may contain disease-specific patches.

Multiple instance learning (MIL) is a powerful technique suitable for this weakly supervised setting. Crucially, once subdivided into patches, all the patches form a "bag" that retains the WSI label. In MIL, each patch within the bag is considered an instance, but the model does not require individual labels for each patch. Instead, the learning process focuses on the overall label of the bag. The key assumption in MIL is that a WSI containing a particular disease will include at least one instance with disease-specific characteristics. Thus, if at least one patch in the bag demonstrates these characteristics, the entire WSI is classified as positive. By analyzing the features extracted from each patch and leveraging the single slide-level label, MIL algorithms can learn to identify these distinguishing features. In this way, they can make accurate predictions at the WSI level and effectively use the limited labeling information available without manually labeling each patch or introducing noisy labels.

2.1.1 MIL Aggregation

In the literature, there are two main approaches to multiple instance learning for histopathology tasks, using whole slide images. These approaches are called embedding-based MIL (bag-level MIL) and instance-based MIL. Both approaches include a first step of feature extraction from patches. Then, the extracted features, called feature embeddings, are aggregated to obtain a slide-level prediction.

The instance-based MIL attempts to identify a subset of representative WSI patches contributing significantly to the overall classification. In this method, the feature embedding of each patch is assigned a score indicating its importance within the bag. These individual patch predictions are then aggregated using max-pooling or mean-pooling [24]. However, instance-based MIL might misclassify cancers with few predictive patches, highlighting its limitations and leading to a preference for embedding-based approaches in the literature.

Embedding-based MIL focuses on creating a final representation encompassing the entire WSI (bag) by summarizing individual patch embeddings. The resulting bag representation is then fed into a classifier for the final prediction. Various pooling strategies can be used for aggregation, such as max-pooling [25, 26] distribution-pooling [27], and attention pooling mechanisms [28, 24, 29]. Attention mechanisms such as classical attention [24] or clustering-constrained attention [28] have proven superior to mean and max pooling, as they offer interpretability through the attention scores assigned to the individual tiles.

The development of vision transformers (ViTs) in computer vision has led to the proposal of MIL pooling methods that utilize these architectures [30, 31, 32]. Self-attention in ViTs weights each instance in WSI by computing the relationship between the instances. Although they are very suitable for histopathology, the large number of WSI fields can significantly increase the computational cost of ViT models. In addition, training these models requires a large amount of data. Early studies have investigated transformer architectures that solve these limitations, such as the Nyströmformer used

in [30]. This model uses two layers of vision transformers with positional encoding and performs better than MIL models such as [28, 25, 24]. In addition, multilevel approaches have been proposed in studies [31, 32]. This multilevel approach allows the model to capture features at different scales, which improves its ability to classify high-resolution tissue images. Some studies have combined ViTs with instance selection methods [33, 34]. This integration makes it possible to select the most predictive patches and thus reduce computational costs.

Furthermore, several research combine embedding-based and instance-based methods [25, 35, 36, 37]. These hybrid models aim to leverage instance-based methods to identify critical instances and generate a final bag embedding, thereby fully utilizing the strengths of both approaches. [25] uses max-pooling to select the critical instance and aggregates the final bag embedding using this instance. Additionally, studies explore innovative combinations such as sparse convolutions with max-pooling alongside embedding-based pooling [38] and integrating ViTs with prototypical learning for enhanced bag- and instance-level supervision [35]. These approaches underscore the ongoing evolution and refinement of MIL techniques in computational pathology.

2.2 Feature Extraction Methods

Feature extraction is a crucial step in digital pathology, playing a key role in analyzing large, high-resolution images. This process converts raw patches of whole slide images into meaningful, lower-dimensional representations, facilitating the characterization of tissue structures, cellular properties, and pathological patterns. These extracted features are essential for various WSI tasks, such as classification and prognosis. A major limitation of feature extraction is the inherent trade-off between dimensionality reduction and information preservation. The generalization ability of feature extraction depends on the domain and quality of the training dataset and the availability of labeled data. Therefore, selecting a method that maximizes essential information retention while accurately mapping pathological features in WSIs is crucial.

2.2.1 Convolutional Neural Networks

The digital pathology domain often lacks annotated patch-based or pixel-based datasets for WSIs, and preparing these datasets is labor-intensive and time-consuming due to the large number of patches involved. As a result, traditional supervised convolutional neural networks (CNNs) trained on natural images have been widely used in pathology research for feature extraction, achieving remarkable results in various applications [28, 30, 31, 39, 40, 41]. The most widely used and effective generic CNN model is ResNet (ResNet-14 [31], ResNet-18 [41], and ResNet-50 [28, 30, 40, 39]), which was trained with a large number of natural images (ImageNet). [28] adapted the pre-trained ResNet model by adding an adaptive mean-spacial pooling layer to output feature embeddings of size 1024. Furthermore, [42] used a VGG model pre-trained on ImageNet for feature extraction. Although the CNN models trained on natural image datasets have shown promising results on histopathology tasks through transfer learning, the domain shift between whole slide images and natural images is huge and limits the performance of MIL tasks. The reason is that histopathology images, unlike natural images, have complex morphologic and cellular structures. Therefore, an extraction model trained for a specific data type does not generalize well to the distributions of another image domain. To overcome the domain shift in histopathology images, [43] developed an in-domain pre-trained CNN model called KimiaNet based on the DenseNet121 architecture, which was tuned on 7126 TCGA whole slides in a supervised manner. [44] tested the performance of KimiaNet with a marginal improvement in classification using state-of-the-art MIL methods.

2.2.2 Self-Supervised Learning (SSL)

Conventional methods depend on large datasets with extensive manual annotations. However, in histopathology, such datasets are mostly unavailable. As a result, the available datasets generally only have patient-level or slide-level labels. Self-supervised learning offers an alternative by leveraging the inherent structure and patterns in unlabeled images of whole slides. SSL generates pseudo-labels or supervised

signals from the unlabeled data that guide the model to learn informative features without explicit annotation.

SimCLR, a well-known model, uses contrastive learning on a ResNet architecture. [25] have trained SimCLR [45] with small histopathology images from the Camelyon16 and TCGA lung cancer datasets. More recently, efficient transformer models such as Swin Transformer [46] for SSL in histopathology [47] have been explored. These models offer advantages such as transferable features compared to traditional CNNs. However, they often reach their limits. These models are usually trained on limited histopathology datasets (public or private) and often focus on specific cancer types. This limits their generalizability to broader histopathology domains and requires fine-tuning for specific downstream tasks.

While the scarcity of labeled data presents a challenge, researchers are pioneering the development of self-supervised learning (SSL) foundation models trained on massive slide-level labeled histopathology datasets. These efforts are promising for rare cancers, where only a few samples are available from a single medical center, making it impossible to fine-tune feature extractor models. Histopathology foundation models trained on massive datasets, encompassing a wide range of diseases from common cancers to rare diseases affecting multiple organs, offer a potential solution. These foundation models can be applied to downstream tasks without fine-tuning and provide better results compared to other methods.

A well-known SSL technique is contrastive learning (CL). This method forces the network to bring similar images (e.g., augmented versions of the same image) closer together in a latent space and to drive dissimilar image representations further apart using a contrastive loss function. [48] used the SimCLR model with contrastive learning. In contrast to the limited training datasets of [25, 47], they trained the model with a diverse dataset that includes 22 tissues, utilizing publicly available datasets such as The Cancer Genome Atlas Program (TCGA) and Clinical Proteomic Tumor Analysis Consortium (CPTAC). [49] also used SimCLR with contrastive learning in their model training with the TCGA dataset. [50] introduced CTransPath, a hybrid SSL model

that uses a novel contrastive learning method called semantically-relevant contrastive learning (SRCL). Unlike traditional CL, which only compares two views of the same instance, SRCL finds numerous patches that share similar visual features of WSIs and uses them as positive instances in model training. The model was trained on a large dataset of over 30,000 unlabeled WSIs from TCGA and PAIP, covering 25 anatomical regions and 32 cancer types.

In contrast to the previously mentioned models' single-modality learning, [51] has introduced Pathology Language-Image Pretraining (PLIP), a vision-language model that is trained with contrastive learning on both images and the corresponding text descriptions. For this purpose, a large OpenPath dataset with 208,414 image-text pairs was collected from Twitter and other datasets. Another visual-language histopathology foundation model trained with image-text pairs is CONCH (CONtrastive learning from Captions for Histopathology) [1]. To create image-caption pairs, they used public PubMed research articles. They also included 21,442 WSIs with 350 cancer subtypes, patient reports, and electronic medical records from a private institute and used all this data for model training. The CONCH model outperformed the PLIP model on several downstream tasks.

While previous works used SSL with contrastive learning in their model training, [52] introduces BROW, a novel deep-learning model that uses transformers and self-distillation for histopathology feature extraction. BROW utilizes a hierarchical approach by analyzing image sections at multiple zoom levels and applying data augmentation. [53] presented Phikon, a novel histopathology foundation model that uses iBOT and masked image modeling (MIM) [54]. In MIM, the model is trained to learn random masked image content. This strategy forces iBOT to learn meaningful representations from many unlabeled histopathology data. The Phikon model outperformed CTransPath [50] on several downstream tasks. The authors also show that the proposed model is robust to variations in the downstream tasks and has superior fine-tuning ability.

Some of the above models use publicly available histopathology datasets such

as TCGA and PAIP [50, 53] or combine them with their private datasets to train their models. The public datasets differ in the quality of their digitized slides and have limited sample size to build a foundational model. Additionally, utilizing these public datasets can introduce biases due to the potential for data leakage. To address these challenges, [55] proposed a model trained on nearly 1.5 million H&E-stained WSIs collected from Memorial Sloan Kettering Cancer Center (MSKCC). This dataset is the largest in the literature and includes benign and cancer slides. In this study, transformers were used as the backbone, and DINOv2 was used as the SSL approach. Like iBOT, the DINOv2 model is another SOTA SSL method and uses self-distillation and masked image modeling to create an image embedding space. This model performed better than [50, 53, 51] on several downstream tasks.

Another study that does not use publicly available datasets but a private dataset is Prov-GigaPath [56]. They proposed a model using Vision Transformers as the backbone for training with a dataset of 171,189 H&E WSIs with 31 tissue types. The model's architecture consists of a DINOv2 model for patch embedding followed by a slide-level embedder called LongNet. Gigapath performed superior to other SSL models [50, 49] in various tumor subtyping tasks. UNI [57] is another foundation model trained with a ViT-L backbone and DINOv2 SSL approach. This model was trained on a diverse private dataset of 100,426 FFPE H&E stained WSIs called Mass-100K with 20 major tissues. They tested the model on 34 different downstream tasks and performed better than [50, 49] on most benchmarks.

3. MATERIALS AND METHODS

3.1 Datasets

For this study, we compiled three datasets of hematoxylin and eosin (H&E) stained WSIs from three institutions: Ege University Hospital Pathology Laboratory (EUH dataset), İstanbul University Hospital Pathology Laboratory (IUH dataset), and Başkent University Hospital Pathology Laboratory (BUH Dataset). All datasets include four subtypes of CHL. The WSIs of all datasets are labeled at slide level, which means different slides from the same patient have been included. The datasets were divided into training, validation (internal validation), and testing (external validation) groups. The data collected from one center was designated the test set, providing the model with previously unseen data to evaluate its ability to generalize across different centers. Datasets collected from the other two centers were used for training and validation with an 80/20 split ratio. All slides from a single patient were consistently assigned to either the training or the validation set, ensuring that slides from the same patient did not appear in both sets. The summary of the datasets can be found in Figure 3.1B.

EUH dataset: This dataset contains 714 weakly-labeled H&E-stained biopsy images from 393 patients, with an average of 2 biopsy slides per patient included to compensate for the rarity of CHL subtypes. The dataset contains a mixture of file formats, with 621 WSIs in .svs format and the remainder in .ndpi format. Two board-certified pathologists thoroughly reviewed the slide-level labels to minimize labeling-related bias. We selected this dataset to train and validate the proposed and other models used in this research.

IUH dataset: This dataset consists of 365 H&E-stained biopsy slides from 206 patients, containing an average of 2 biopsy slides per patient. All slides in this dataset are in .ndpi format. Similar to the EUH dataset, the labels of the WSIs were reviewed

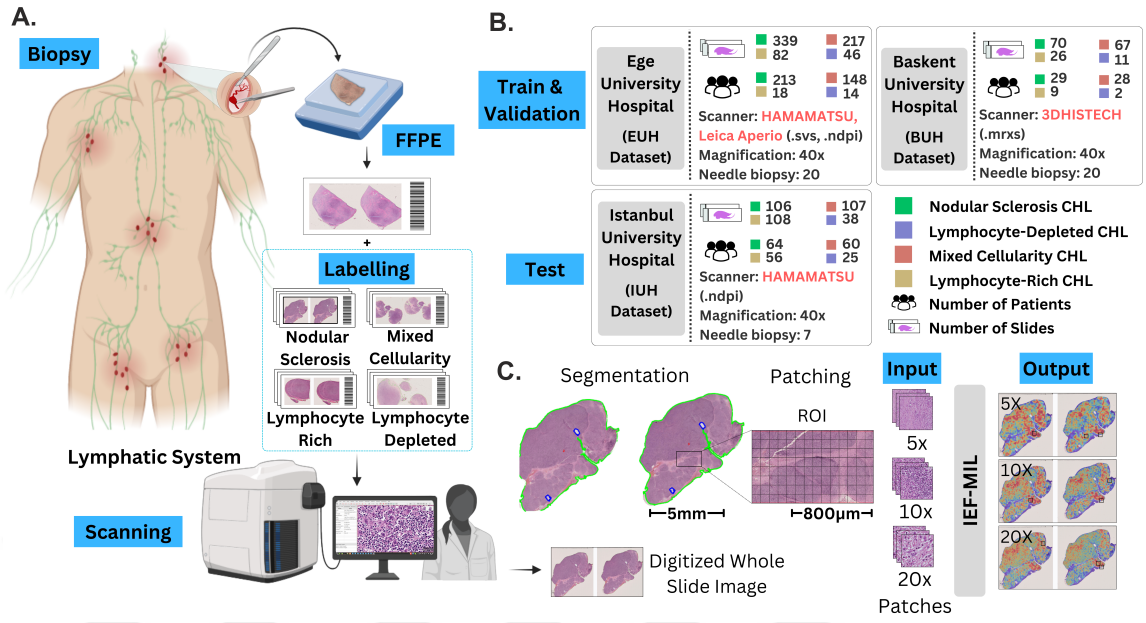

Figure 3.1 Workflow Overview. **A.** Lymph nodes are surgically removed, fixed in formalin, embedded in paraffin (FFPE), sliced, and stained with Hematoxylin and Eosin (H&E). Stained slices are scanned to create digitized whole slide images (WSIs). **B.** The study uses three datasets from Türkiye: EUH, BUH, and IUH, with patient and slide counts noted. WSIs are scanned at 40x magnification. The EUH and BUH datasets are combined for training and validation with 5-fold cross-validation, while the IUH dataset is used as an independent cohort. **C.** Tissue is segmented from WSIs into patches across three magnification levels. These patches are used for training, validation, and testing in a Multiple Instance Learning (MIL) network, which produces slide-level predictions and heatmaps for each magnification.

by two board-certified pathologists from the Ege University Pathology Laboratory to ensure consistency and minimize interobserver variability in labeling. As a result, the labeling of 32 WSIs was revised, representing almost 9% of the IUH dataset. After reviewing the slide labels, six WSIs were excluded from the IUH dataset due to poor tissue quality and insufficient staining. We obtained a final dataset of 359 samples. We selected this dataset as the independent test cohort for our experiments.

BUH dataset: This dataset includes 174 H&E-stained biopsy slides from 68 patients, with an average of 3 biopsy slides per patient. All slides in this dataset are in .mrxs format. To account for the rarity of LDCHL and LRCHL subtypes, we included 2-6 WSIs from one patient. Two board-certified pathologists from the Ege University Pathology Laboratory reviewed the labels of 10% of this dataset. This dataset served as a secondary center for the training and validation processes, increasing the diversity of images, which is critical for creating a generalizable deep learning model.

All datasets were collected according to ethical guidelines. Patient anonymity was ensured by removing all identifiable information during the data collection.

3.2 Whole Slide Image Preprocessing

WSIs, digital histopathological slides, provide high-resolution details of tissue morphology at different magnifications. A typical WSI scanned at 40x magnification has an enormous average size of 100,000 x 100,000 pixels. A WSI scanned at 40x magnification (207,872 x 86,016 pixels) would reduce to half its size at 20x magnification (103,936 x 43,008 pixels), and at lower magnifications, the shape decreases proportionally. Whole slide images have a pyramidal structure in which the same entire slide image is stored at different zoom levels. The number of zoomed-down image levels stored in the pyramid can vary depending on the selected file format, e.g., .svs, .mrxs, .tiff, and .ndpi.

In this study, three scanners were used to digitize the histopathological biopsy slides: Hamamatsu NanoZoomer S60, 3DHistech Pannoramic 250 Flash III, and Leica Aperio AT2. This resulted in three distinct file formats: .ndpi, .mrxs, and .svs, with all slides scanned at 40x magnification. Due to the large size of the slides, a preprocessing pipeline was employed, which included segmentation and patching steps. All preprocessing and feature extraction tasks were carried out using the CLAM tool [28]. **Segmentation:** The Otsu method was utilized to segment the tissue regions of each whole slide image (WSI) from the background. This method effectively removes irrelevant and empty areas, isolating the tissue of interest for subsequent analysis. **Patching:** Following segmentation, the isolated tissue was divided into non-overlapping patches of size 224 x 224 pixels. This patching process was conducted at three different magnifications: 20x, 10x, and 5x (Figure 3.1C).

To achieve a balance between capturing sufficient tissue detail and maintaining computational efficiency, we carefully selected both the size of the patches and the magnification. We opted for a common patch size of 224 x 224, compatible with widely

Table 3.1
Magnifications and pyramid levels for .ndpi, .mrxs, and .svs WSI file formats. Level 0 is 40x magnification. Magnification for each pyramid scale is calculated as Magnification = $\frac{40x}{\text{Scale}}$. Values for lower pyramid levels with extensive decimal places are excluded.

Pyramid level	**1**	**2**	**4**	**8**	**16**	**32**	**64**	**128**	**256**	**512**	**1024**
Magnification	40x	20x	10x	5x	2.5x	1.25x	...	...	...	...	...
.ndpi	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
.mrxs	✓	✓	✓	✓	✓	✓	✓	✓	✓		
.svs	✓		✓		✓		✓				

used pre-trained feature extraction architectures in deep learning applications. To ensure comprehensive analysis, we chose a range of magnifications (20x, 10x and 5x). This allows us to capture the broader tissue context at lower magnifications and finer cellular details at higher magnifications. In addition, we chose 20x magnification over 40x magnification because patches at 20x magnification contain more morphological structures and preserve contextual information, ultimately resulting in fewer patches and, therefore, fewer computational resources compared to 40x magnification.

Dealing with format-specific discrepancies: The file formats .ndpi, .mrxs, and .svs each include varying magnification scales. Specifically, .ndpi and .mrxs files offer eleven and nine magnification levels, respectively, whereas .svs files are limited to four magnification levels (Table 3.1). We used the following strategies to compensate for this discrepancy and obtain patches with the desired magnifications (20x and 5x) from .svs files. For patching at 20x magnification, we generated patches of shape (448 x 448) at 40x and resized them to (224 x 224) during feature extraction. To obtain patches at 5x magnification, the WSIs were cropped into tiles of shape (448 x 448) at 10x. These patches were then resized to (224 x 224) during feature extraction. For .ndpi and .mrxs files, patching was performed directly at the desired size (224 x 224) without overlap for all magnifications.

Feature extraction: After completing the patching stage, we derived low-dimensional features from the patches utilizing various models for feature extraction. These models were employed directly on our dataset without any modifications or fine-tuning. The extraction processes yield feature embeddings of varying dimensions, as

indicated in Table 3.2. Based on the feature extraction techniques used, which are elaborated in [53, 1, 57, 56, 55, 50], we normalized the images of all patches using a uniform mean vector (0.485, 0.456, 0.406) and standard deviation vector (0.229, 0.224, 0.225). This normalization process is crucial for centering and scaling the data, facilitating more effective model training. In our research, we used only the class-token of Virchow [55] and Phikon [53] models.

Table 3.2
Overview of feature extractor models used in this study. The asterisk (*) denotes the number of patches instead of WSIs.

	Model			Dataset		
	Architecture	SSL method	Size	Source	Size	Feature length
Phikon	ViT-B	iBOT	86M	TCGA	6K	768
CONCH	ViT-B	iBOT	86M	PMC-Path + EDU	1.2M*	512
UNI	ViT-L	DINOv2	307M	Mass-100K	100K	1024
Virchow	ViT-H	DINOv2	632M	MSKCC	1.5M	1280
CTransPath	Swin T.	SRCL	28M	TCGA + PAIP	32K	768
Prov-GigaPath	ViT-G	DINOv2	1B	Providence	171K	1536

Among the models, Virchow and CONCH both include whole slide images of lymphoma cancer, though they lack details on the specific types included. Phikon and CTransPath datasets each contain WSIs of one type of non-Hodgkin lymphoma but do not have any Hodgkin lymphoma samples. UNI's training dataset comprises WSIs from the lymphatic system without Hodgkin or non-Hodgkin lymphoma images, whereas Prov-GigaPath notably excludes any WSIs from the lymphatic system, thus not including lymphoma cases. Detailed specifications about the source of data, the size of the dataset used for training, the architecture of each model, along with the SSL methodology employed, and the length of the feature vectors are presented in Table 3.2.

3.3 Model Architecture

The overall pipeline of the proposed model **Instance & Embedding Fused Multiple Instance Learning (IEF-MIL)** is given in Figure 3.2. The model comprises two components: an SSL-based feature embedder module and a dual-stream MIL pooling module. The dual-stream MIL model has two branches; one utilizes an embedding-based pooling with global self-attention, and the other performs an instance-based mean pooling.

3.3.1 Self-Supervised Vision Transformer as a Feature Extractor

For this study, we performed transfer learning from a vision-language self-supervised learning foundation model called CONCH proposed in [1]. They created image-caption pairs using public PubMed research articles to train this multimodal model. They also included 21,442 WSIs with 350 cancer subtypes, patient reports, and electronic medical records from a private institute. The model consists of an image encoder and a text decoder part. The pre-training of the image encoder (feature extractor) is performed in two steps: 1. pre-training of the image encoder with the iBOT method [54] and 2. pre-training of the image encoder and the unimodal and multimodal text decoders with CoCa [58]. The basic framework of the image encoder is a Vision Transformer Base (ViT-B) model. The ViT-B model consists of 12 layers of encoders. Each encoder consists of a Multi-Head-Self-Attention (MSA) with 12 heads and a Multi-Layer-Perception (MLP) block, a feed-forward network. The projection head consists of 3 layers of multi-layer perception and a bottleneck with $l2$ normalization. The patch size is 16, and the generated embedding size is 768.

The vision encoder of the proposed model is first trained with iBOT, which uses the Masked Image Modeling (MIM) method. iBOT masks random patches of the image and requires the model to reconstruct the missing patches. The objective is to learn meaningful image representations by minimizing the discrepancy between the masked and complete image token distributions. iBOT achieves this by using a self-

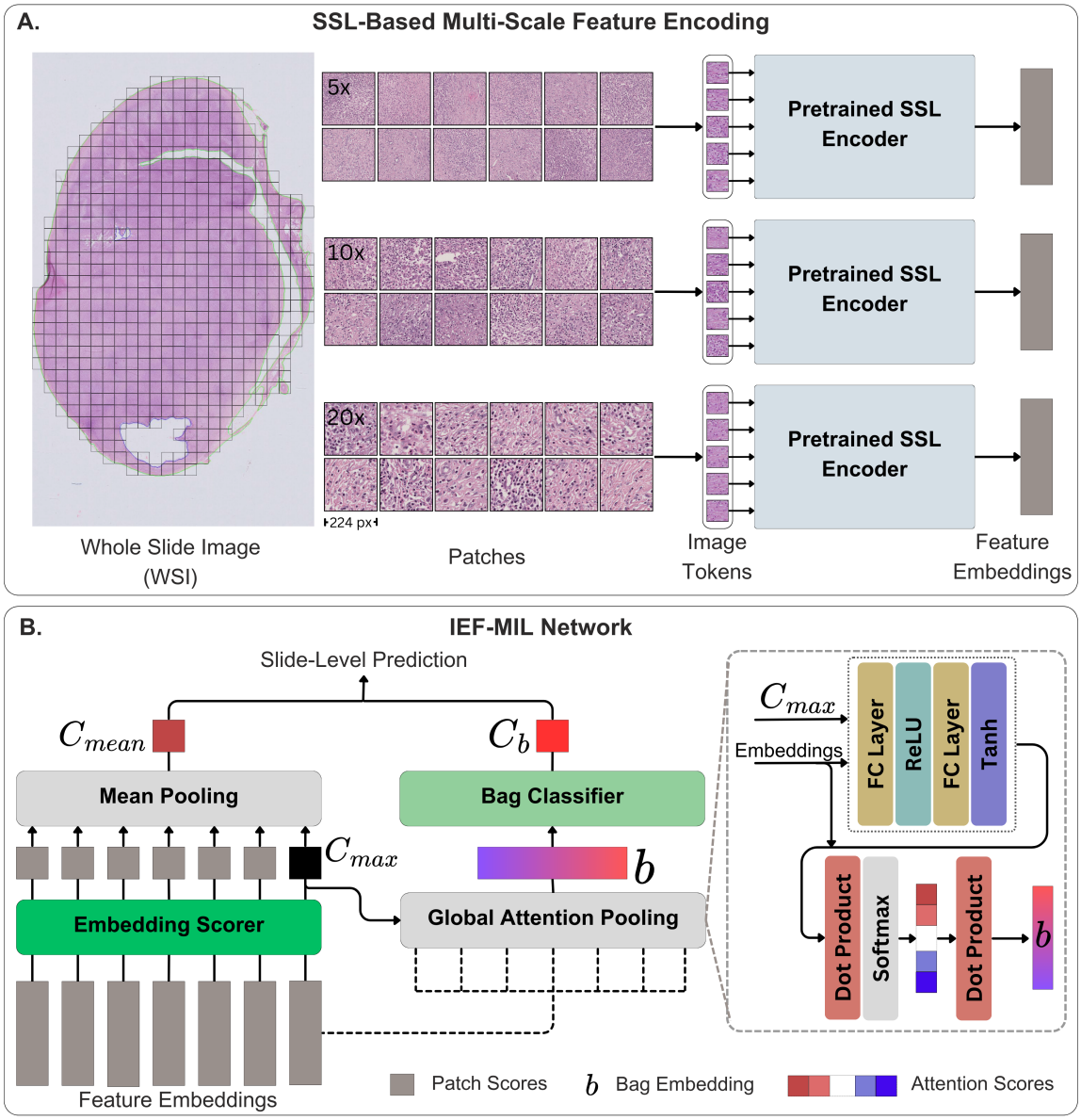

Figure 3.2 Overview of the Proposed MIL Pipeline. A. Preprocessing step showing multi-scale feature extraction from patches at 5x, 10x, and 20x magnifications, with embeddings concatenated for model training, validation, and testing. The CONCH model [1] is used for feature extraction with transfer learning. **B.** Architecture of the IEF-MIL model, featuring instance-based pooling (mean pooling of embeddings) and embedding-based pooling (global attention pooling). The final slide score and prediction are computed by averaging the mean and bag scores.

distillation technique with two identical networks: a student network (target) and a teacher network (tokenizer). The student network tries to predict the masked content of the patches by using the Exponential Moving Average (EMA) for momentum updates. Meanwhile, the teacher network trained on complete images serves as a reference point. Unlike predefined offline tokenizers, the teacher network is updated during the training of the student networks and acts as an online tokenizer. To pre-train the iBOT model,

a projection head is added to the backbones of both the student and teacher networks during iBOT training. Patches of 256 x 256 at 20x magnification were generated from whole slide images for training the unimodal image encoder. Likewise, patient reports and electronic records were used to train the unimodal text decoder.

The second step is the pre-training of the image encoder (student network of iBOT) and the text encoder with the CoCa SSL method [58]. The CoCa training requires an unimodal image encoder, an unimodal text decoder, and a multi-modal text decoder. Two losses are calculated during the training of this model. A contrastive loss $\mathcal{L}_{Con}$ is calculated between the class tokens of the unimodal image encoder and the text decoder. With the contrastive loss, the text and the corresponding image representations are aligned to the same feature space. The representations learned from the image encoder and text decoder are then fed into the multimodal decoder to generate a unified image-text representation. The second loss, the so-called captioning loss $\mathcal{L}_{Cap}$, is calculated between the target text representation and the generated unified representation. In the end, these two losses are combined: $\mathcal{L}_{CoCa} = \lambda_{Con} \cdot \mathcal{L}_{Con} + \lambda_{Cap} \cdot \mathcal{L}_{Cap}$ where λ_{Con} and λ_{Cap} are hyper-parameters that weight two losses. An image size of 448 x 448 is used to train the multimodal model with image-caption pairs obtained from open-access PubMed publications. Further training parameters for the two-stage training of the model can be found under [1].

To extract features from the patches of CHL whole slide images, the vision transformer backbone of COCNH was used as shown in Figure 3.2A.

3.3.2 MIL Aggregator

The main objective of the multi-class multiple instance learning problem is to predict a bag label Y for the given bag X. X is a whole slide image bag with N instances. The bag of patches is $X = \{x_1, x_2, \ldots, x_N\}$. An embedding vector is obtained using a feature extractor f such that $x'_i = f(x_i) \in \mathbb{R}^{L \times 1}$. The bag of features for all instances is $f_X = \{x'_1, x'_2, \ldots, x'_N\}$, and the bag label is $Y \in \{0, 1\}^K$. Each

instance is a vector with dimensions L, and $Y = [Y^1, Y^2, \ldots, Y^K]$ is a one-hot encoded label vector where $\sum_{k=1}^{K} Y^k = 1$. For example, the bag X belongs to class k where $k \in \{1, 2, \ldots, K\}$, and the class label is set to 1 in a one-hot vector ($Y^k = 1$).

Feature vectors extracted from different magnifications (5x, 10x, and 20x) are concatenated to create a bag of embeddings. The bag is then processed by the MIL model Figure 3.2B, which employs a dual-stream approach:

1. **Instance-Based Pooling**

   In this pooling, an embedding scorer $h(\cdot)$ calculates a score for each instance x_i using the feature vectors x_i'. The mean value of the individual instance scores is then calculated using the mean pooling operation g_{mean}:

$$C_{\text{mean}} = g_{\text{mean}}\left(h(x_1'), \ldots, h(x_N')\right) = \frac{1}{N}\sum_{i=1}^{N} W_0 x_i' \tag{3.1}$$

   where W_0 is the weight vector obtained from the embedding scorer $h(\cdot)$, and g_{mean} represents the mean pooling operation.

2. **Embedding-Based Pooling**

   The embedding-based pooling part of the model uses global attention pooling. The global attention pooling first selects the instance embedding with the maximum score ($C_{\max}$) as the critical instance. It creates a bag embedding b by aggregating all feature vectors in the bag f_X using this critical instance.

   All instances in f_X, including the critical instance, are used to create query $q_i \in \mathbb{R}^{L\times 1}$ and information $v_i \in \mathbb{R}^{L\times 1}$ vectors using a non-linear network denoted as $h_b(\cdot)$. W_q and W_v are weight matrices where $q_i = W_q x_i'$ and $v_i = W_v x_i'$, with $i = 1, \ldots, N$. Then, the similarity measurement between the critical instance q_{max} and instance q_i is:

$$S_i = \frac{\exp(\langle q_i, q_{max}\rangle)}{\sum_{k=1}^{N}\exp(\langle q_k, q_{max}\rangle)} \tag{3.2}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of two query vectors. The bag embedding b is obtained by summing the information vectors v_i that are weighted with the calculated similarity scores.

$$b = \sum_{i=1}^{N} S_i v_i \tag{3.3}$$

The aggregated bag representation b is fed into the bag classifier, which computes the bag score C_b:

$$C_b = g_b(x'_1, \ldots, x'_N) = W_b \sum_{i=1}^{N} S_i v_i = W_b b \tag{3.4}$$

Here, W_b is the weight vector derived from the classifier, and g_b represents the embedding-based pooling operation.

The final slide-level score is calculated by averaging the mean instance score C_{mean} and the embedding-based score C_b:

$$C_{\text{final}} = \frac{1}{2}(C_{\text{mean}} + C_b) \tag{3.5}$$

This approach leverages both instance-based and embedding-based information to provide a robust classification for the WSI bag.

3.4 Training Details and Hyperparameters

In this study, we employed the CLAM tool [28] for all project stages, including patching, feature extraction, training, testing, and inference. We conducted a 5-fold cross-validation on the training dataset. The proposed model, IEF-MIL, and other MIL models were trained from scratch. The training parameters include Adam optimizer, cross-entropy loss function, 1 batch size, and 2e-4 learning rate. Early stopping was implemented with a patience of 15 epochs and a stopping epoch of 50, resulting in a

variable number of training epochs with a maximum limit of 100. The patience is the number of epochs to wait after no improvement in validation loss is observed, while the stopping epoch defines the earliest epoch to stop training if no improvement is observed. The random seed was set to 1.

The feature extractor models used in this thesis were sourced from various repositories. The code and weights for the CTransPath model were obtained from their GitHub project repository. Additionally, HuggingFace (https://huggingface.com) provided the embedder model codes for Phikon, CONCH, UNI, Prov-GigaPath, and Virchow. Furthermore, the codes of CLAM, TransMIL, DSMIL and ACMIL were taken from their GitHub project repositories.

3.5 Evaluation Metrics

To evaluate the performance of our multiclass classification model on an imbalanced dataset, we employed multiple metrics. These included the weighted F1 score, ROC-AUC score, confusion matrix, and precision-recall (PR) curves per class. Given the imbalance in our dataset, where one class significantly outnumbers the others, many true negative predictions may overshadow the impact of false positives on the model's evaluation. This imbalance could lead to an overly favorable AUC score, as the AUC score is less sensitive to class imbalance. To address this issue and accurately reflect the model's actual performance, we also employed the weighted F1 score and PR curves.

The weighted F1 score calculates the harmonic mean of precision and recall, given by F1 Score $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$, across all classes and provides a comprehensive assessment. The PR curves illustrate the trade-off between precision (Precision $= \frac{\text{TP}}{\text{TP+FP}}$) and recall (Recall $= \frac{\text{TP}}{\text{TP+FN}}$). In addition, the confusion matrix provided a raw picture of correctly classified (true-positive and true-negative) and misclassified instances (false-positive and false-negative) across classes.

All metrics were computed using Scikit-learn (v1.3.2).

3.6 System Specifications

- Hardware: This thesis utilized high-performance graphics processing units (GPUs) to accelerate deep learning computations. The setup featured one NVIDIA GeForce RTX 4090 with 24 GB of memory and two NVIDIA GeForce RTX 3090 GPUs, each with a memory of 24 GB. Experiments were conducted on two systems with x86_64 CPU architecture and eight cores. A single GPU was employed to train MIL and feature extraction models. Approximately 10 TB of hard disk space was allocated for the study, primarily to store whole slide images and associated processed files (.h5 and .pt).
- Software: The computing environment was based on the Ubuntu 22.04 LTS operating system. Python (v3.10) and PyTorch (v2.3.1 with CUDA 11.8) were used, along with Pip as the package manager. Key packages included openslide-python (v1.3.1) and pytorch-lightning (v2.3.2).

4. RESULTS

In this study, we evaluate the performance of our proposed IEF-MIL model by comparing it with five other state-of-the-art MIL models: CLAM-SB, CLAM-MB [28], TRANSMIL [30], ACMIL-GA [29], and DSMIL [25]. All models were trained from scratch to ensure a fair comparison. For the training of CLAM-SB, CLAM-MB, TRANSMIL, and ACMIL-GA, we utilized feature vectors derived from a single magnification (20x). In contrast, DSMIL was trained using a combination of features from both 10x and 20x magnifications in alignment with its architectural specifications [25]. Our proposed IEF-MIL model leverages feature embeddings concatenated from 5x, 10x, and 20x magnifications, which were extracted using foundation models through transfer learning. This multi-magnification approach allows for a more comprehensive feature representation.

Additionally, we conducted an extensive evaluation of self-supervised learning foundation models [55, 1, 50, 53, 56, 57] on the IUH test set to further assess their performance. We provide a qualitative assessment of the IEF-MIL model through heatmaps generated from attention scores, offering insights into its decision-making process. Furthermore, ablation studies were performed to analyze the impact of integrating features from different magnifications on the overall model performance.

4.1 Classification Results on Test Set

Table 4.1 presents the results from a 5-fold cross-validation, highlighting the performance of all trained models using feature vectors extracted from six different models. Notably, our IEF-MIL model outperformed other advanced MIL models when trained with CONCH features. Specifically, our model achieved an AUC of 0.905 ± 0.016 and a weighted F1 score of 0.689 ± 0.05 on the testing dataset, demonstrating superior performance metrics. The CLAM-MB model, also trained with CONCH features, recorded

Table 4.1
Comparison of AUC (bottom) and F1 (top) scores (mean ± SD) for the IEF-MIL and other MIL models on a test set using features from six backbones. The top-performing MIL model for each backbone is underlined.

MIL Model	Backbone					
	CONCH	CTransPath	Prov-GigaPath	Phikon	UNI	Virchow
CLAM-SB	0.641±0.038	0.51±0.039	0.554±0.07	0.542±0.035	0.518±0.039	0.465±0.106
	0.889±0.019	0.831±0.022	0.861±0.011	0.829±0.022	0.844±0.009	0.782±0.045
CLAM-MB	0.683±0.033	0.532±0.035	0.545±0.077	0.635±0.034	0.591±0.052	0.529±0.061
	0.9±0.012	0.849±0.019	0.868±0.01	0.876±0.011	0.854±0.035	0.796±0.036
DSMIL	0.639±0.042	0.535±0.036	0.559±0.019	0.603±0.041	0.52±0.035	0.534±0.041
	0.878±0.029	0.829±0.015	0.842±0.016	0.846±0.027	0.831±0.019	0.821±0.021
TRANSMIL	0.638±0.086	0.501±0.068	0.541±0.052	0.618±0.04	0.516±0.095	0.453±0.072
	0.88±0.031	0.806±0.022	0.851±0.037	0.859±0.011	0.84±0.035	0.738±0.074
ACMIL-GA	0.595±0.094	0.482±0.057	0.515±0.079	0.639±0.027	0.535±0.079	0.485±0.137
	0.887±0.024	0.825±0.027	0.856±0.024	0.883±0.014	0.848±0.023	0.774±0.094
IEF-MIL	0.689±0.05	0.548±0.047	0.576±0.028	0.643±0.041	0.525±0.027	0.545±0.052
(ours)	0.905±0.016	0.85±0.014	0.873±0.01	0.885±0.007	0.862±0.014	0.831±0.03

mean AUC and weighted F1 scores of 0.683 ± 0.033 and 0.591 ± 0.052, respectively. Our IEF-MIL model also showed improved results over other MIL models utilizing feature embeddings from CTransPath, Phikon, Prov-GigaPath, and Virchow. CLAM-MB with UNI features achieved a weighted F1 score of 0.591 ± 0.052, outperforming the IEF-MIL model in this context.

Additionally, we evaluated various state-of-the-art (SOTA) foundation models trained with self-supervised learning. The IEF-MIL model with CONCH embeddings achieved an F1 score of 0.689 ± 0.05, outperforming other histopathology feature embedders. This highlights the superior generalizability and robustness of the CONCH SSL model for downstream tasks across diverse staining and scanning conditions. The Phikon model, while not as effective as CONCH, still performed well, with AUC and F1 scores of 0.885 ± 0.007 and 0.643 ± 0.041, respectively, when used with the IEF-MIL

model.

Examining the class-specific performance of our model trained with CONCH features, Figure 4.1 provides insights into individual class performance via precision-recall curves and an average confusion matrix. The precision-recall curves reveal the impact of false positives on class-wise performance. Despite the optimistic AUC value of 0.905 ± 0.016, the confusion matrix and precision-recall curves explain the lower F1 score of 0.689 ± 0.05, primarily due to class imbalance.

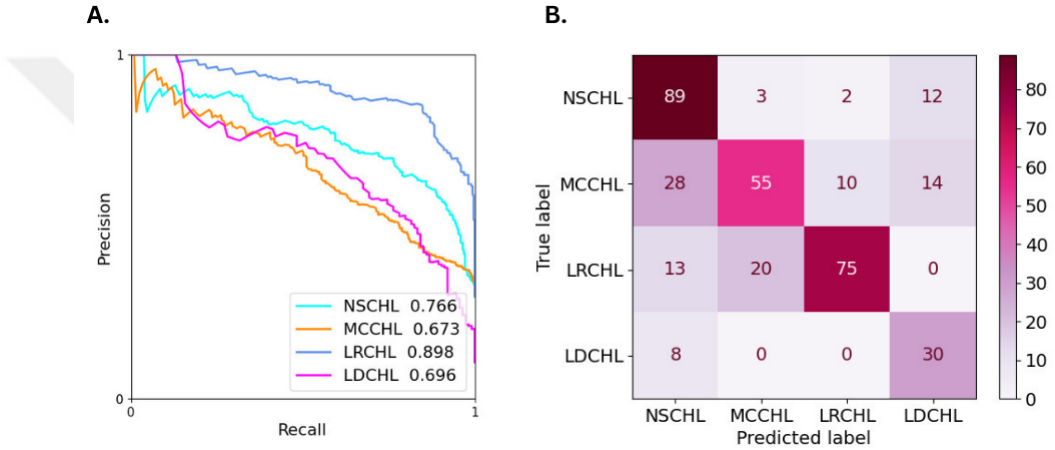

Figure 4.1 Class-based performance evaluation of the IEF-MIL model trained with CONCH features. **A.** Precision-recall curves and the area under the precision-recall curve (AUC-PR) provided for each class. **B.** Confusion matrix depicting the raw results of model predictions.

The IEF-MIL model exhibits varying performance in classifying the four subtypes of Classical Hodgkin Lymphoma (CHL). Among these, the LRCHL subtype demonstrates the highest precision and recall, achieving an AUC-PR of 0.898 and a substantial number of true positives, indicating its strong predictive capability. The NSCHL subtype also performs well, with an AUC-PR of 0.766 and a notable number of true positives, although it has slightly higher rates of false negatives and false positives compared to LRCHL.

In contrast, the LDCHL subtype shows moderate performance with an AUC-PR of 0.696, maintaining a balance between false positives and false negatives. The MCCHL subtype, with the lowest AUC-PR of 0.673, faces challenges due to a higher incidence of false negatives and false positives. The LDCHL subtype constitutes only

7.6% of the dataset, while the MCCHL subtype represents 33.8% of CHL samples. Despite the model being trained with fewer LDCHL samples, LDCHL outperforms MCCHL. This discrepancy is attributed to LDCHL's relative ease of classification, whereas MCCHL exhibits greater feature overlap with other classes, complicating its differentiation.

4.2 Ablation Study: The Performance of Multiple-Scales

We conducted ablation studies to evaluate how different magnification levels impact the performance of the IEF-MIL model. Our analysis explored three configurations: (i) using exclusively 20x magnification, (ii) combining 10x and 20x magnifications, and (iii) integrating 5x, 10x, and 20x magnifications. Figure 4.2 illustrates the average F1 scores and their standard deviations for each configuration.

The results indicate that the IEF-MIL model benefits significantly from incorporating features across various magnification levels. Notably, integrating multiple magnifications within the model training with CONCH features led to superior performance. This advantage is attributed to the extensive and varied training dataset of CONCH, which encompasses histopathology images across multiple magnifications within the PubMed dataset, thereby enhancing the model's robustness. The enhanced performance of the CONCH model, which utilized features across three scales, contrasts with other feature extractors that typically rely on single magnification levels, often 20x, with a fixed patch size [55, 50, 53, 56, 57].

The most notable performance was achieved with the CONCH model incorporating all three magnifications, resulting in an F1-score of 0.689 ± 0.05. This suggests that incorporating lower magnification patches (5x and 10x) significantly enhances model performance when using CONCH features. In contrast, similar improvements were not observed with other feature extractors. For example, Phikon features performed best with a combination of 10x and 20x magnifications, a trend also observed with UNI, Virchow, and CTransPath feature extractors. Meanwhile, the Prov-GigaPath model,

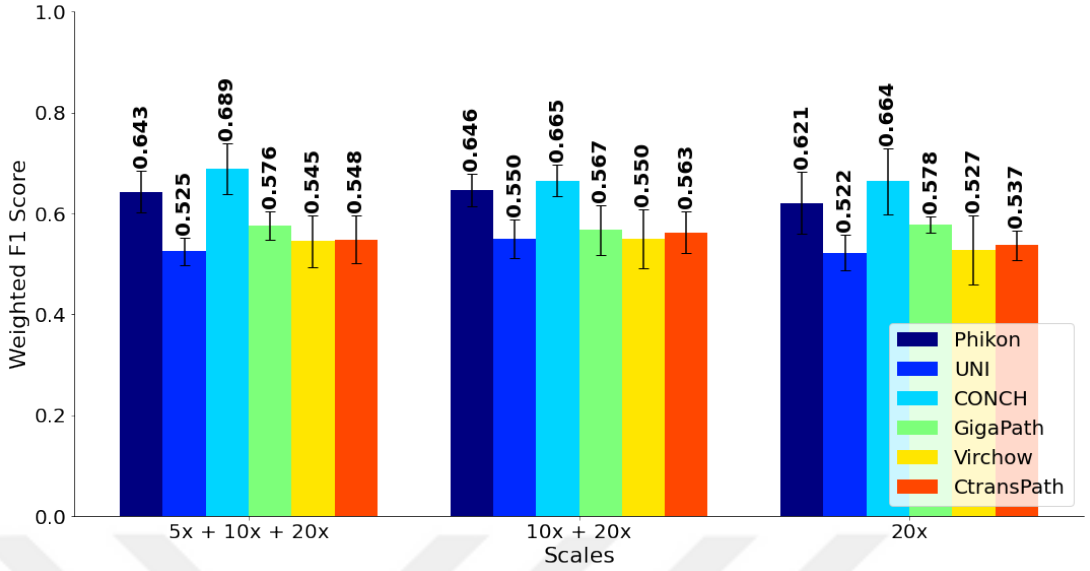

Figure 4.2 Comparison of F1 scores across different magnification combinations of feature vectors from six histopathology foundation models trained on the IEF-MIL model.

which utilized features from a single magnification (20x), outperformed configurations that combined different magnifications.

These findings underscore the value of incorporating patches from multiple scales to capture a broader spectrum of CHL subtype morphologies, which contributes to more accurate classification. Using diverse magnification levels enriches the model's ability to distinguish between features and improves overall performance.

4.3 Interpretability with Heatmaps

Figure 4.3 presents heatmaps for three magnifications (5x, 10x, and 20x) generated by the IEF-MIL model for the NSCHL, MCCHL, LRCHL, and LDCHL subtypes. These heatmaps, derived from attention scores, offer insights into the model's classification approach by highlighting regions of interest. The top three predictive patches are marked with red boxes, providing a closer view of the model's class-specific features and focal points. By displaying heatmaps at multiple magnifications, we improve interpretability, allowing for a detailed assessment of the model's performance across

different zoom levels. Red pixels in the heatmaps indicate areas with high attention scores, while blue pixels represent regions with low attention.

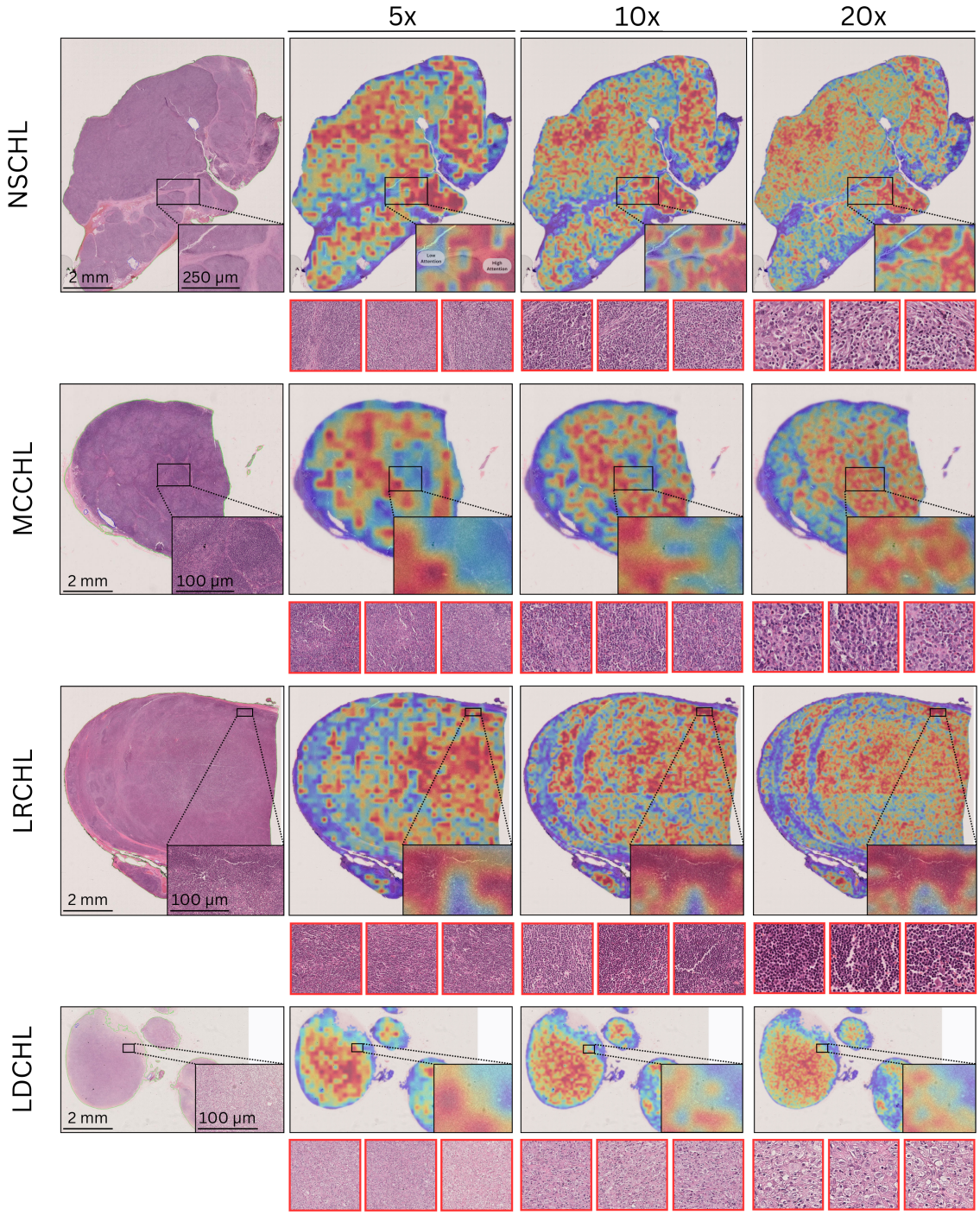

Figure 4.3 Heatmap Visualization of Correctly Classified Subtypes. The figure shows three magnification heatmaps for NSCHL, MCCHL, LRCHL, and LDCHL subtypes. Each heatmap displays top-scoring patches, with black squares indicating ROIs. Red areas highlight high-attention regions critical for classification, while blue areas show low attention. The heatmaps illustrate the model's focus and its alignment with subtype-specific features.

The heatmaps demonstrate that the IEF-MIL model effectively identifies key

morphological features critical for classifying CHL subtypes. This detection aligns well with pathologists' diagnostic patterns, underscoring the model's practical utility. Specifically:

- **NSCHL:** The model consistently focuses on fibrotic bands and nodular areas across all magnifications, with zoomed-in regions of interest (ROIs) clearly showing these features. The highlighted patches, including fibrotic bands and inflammatory areas with HRS cells, support this prediction.
- **LRCHL:** Given the lower frequency of HRS cells, the model emphasizes lymphocyte-rich areas as key predictive features. The most predictive patches reveal inflammatory cells, such as histiocytes, amidst abundant lymphocytes.
- **LDCHL:** The model highlights regions with minimal lymphocytes and HRS cells, focusing on areas devoid of these components. This focus reflects the diagnostic significance of these sparse elements. Zoomed-in ROIs confirm the model's ability to accurately identify high-value diagnostic areas, such as those with numerous HRS cells and fewer lymphocytes.
- **MCCHL:** The model emphasizes a rich inflammatory environment characterized by eosinophils, histiocytes, plasma cells, neutrophils, and HRS cells, particularly at 10x and 20x magnifications. The 20x magnification patches provide a clearer view of these cell types. This dense inflammatory milieu, distinct from other subtypes, demonstrates the model's capability to capture subtle yet significant differences in the tumor microenvironment.

Overall, the IEF-MIL model's heatmaps effectively capture the complexity of CHL subtypes, reinforcing its reliability by mirroring pathologists' analysis and highlighting critical differences in the tumor microenvironment.

5. DISCUSSION

The application of deep learning in the subtyping of CHL represents a transformative advancement in digital pathology. This study is pioneering in its approach, setting itself apart from previous research that either categorized CHL as a single class [13], classified it within the negative class of binary classification of diffuse large B-cell lymphoma [16, 20], or addressed multiclass classification involving only two CHL subtypes [23]. None of these studies focused on the subtyping of CHL subtypes. Our work utilizes the Instance & Embedding Fused Multiple Instance Learning (IEF-MIL) model, which demonstrates a high level of accuracy and robustness in distinguishing between the various subtypes of CHL using H&E-stained WSIs. The IEF-MIL model is distinguished by its innovative approach to dual-stream multiple instance learning, incorporating global attention pooling and mean pooling strategies. Additionally, the model leverages patches across multiple magnifications to effectively mimic the multi-scale analysis performed by pathologists. This method allows the model to consider both the tumor microenvironment and neoplastic cells, which are crucial in accurately subtyping CHL.

One primary challenge this study addresses is the inherent complexity and heterogeneity of Classical Hodgkin Lymphoma subtypes. The CHL spectrum includes the NSCHL, MCCHL, LDCHL, and LRCHL subtypes, each characterized by distinct histopathological features. However, these subtypes can present overlapping characteristics, complicating accurate diagnosis even for experienced pathologists. The model demonstrated remarkable accuracy in classifying these subtypes, showcasing the power of deep learning to reduce diagnostic variability and improve clinical outcomes. Notably, the model excelled with the rare LDCHL subtype, representing only 7.6% of the dataset, outperforming the more prevalent MCCHL (33.8%). This highlights the IEF-MIL model's potential to enhance the detection and diagnosis of rare subtypes. Additionally, the performance of MCCHL can be attributed to its morphological characteristics and its shared inflammatory features with other subtypes, particularly NSCHL

and LDCHL.

Moreover, this study emphasizes the importance of employing multi-scale approaches in histopathological analysis. By leveraging CONCH features at varying magnifications (5x, 10x, and 20x), our model provides a nuanced and comprehensive examination of histological patterns, mirroring the diagnostic approach used in clinical practice. This multi-scale analysis is especially suitable in Classical Hodgkin Lymphoma, where detecting distinctive features – such as fibrotic bands in NSCHL or the characteristic inflammatory milieu in MCCHL – necessitates detailed examination at multiple magnification levels. Such an approach mirrors the traditional diagnostic practice and enhances the model's capability to capture and differentiate subtle histological details critical for accurate diagnosis.

Our results reveal substantial variability in model performance, contingent upon MIL strategies and the choice of embedders. The IEF-MIL model consistently achieves high F1 scores, especially with the CONCH embedder (0.689 ± 0.05). CLAM-MB [28] also performs competitively, reaching a top score of 0.683 ± 0.033 with CONCH, and closely approximates IEF-MIL's results. TRANSMIL [30], utilizing transformers for MIL pooling, exhibits variable performance, likely due to high data requirements of transformer models, resulting in lower scores with some embedders. ACMIL-GA [29], which also uses attention pooling, and DSMIL [25], combining max and attention pooling, do not match the performance of IEF-MIL.

Despite notable advancements, our study highlights several limitations and identifies critical areas for future research. We utilized several histopathology foundational models, pre-trained on extensive and diverse datasets, for feature extraction through transfer learning. However, these foundation models either include a range of lymphoma cancers [55, 1, 53, 50, 57] or exclude lymphoma and lymphatic cancers [56], with none specifically including CHL samples, which are morphologically distinct from other lymphoma types. While these pre-trained models are valuable, their reliance may not fully capture the variability inherent in specific cancer types, particularly those not represented in the pretraining datasets. This limitation underscores the necessity for

further fine-tuning and customizing these models to enhance their robustness and clinical applicability. To improve the performance of our MIL pipeline, obtaining a larger, more diverse dataset for the fine-tuning of these foundational models is essential. As discussed in [1], tailored fine-tuning of these models can significantly enhance CHL subtyping, offering advantages beyond those achieved through transfer learning alone.

In conclusion, this study provides a significant contribution to integrating deep learning with medical diagnostics. Applying the IEF-MIL model for Classical Hodgkin Lymphoma subtyping demonstrates the transformative impact of advanced machine learning techniques in improving diagnostic accuracy and reducing variability in CHL assessments. By demonstrating the model's efficacy in enhancing diagnostic precision, this research underscores the potential of deep learning to revolutionize CHL diagnostics and lays a robust foundation for future advancements in the field. The findings pave the way for further innovations, promising to advance the frontier of computational pathology and precision medicine.

6. CONCLUSION

This study presents a novel MIL pipeline specifically designed for Classical Hodgkin Lymphoma subtyping. Our results demonstrate the potential of leveraging SOTA SSL foundation models in histopathology. While the initial findings are promising, they also highlight the need for further refinement and optimization of these models to achieve superior performance in CHL subtyping. Future research should prioritize expanding the dataset to facilitate more precise fine-tuning of these foundation models. This will increase CHL subtyping accuracy, facilitating the development of more effective diagnostic tools.

Additionally, we introduce a comprehensive dataset comprising 1,247 whole slide images of CHL, including 95 LDCHL, 216 LRCHL, 421 MCCHL, and 515 NSCHL cases, sourced from three medical centers. This dataset contributes significantly to CHL research by addressing the critical need for extensive, annotated datasets encompassing a broad range of CHL subtypes, including rarer forms. This dataset will be a crucial asset, advancing clinical and computational CHL research and ultimately enhancing diagnostic accuracy and treatment outcomes.

REFERENCES

1. Lu, M. Y., B. Chen, D. F. Williamson, R. J. Chen, *et al.*, "A visual-language foundation model for computational pathology," *Nature Medicine*, Vol. 30, pp. 863–874, 2024.

2. Ansell, S. M., "Hodgkin lymphoma: 2023 update on diagnosis, risk-stratification, and management," *American Journal of Hematology*, Vol. 97, no. 11, pp. 1478–1488, 2022.

3. Sung, H., J. Ferlay, R. L. Siegel, M. Laversanne, *et al.*, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, Vol. 71, no. 3, pp. 209–249, 2021.

4. Wang, H. W., J. P. Balakrishna, S. Pittaluga, and E. S. Jaffe, "Diagnosis of hodgkin lymphoma in the modern era," *British Journal of Haematology*, Vol. 184, no. 1, pp. 45–59, 2019.

5. Swerdlow, S. H., E. Campo, S. A. Pileri, N. L. Harris, *et al.*, "The 2016 revision of the World Health Organization classification of lymphoid neoplasms," *Blood*, Vol. 127, no. 20, pp. 2375–2390, 2016.

6. Brice, P., E. de Kerviler, and J. W. Friedberg, "Classical hodgkin lymphoma," *The Lancet*, Vol. 398, no. 10310, pp. 1518–1527, 2021.

7. Takahara, T., A. Satou, T. Tsuzuki, and S. Nakamura, "Hodgkin lymphoma: Biology and differential diagnostic problem," *Diagnostics(Basel)*, Vol. 12, no. 6, p. 1507, 2022.

8. Eberle, F. C., H. Mani, and E. S. Jaffe, "Histopathology of hodgkin's lymphoma," *The Cancer Journal*, Vol. 15, no. 2, pp. 129–137, 2009.

9. Syrykh, C., A. Abreu, N. Amara, A. Siegfried, *et al.*, "Accurate diagnosis of lymphoma on whole-slide histopathology images using deep learning," *NPJ Digital Medicine*, Vol. 3, no. 1, p. 63, 2020.

10. Steinbuss, G., M. Kriegsmann, C. Zgorzelski, A. Brobeil, *et al.*, "Deep learning for the classification of non-hodgkin lymphoma on histopathological images," *Cancers*, Vol. 13, no. 10, p. 2419, 2021.

11. Irshaid, L., J. Bleiberg, E. Weinberger, J. Garritano, *et al.*, "Histopathologic and machine deep learning criteria to predict lymphoma transformation in bone marrow biopsies," *Archives of Pathology & Laboratory Medicine*, Vol. 146, no. 2, pp. 182–193, 2022.

12. Mohlman, J. S., S. D. Leventhal, T. Hansen, J. Kohan, *et al.*, "Improving Augmented Human Intelligence to Distinguish Burkitt Lymphoma From Diffuse Large B-Cell Lymphoma Cases," *American Journal of Clinical Pathology*, Vol. 153, no. 6, pp. 743–759, 2020.

13. Shankar, V., X. Yang, V. Krishna, B. Tan, *et al.*, "Lymphoml: An interpretable artificial intelligence-based method identifies morphologic features that correlate with lymphoma subtype," in *Machine Learning for Health (ML4H)*, pp. 528–558, 2023.

14. Bai, J., H. Jiang, S. Li, X. Ma, *et al.*, "Nhl pathological image classification based on hierarchical local information and googlenet-based representations," *BioMed Research International*, Vol. 2019, 2019.

15. El Achi, H., T. Belousova, L. Chen, A. Wahed, *et al.*, "Automated diagnosis of lymphoma with digital pathology images using deep learning," *Annals of Clinical & Laboratory Science*, Vol. 49, no. 2, pp. 153–160, 2019.

16. Basu, S., R. Agarwal, and V. Srivastava, "Deep discriminative learning model with calibrated attention map for the automated diagnosis of diffuse large b-cell lymphoma," *Biomedical Signal Processing and Control*, Vol. 76, p. 103728, 2022.

17. Khelil, H., A. El Moumene Zerari, and L. Djerou, "Accurate diagnosis of non-hodgkin lymphoma on whole-slide images using deep learning," in *2022 IEEE 9th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, pp. 447–451, IEEE, 2022.

18. Hamdi, M., E. M. Senan, M. E. Jadhav, F. Olayah, *et al.*, "Hybrid models based on fusion features of a cnn and handcrafted features for accurate histopathological image analysis for diagnosing malignant lymphomas," *Diagnostics*, Vol. 13, no. 13, p. 2258, 2023.

19. Miyoshi, H., K. Sato, Y. Kabeya, S. Yonezawa, *et al.*, "Deep learning shows the capability of high-level computer-aided diagnosis in malignant lymphoma," *Laboratory Investigation*, Vol. 100, no. 10, pp. 1300–1310, 2020.

20. Li, D., J. R. Bledsoe, Y. Zeng, W. Liu, *et al.*, "A deep learning diagnostic platform for diffuse large b-cell lymphoma with high accuracy across multiple hospitals," *Nature Communications*, Vol. 11, no. 1, p. 6004, 2020.

21. Brancati, N., G. De Pietro, M. Frucci, and D. Riccio, "A deep learning approach for breast invasive ductal carcinoma detection and lymphoma multi-classification in histological images," *IEEE Access*, Vol. 7, pp. 44709–44720, 2019.

22. Jianfei, Z., C. Wensheng, G. Xiaoyan, W. Bo, *et al.*, "Classification of digital pathological images of non-hodgkin's lymphoma subtypes based on the fusion of transfer learning and principal component analysis," *Medical Physics*, Vol. 47, pp. 4241–4253, 2020.

23. Hashimoto, N., D. Fukushima, R. Koga, Y. Takagi, *et al.*, "Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3851–3860, 2020.

24. Ilse, M., J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136, 2018.

25. Li, B., Y. Li, and K. W. Eliceiri, "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2021.

26. Campanella, G., M. Hanna, L. Geneslaw, A. Miraflor, *et al.*, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Medicine*, Vol. 25, p. 1, 2019.

27. Oner, M. U., J. M. S. Kye Jet, H. K. Lee, and W. K. Sung, "Distribution based mil pooling filters: Experiments on a lymph node metastases dataset," *Medical Image Analysis*, Vol. 87, p. 102813, 2023.

28. Lu, M. Y., D. F. Williamson, T. Y. Chen, R. J. Chen, *et al.*, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering*, Vol. 5, no. 6, pp. 555–570, 2021.

29. Zhang, Y., H. Li, Y. Sun, S. Zheng, *et al.*, "Attention-challenging multiple instance learning for whole slide image classification," *ArXiv*, Vol. abs/2311.07125, 2023.

30. Shao, Z., H. Bian, Y. Chen, Y. Wang, *et al.*, "Transmil: Transformer based correlated multiple instance learning for whole slide image classification," *Advances in Neural Information Processing Systems*, Vol. 34, pp. 2136–2147, 2021.

31. Zhang, R., Q. Zhang, Y. Liu, H. Xin, *et al.*, "Multi-level multiple instance learning with transformer for whole slide image classification," *arXiv preprint*, Vol. 2306.05029, 2023.

32. Xiong, C., H. Chen, J. J. Sung, and I. King, "Diagnose like a pathologist: Transformer-enabled hierarchical attention-guided multiple instance learning for whole slide image classification," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pp. 1587–1595, 2023.

33. Zhang, J., C. Hou, W. Zhu, M. Zhang, *et al.*, "Attention multiple instance learning with transformer aggregation for breast cancer whole slide image classification," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1804–1809, 2022.

34. Gao, C., Q. Sun, W. Zhu, L. Zhang, *et al.*, "Transformer based multiple instance learning for wsi breast cancer classification," *Biomedical Signal Processing and Control*, Vol. 89, p. 105755, 2024.

35. Ren, Q., Y. Zhao, B. He, B. Wu, *et al.*, "Iib-mil: Integrated instance-level and bag-level multiple instances learning with label disambiguation for pathological image analysis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 560–569, Springer, 2023.

36. Tan, L., H. Li, J. Yu, H. Zhou, *et al.*, "Colorectal cancer lymph node metastasis prediction with weakly supervised transformer-based multi-instance learning," *Medical & Biological Engineering & Computing*, Vol. 61, no. 6, pp. 1565–1580, 2023.

37. Gul, A. G., O. Cetin, C. Reich, N. Flinner, *et al.*, "Histopathological image classification based on self-supervised vision transformer and weak labels," in *Medical Imaging 2022: Digital and Computational Pathology*, Vol. 12039, pp. 366–373, SPIE, 2022.

38. Lerousseau, M., M. Vakalopoulou, E. Deutsch, and N. Paragios, "Sparseconvmil: sparse convolutional context-aware multiple instance learning for whole slide image classification," in *MICCAI Workshop on Computational Pathology*, pp. 129–139, PMLR, 2021.

39. Yu, J., T. Ma, Y. Fu, H. Chen, *et al.*, "Local-to-global spatial learning for whole-slide image representation and classification," *Computerized Medical Imaging and Graphics*, Vol. 107, p. 102230, 2023.

40. Zhang, H., Y. Meng, Y. Zhao, Y. Qiao, *et al.*, "Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18802–18812, 2022.

41. Rymarczyk, D., A. Borowa, J. Tabor, and B. Zielinski, "Kernel self-attention for weakly-supervised image classification using deep multiple instance learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1721–1730, 2021.

42. Hashimoto, N., D. Fukushima, R. Koga, Y. Takagi, *et al.*, "Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with non-annotated histopathological images," *CoRR*, Vol. abs/2001.01599, 2020.

43. Riasatian, A., M. Babaie, D. Maleki, S. Kalra, *et al.*, "Fine-Tuning and training of densenet for histopathology image representation using TCGA diagnostic slides," *Medical Image Analysis*, Vol. 70, p. 102032, 2021.

44. Chitnis, S. R., S. Liu, T. Dash, T. T. Verlekar, *et al.*, "Domain-specific pre-training improves confidence in whole slide image classification," in *2023 45th Annual International Conference of the IEEE Engineering in Medicine amp; Biology Society (EMBC)*, pp. 1–4, IEEE, 2023.

45. Chen, T., S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*, pp. 1597–1607, PMLR, 2020.

46. Liu, Z., Y. Lin, Y. Cao, H. Hu, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

47. Cai, H., X. Feng, R. Yin, Y. Zhao, *et al.*, "Mist: multiple instance learning network based on swin transformer for whole slide image classification of colorectal adenomas," *The Journal of Pathology*, Vol. 259, no. 2, pp. 125–135, 2023.

48. Ciga, O., T. Xu, and A. L. Martel, "Self supervised contrastive learning for digital histopathology," *Machine Learning with Applications*, Vol. 7, p. 100198, 2022.

49. Azizi, S., L. A. Culp, J. Freyberg, B. Mustafa, *et al.*, "Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging," *Nature Biomedical Engineering*, Vol. 7, no. 6, pp. 756–779, 2023.

50. Wang, X., S. Yang, J. Zhang, M. Wang, *et al.*, "Transformer-based unsupervised contrastive learning for histopathological image classification," *Medical Image Analysis*, Vol. 81, p. 102559, 2022.

51. Huang, Z., F. Bianchi, M. Yuksekgonul, T. J. Montine, *et al.*, "A visual–language foundation model for pathology image analysis using medical twitter," *Nature Medicine*, Vol. 29, no. 9, pp. 2307–2316, 2023.

52. Wu, Y., S. Li, Z. Du, and W. Zhu, "Brow: Better features for whole slide image based on self-distillation," *CoRR*, Vol. abs/2309.08259, 2023.

53. Filiot, A., R. Ghermi, A. Olivier, P. Jacob, *et al.*, "Scaling self-supervised learning for histopathology with masked image modeling," *medRxiv*, 2023.

54. Zhou, J., C. Wei, H. Wang, W. Shen, *et al.*, "ibot: Image bert pre-training with online tokenizer," *CoRR*, Vol. abs/2111.07832, 2021.

55. Vorontsov, E., A. Bozkurt, A. Casson, G. Shaikovski, *et al.*, "A foundation model for clinical-grade computational pathology and rare cancers detection," *Nature Medicine*, pp. 1–12, 2024.

56. Xu, H., N. Usuyama, J. Bagga, S. Zhang, *et al.*, "A whole-slide foundation model for digital pathology from real-world data," *Nature*, Vol. 630, pp. 181–188, 2024.

57. Chen, R. J., T. Ding, M. Y. Lu, D. F. Williamson, *et al.*, "Towards a general-purpose foundation model for computational pathology," *Nature Medicine*, Vol. 30, pp. 850–862, 2024.

58. Yu, J., Z. Wang, V. Vasudevan, L. Yeung, *et al.*, "Coca: Contrastive captioners are image-text foundation models," *arXiv preprint*, Vol. abs/2205.01917, 2022.

59. Chen, R. J., C. Chen, Y. Li, T. Y. Chen, *et al.*, "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16144–16155, 2022.

60. Wu, W., Z. Zhu, B. Magnier, and L. Wang, "Clustering-based multi-instance learning network for whole slide image classification," in *International Workshop on Computational Mathematics Modeling in Cancer Analysis*, pp. 100–109, Springer, 2022.

61. Qu, L., X. Luo, S. Liu, M. Wang, *et al.*, "Dgmil: Distribution guided multiple instance learning for whole slide image classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 24–34, Springer, 2022.

62. Chen, Y., Z. Shao, H. Bian, Z. Fang, *et al.*, "dmil-transformer: Multiple instance learning via integrating morphological and spatial information for lymph node metastasis classification," *IEEE Journal of Biomedical and Health Informatics*, Vol. 27, no. 9, pp. 4433–4443, 2023.

63. Keshvarikhojasteh, H., J. P. Pluim, and M. Veta, "Multi-head attention-based deep multiple instance learning," in *MICCAI Workshop on Computational Pathology with Multimodal Data (COMPAYL)*, 2024.

64. Tang, W., F. Zhou, S. Huang, X. Zhu, *et al.*, "Feature re-embedding: Towards foundation model-level performance in computational pathology," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11343–11352, 2024.

65. Zhao, Y., Z. Lin, K. Sun, Y. Zhang, *et al.*, "Setmil: spatial encoding transformer-based multiple instance learning for pathological image analysis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 66–76, Springer, 2022.

66. Cheson, B. D., R. I. Fisher, S. F. Barrington, F. Cavalli, *et al.*, "Recommendations for initial evaluation, staging, and response assessment of hodgkin and non-hodgkin lymphoma: the lugano classification," *Journal of Clinical Oncology*, Vol. 32, no. 27, pp. 3059–3068, 2014.