College of Computer and Information Sciences
Computer Science Department

"Pathology Image Classification Using AI"

CSC 496– Final Report

Prepared by:

Saja Alzahrani	444200512
Shatha Alabbad	444200515
Hessa Aldekhiel	444200593
Maram Alsheddi	444200628
Sadeem Almutairi	444200865

Supervised by:

Dr. Sultanah Alotaibi

Research project for the degree of bachelor's in computer science

First Semester 1447

I. English Abstract

Digital pathology (DP) has rapidly advanced to transform traditional pathology methods into computational pathology (CP), with the introduction of Whole-slide images (WSI), which is the process of scanning glass slides to produce high-resolution digital images of entire pathology slides, allowing pathologists to view and analyze microscopic images digitally. By incorporating Artificial Intelligence (AI) and Deep Learning (DL), CP has not only helped automate parts of the diagnostic process but also shown the potential to improve efficiency, consistency, and scalability in clinical practice. However, there are still major limitations. The large size of pathology datasets requires a lot of computational power to avoid the possibility of slowing down experimentation and development. Another major limitation is the lack of performance generalization when models are applied to data beyond their training. This is often caused by the lack of diversity in available datasets and the technical variations in image acquisition, which can cause biases that reduce the reliability and robustness of classification models. This project focuses on implementing a DL model, selected based on the literature review, to automate the classification of a publicly available WSI dataset. By focusing on improvements in data preparation, model training, and evaluation, this project aims to provide a working classification pipeline, provide empirical analysis, benchmark models with a public dataset, therefore contributing a practical solution for CP.

II. Arabic Abstract

تقدم علم الأمراض الرقمي بسرعة كبيرة ليحول الأساليب التقليدية في علم الأمراض إلى علم أمراض حاسوبي، وذلك من خلال إدخال تقنية الشرائح الكاملة (WSI)، وهي عملية مسح الشرائح الزجاجية لإنتاج صور رقمية عالية الدقة لكامل الشريحة المرضية، مما يمكن أخصائيين علم الأمراض من عرض الصور المجهرية وتحليلها بشكل رقمي. ومن خلال دمج الذكاء الاصطناعي (AI) والتعلم العميق (DL)، لم يقتصر دور علم الأمراض الحاسوبي على أتمتة أجزاء من عملية التشخيص فحسب، بل أظهر أيضًا إمكانية تحسين الكفاءة والاتساق وقابلية التوسع في الممارسة السريرية. ومع ذلك، لا تزال هناك قيود رئيسية. فالحجم الكبير لبيانات علم الأمراض يتطلب قدرة حسابية عالية لتجنب بطء عمليات التجريب والتطوير. كما تتمثل إحدى القيود المهمة الأخرى في ضعف قدرة النماذج على التعميم عند تطبيقها على بيانات تختلف عن تلك التي تم تدريبها عليها. وغالباً ما ينتج ذلك عن نقص التنوع في مجموعات البيانات المتاحة والاختلافات التقنية في طرق التقاط الصور، مما قد يؤدي إلى تحيزات تقلل من موثوقية النماذج وقوتها في التصنيف. يركز هذا المشروع على أتمتة تصنيف صور شرائح علم الأمراض باستخدام تقنيات التعلم العميق، وذلك من خلال استخدام ومعالجة مجموعة بيانات عامة من الشرائح الكاملة (WSI)، ثم تنفيذ النماذج العميقة وتدريبها وتقييم أدائها. النتيجة المتوقعة هي تسمية تصنيف لكل شريحة مثل "ورم" أو "طبيعي"، حيث سيتم تحديد الفئات الدقيقة بناءً على مجموعة البيانات المختارة. ومن خلال التركيز على تحسين إعداد البيانات، وتدريب النماذج، وعمليات التقييم، يهدف هذا المشروع إلى تطوير نظام تصنيف عملي ومتكامل، وتقديم تحليل تجريبي ومقارنة بين النماذج ومجموعة البيانات المتاحة، مما يساهم في تقديم حل فعال وواقعي في مجال علم الأمراض الحاسوبي.

Table of Contents

List of Tables

List of Figures

List of Notations

AI	Artificial intelligence
AUC	Area under the curve
BCE	Binary Cross Entropy
CNNs	Convolutional neural networks
CP	Computational pathology
CTE	CNN-based tile encoder
DP	Digital pathology
DL	Deep Learning
HOG	Histogram of Oriented Gradients
H&E	Hematoxylin and Eosin
LIME	Local interpretable model-agnostic explanations
MIL	Traditional Multiple Instance Learning
ML	Machine learning
PTO	Pyramid Tiling with Overlap
RF	Random Forest
ROI	Region of interest
SHAP	Shapley additive explanations
TFA	Transformer-based feature aggregator
WSIs	Whole Slide Images
XAI	Explainable Artificial Intelligence

Chapter 1: Introduction

In recent years, the way pathologists analyze tissues has changed dramatically with the rise of digital tools and AI. Instead of relying only on microscopes, tissue slides can now be scanned into Whole Slide Images (WSIs), which allow computers to detect complex visual patterns and assist in diagnosing diseases. This digital shift has made it possible to apply AI and DL in pathology, offering faster and more consistent diagnostic support compared to manual examination. However, manual analysis of pathology slides remains time-consuming and prone to human variability. Developing AI-driven classification models can help overcome these limitations by improving diagnostic consistency, reducing workload, and enabling faster clinical decision-making [1].

DL models, especially those based on Convolutional Neural Networks (CNNs), have shown strong potential in pathology image classification due to their strong feature extraction and pattern recognition capabilities, often outperforming traditional approaches [2], [3]. More recently, Transformer-based architectures have been used to analyze large-scale WSIs with improved spatial understanding, achieving promising results in classification and staging [4]. These developments demonstrate that AI can play an important role in assisting pathologists and reducing diagnostic workload.

However, even with these advances, several limitations still exist. One major issue is variability in staining and image quality between slides, scanners, or medical centers, which makes it difficult for models to generalize well to new data [5], [6]. Another concern is the “black-box” problem, as many DL models can make accurate predictions but lack interpretability, which limits their acceptance by clinicians [7]. In addition, WSIs are extremely large, often exceeding gigapixel size, which creates computational challenges during training and inference [8].

Building upon recent developments in DP and AI, this research aims to develop a DL-based pathology image classification approach for WSIs using a publicly available dataset to classify normal and tumor tissues.

1.1 Problem Statement

In the medical field, pathology forms an important basis for accurate diagnosis and effective treatment, especially for life-threatening diseases like cancer. Throughout the years, pathologists have been analyzing pathology slides manually, and even with the introduction of WSIs, this process is often slow, inconsistent and requires precision to analyze manually, along with the high chance of misdiagnosis due to human error. These challenges have motivated the integration of AI and DL into the diagnostic workflow to improve accuracy, reduce workload, and provide more consistent results.

Although AI and DL have shown great advances, current adaptations in pathology still face major limitations. Pathology dataset are often very large, which can make storage, preprocessing, and training computationally demanding, slowing down the experimentation and development [9]. Additionally, a major challenge is the lack of performance generalization, where a model trained to perform well in one dataset struggles when applied to slides from other institutions due to variations in image acquisition, such as different staining and scanning parameters used across laboratories and a lack of dataset diversity [10]. These factors cause the models to be biased, which in turn reduces the reliability and robustness of classification models [9], [11].

The aim of this project is to develop and evaluate a DL approach to automatically classify pathology slides from a publicly available WSI dataset into categories “Tumor” or “Normal”. To achieve that, necessary preprocessing methods will be applied to the dataset, then the training, followed by the evaluation of the selected DL models. The expected output will be a classification label for each slide. The contribution of this project includes the development of a working classification pipeline along with a comparative benchmarking of multiple models.

1.2 Goals and Objectives

The goal of the project is to develop a DL-based classification pipeline for pathology image classification, with the aim of improving diagnostic precision, reducing pathologists’ workload, and supporting medical research and education.

To achieve the goal, the overall objectives are as follows:

1. Conduct a literature review of current research related to AI-based pathology image classification to identify methods, gaps, challenges in the field, including a review of available datasets and models used in pathology image classification.
2. Examine existing DL models for pathology image classification to identify their strengths, weaknesses, and select the most suitable models for our problem.
3. Select and prepare a publicly available WSI dataset, applying various preprocessing methods to address variability.
4. Design a formal experiment for classifying pathology images, defining the evaluation and training approaches, and performance metrics, such as accuracy, precision, recall, and F1-score to evaluate and compare results across models.
5. Implement the selected models within the designed classification pipeline.
6. Conduct a comparative analysis of the implemented models by training and evaluating them and applying preprocessing strategies, benchmarking their performance against each other and existing studies to identify effective combinations.

1.3 Proposed Solution

For this research problem, we suggest developing a DL-based pathology image classification approach using publicly available dataset. The proposed pipeline includes the main stages of data preprocessing, model training, and evaluation to classify each slide as “Tumor” or “Normal.” This approach is suitable for the problem because DL can automatically extract visual features from pathological images and learn complex patterns that support accurate classification. As part of our contribution, we plan to benchmark different DL models to compare their performance and improve the overall consistency of classification results.

1.4 Research Scope

In our research, we focus on classifying pathology images using DL. We aim to classify WSIs into labels such as Tumor and Normal, using publicly available dataset. The evaluation will be performed using DL models on standard performance metrics.

The research aims to address the following questions:

- How effectively can lightweight models with different feature extractors classify WSIs as Tumor or Normal?

To address this question, this research relies on a publicly available dataset and focuses on empirical analysis.

1.5 Research Significance

The significance of this research is represented by designing a DL-based classification pipeline for pathology images using a publicly available WSIs dataset, reducing the dependence on manual examination of tissue slides by pathologists, which is often time-consuming and prone to human error. The research will have several benefits. First, provides a framework capable of classifying slides into categories (Tumor vs. Normal). Furthermore, this work contributes to advancing DP and supports the integration of AI to assist pathologists and improve the efficiency of image classification in healthcare.

1.6 Ethical and Social Implications

The development of AI applications in DP involves several ethical, social, and privacy considerations that must be addressed throughout the research process. Since our research focuses on classifying WSIs from publicly available dataset and pretrained or standard models, most ethical risks already controlled, it is essential to ensure that all data used are anonymized and collected in compliance with ethical research standards. The dataset used for this research do not contain any personally identifiable information, which minimizes privacy risks and ensures responsible data usage.

From a social perspective, this research aims to support medical professionals by improving diagnostic efficiency and consistency. By assisting pathologists in classifying tissue samples more effectively, AI-based models can help reduce workload and enhance accessibility to healthcare services, particularly in under-resourced or high-demand medical environments. However, it is important to emphasize that the use of AI in pathology is intended to complement and not replace the expertise of human specialists.

Overall, integrating AI in medical workflows requires fairness and accountability. using balanced datasets and transparent metrics to minimize bias and ensure reliable results, while maintaining data confidentiality and following ethical research standards.

The rest of this document is organized into five chapters. Chapter 2 presents the Background including an overview of pathology and the diagnostic process, it also reviews publicly available datasets, preprocessing methods, the technical foundation of WSI analysis, and performance Measures. Chapter 3 reviews related works on pathology image classification divided into seven parts: ML, DL, CNN-Based Approaches, Transformer and Attention-Based Architectures, Large-Scale and Self-Supervised Learning Frameworks, Foundation Models, and Hybrid Methodologies. Chapter 4 presents the methodology that will be implemented in this research, including the overall WSI classification pipeline, the dataset, the preprocessing steps, and the feature extraction backbones. Chapter 5 introduces experimental design, detailing the dataset split, the experimental conditions, the training configuration, and the evaluation protocol that will be used to assess and compare the performance of the proposed models. Chapter 6 concludes the document and provides References.

Chapter 2: Background

This chapter outlines the fundamental concepts and key terminologies related to pathology image classification using AI. It provides the necessary background on DP, DL techniques, and datasets to establish a clear understanding of the research context.

2.1 Introduction to Pathology and Digital Pathology

Pathology is centered on the accurate analysis of microscopic features within tissue samples to provide a diagnosis, which is essential for guiding patient prognosis and therapeutic planning. The diagnosis critically relies on the pathologist's expert interpretation of microscopic tissue features [12].

The field has undergone a paradigm shift towards DP. This approach is centered on WSI, also known as virtual microscopy, which is the process of scanning a conventional glass slide to generate a single, high-resolution digital image [13]. These digital versions can be viewed and analyzed on a computer, offering a way to replicate the experience of traditional light microscopy and enabling CP [9].

The digitization of pathology slides produces several advantages. It enhances collaboration and accessibility, allowing digital slides to be easily shared and viewed remotely for telepathology, expert consultations, and educational purposes, thereby improving workflow efficiency [14]. Furthermore, it eliminates the risks associated with the physical degradation, breakage, or loss of glass slides [13], [14] . Arguably, the most important advantage is that WSI provides the foundation for CP, enabling the development and application of AI-driven image analysis tools [9], [15]. These tools can provide objective, reproducible, and precise quantitative measurements, which help augment the pathologist's diagnosis and improve efficiency [15], [16].

2.2 Challenges in Pathology Image Analysis

Despite the importance of traditional pathology, its reliance on human expertise introduces several inherent challenges that can affect diagnostic consistency and efficiency. The most prominent challenge is the inherent subjectivity of visual interpretation. This subjectivity often results in significant inter-observer variability, where multiple expert pathologists may assign different interpretations to the same tissue sample [15]. These discrepancies are particularly prevalent in complex tasks, such as tumor classification, and can potentially impact crucial patient management decisions. Furthermore, the high workload makes the labor-intensive process of manual diagnosis prone to pathologist fatigue, which increases the risk of diagnostic delays and errors. Additionally, the manual assessment by pathologists is limited in providing precise quantitative data, as visual estimates often lack the accuracy and reproducibility required to support modern diagnostics [15], [16].

2.3 Emergence of AI and DL in Medical Image Analysis

Recently, AI and DL have revolutionized medical image analysis, transforming traditional manual pathology into computational approach. With introduction of DP and WSI, tissue slides can now be digitized for automated analysis, improving diagnostic speed and consistency. CNN has shown strong performance in detecting and classifying tissues, outperforming traditional image processing techniques [28]. XAI methods also enhance the transparency of these models. Allowing clinicians to better understand AI decisions and increase trust in clinical setting [20]. More recently, hybrid and multimodal have extended AI capabilities to integrate both image and textual data for diagnostic reasoning, which is a significant step forward in CP [30].

2.4 DL in Computational Pathology

CP integrates DP, AI, and DL to extract meaningful diagnostic and predictive information from pathology images. It focuses on developing algorithms capable of automatically detecting, classifying, and interpreting tissue patterns to assist disease

diagnosis and outcome prediction. Through the extraction and analysis of quantitative image features from histopathology slides, CP enables the identification of potential biomarkers and helps link visual tissue characteristics with clinical outcomes.

Through DL, CP has achieved remarkable improvements in key tasks such as classification, segmentation, and feature analysis, often reaching human-level performance. In addition to improving diagnostic accuracy, these techniques contribute to standardizing pathology evaluations, reducing observer variability, and enhancing the reliability of clinical decision-making [10].

The direct use of WSIs for computational analysis presents its own set of challenges. Their massive size, often reaching gigapixel resolutions, makes processing them computationally expensive [11]. Furthermore, they can suffer from various artifacts, such as tissue folds and blurriness, and often show significant variations in color and staining intensity due to different lab protocols [17]. These issues can negatively impact the performance of analytical models, making preprocessing an essential step.

Before training, WSI often needed number of preprocessing steps, including tissue detection, artifact removal, and color normalization to help model focus on the most tissue regions by removing background or noise. To achieve this, studies using Techniques on TCGA, comparing traditional image processing, such as color thresholding and morphological operations, with hybrid approaches that combine DL and classical algorithms to efficiently identify tissue areas, showing the importance of preprocessing in optimizing computational resources and model performance in pathology image classification [18].

2.5 Datasets

To develop robust DL models in CP, there needs to be large, publicly available datasets that are well annotated. For any new model, these datasets are used for training and validation, meaning that the models are highly dependent on the quality of the datasets.

One of the most well-known is The Cancer Genome Atlas (TCGA), a large-scale public cancer database that includes genomic and histopathological WSIs across more than 30 cancer types. Subsets like TCGA-BRCA (Breast Invasive Carcinoma) and TCGA-LUSC (Lung Squamous Cell Carcinoma) are frequently used. They consist of thousands of H&E-stained WSIs scanned at $20\times$ and $40\times$ magnification, with resolutions up to 100,000 × 100,000 pixels [19], [20]. These large, high-quality datasets make TCGA a great resource for cancer classification and tumor detection tasks.

While TCGA provides a large range of cancer types, several other public datasets are used for specific diagnostic tasks. For example, Camelyon17 focuses on detecting breast cancer metastasis in lymph node WSIs [8], while LC25000 contains lung and colon tissue patches [21]. Other specialized datasets include ISIC and HAM10000, which provide thousands of labeled images of skin lesions, and OCT2017, which contains Optical Coherence Tomography (OCT) retinal images [1]. Furthermore, KVASIR and Hyper Kvasir are large, open-access datasets of labeled endoscopic images for disease detection [4], [22].

In addition to these public datasets, Some studies rely on private or custom collected datasets that consist of (H&E) stained histological images, such as those from Alberta Precision Labs, Leeds University, and the Cancer Imaging Archive (TCIA) [2]. They are usually gathered by researchers from medical institutions for a specific goal that the public data does not provide.

Finally, some datasets are provided by data science platforms like Kaggle and Figshare [1] and offer large collections of medical imaging datasets of various sizes making them a valuable source for many researchers in this field.

The table below provides an overview of a selection of prominent datasets, highlighting their scale, tissue type, and primary tasks.

Table 1: Overview of a Selection of Datasets Used in Pathology

Data Modality	Dataset	Number of Samples	Size (pixels)	Tissue/Organ	Classification Labels or Task	Reference
WSIs	**TCGA-BRCA**	3,112	100,000 * 100,000	Breast	Invasive Carcinoma (Tumor vs. Normal)	**[19]**
	TCGA-LUSC	1,612	100,000 * 100,000	Lung	Squamous Cell Carcinoma (Tumor vs. Normal)	**[20]**
	Camelyon17	500	100,000 * 100,000	Breast (Lymph Node)	Metastasis Detection	**[8]**
	ISIC	>25,000	various sizes	Skin	Skin Lesion Classification	**[1]**
Patches	**LC25000**	25,000	768 * 768	Lung & Colon	5 Classes (e.g., Tumor, Normal)	**[21]**
Endoscopic Images	**KVASIR**	1000	Up to 1920 * 1072	Gastrointestinal	Endoscopic Disease Detection	**[22]**
	Hyper Kvasir	10,662	Up to 1920 * 1072	Gastrointestinal	Endoscopic Disease Detection	**[4]**

As shown by Table 1, there is a variety in datasets in terms of data modality, scope, scale and overall task. In this project, we will utilize datasets with many high-resolution WSIs and clear "Tumor" vs. "Normal" labels, making them ideal for the development and benchmarking of our classification pipeline.

2.6 Technical Foundations for WSI Analysis

WSI analysis primarily relies on DL techniques, particularly CNNs, which are DL architectures composed of convolution and pooling layers followed by fully connected layers, designed to automatically learn visual patterns from images [23]. CNN architectures such as ResNet18, EfficientNetV2, InceptionV3, and VGG16 are widely used for WSI classification, as they effectively capture both fine-grained and high-level tissue features essential for diagnostic accuracy [2], [22], [23].

The Vision Transformer (ViT) applies a pure transformer directly to images by splitting an image into fixed-size patches and processing these patches as a sequence of patch embeddings using self-attention [24]. Recent developments have introduced Transformer-based architectures such as Wave-ViT [4], which enhance the ability to model long-range spatial relationships between image regions, improving classification accuracy for complex WSIs. Hybrid models like MobileNetV2 [7] combining CNN and transformer components, have shown improved performance by integrating both local and global context understanding. Furthermore, XAI techniques such as Grad-CAM and Saliency Maps are often used to highlight the regions influencing model decisions, improving interpretability and clinical trust. In addition, foundation and self-supervised learning models have emerged as promising approaches, enabling feature learning from large unlabeled datasets and improving model generalization across diverse pathology domains.

Alongside these architectures, weakly supervised learning methods such as Attention-based Multiple Instance Learning (AMIL) have become widely used in WSI classification, as they operate on slide-level labels without requiring patch-level annotations. AMIL provides an effective framework for aggregating patch embeddings and identifying the most informative regions within a slide, making it suitable for large-scale WSI analysis [25].

These technical foundations together support the development of a DL based pipeline for WSI classification.

2.7 Performance Measures

To evaluate the performance of classification models in WSI analysis, several standard measures are commonly used. These metrics help assess how accurately and effectively a model can identify different classes within pathology. These measures shown in Table 2 will later be used to assess and compare the performance of the developed DL models in this research.

Table 2: Performance Measures

Metrics	Description
(2.1) Accuracy (%)	Measures the percentage of correct predictions among all samples
(2.2) Precision (%)	Measures of how many of the predicted positive are truly positive
(2.3) Recall (Sensitivity) (%)	Measures of how many of the predicted positive cases were correctly identified
(2.4) F1-score	Balances precision and recall for imbalanced datasets
(2.5) Specificity (%)	Measures how well the model identifies negative cases
AUC	Indicates the model's ability to distinguish between classes

These measures listed in Table 2 focus mostly on the classifier's capability to predict correctly and they are built from the confusion matrix, which contains correctly and incorrectly predicted examples.

- True Positive (TP): The case in which the classifier predicted yes, and it is actually yes
- True Negative (TN): The case in which the classifier predicted no, and it is actually no.
- False Positive (FP): The case in which the classifier predicted yes, and it is actually no.
- False Negative (FN): The case in which the classifier predicted no, and it is actually yes.

The corresponding evaluation metrics are computed with these equations as follow:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

$$Precision = \frac{TP}{TP + FN} \tag{2.2}$$

$$Recall\ (Sensitivity) = \frac{TP}{TP + FN} \tag{2.3}$$

$$F1 - score = 2\ \times \frac{presicion\ \times recall}{presicion\ + recall} \tag{2.4}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2.5}$$

2.8 Current Limitations and Open Challenges

Despite the remarkable progress of AI in pathology, several limitations and open challenges remain. One major challenge is generalizability, as models trained on data from one institution often show reduced performance when applied to data from other sources. This issue arises from variations in sample preparation, staining methods, and scanning devices [5], [12]. In addition, effective training requires large amounts of accurately annotated data, which is both costly and time-consuming for experts. The massive size of WSIs also poses a major computational challenge, highlighting the need for more efficient models and architectures capable of processing such large data effectively [8], [25], [26]. Another key barrier to clinical adoption is the "black box" nature of many DL models. This lack of transparency reduces clinicians' trust and emphasizes the importance of developing Explainable Artificial Intelligence (XAI) methods to make model decisions more interpretable [7], [22], [27]. Finally, future research should focus on integrating pathological image data with other multimodal data, such as genomic [10], [28] and radiological data [1], to achieve a more comprehensive understanding of the patient's condition.

2.9 Summary

In Summary this chapter provided an overview of the key concepts, challenges, and methods relevant to digital and CP. It explained how the shift from traditional pathology to DP, supported by WSI, enabled the application of AI and DL for automated tissue classification. The chapter also discussed common preprocessing techniques, datasets and major DL architectures. Despite significant progress, challenges such as generalizability, computational complexity, and model interpretability remain open research areas.

The following chapter builds upon this foundation by reviewing existing research on AI-based pathology image classification, highlighting different approaches, models, and limitations that guide the design of our proposed pipeline.

Chapter 3: Literature Review

The rise of AI in DP has introduced a new era of precision and automation in pathology image classification, offering both remarkable advancements and notable challenges. AI-driven models have improved diagnostic accuracy and reduced human workload, transforming traditional pathology workflows. However, these systems often function as 'black boxes', raising concerns about transparency, model interpretability, and clinical reliability, emphasizing the need for explainable and generalizable AI solutions in healthcare.

3.1 Overview of AI in Pathology

AI has become an essential tool in DP, transforming how tissue slides are analyzed and interpreted. As discussed in Chapter 2, DL techniques such as CNNs have shown strong performance in classifying pathology images into categories such as normal or tumor. These models help pathologists make faster and more consistent diagnoses while reducing manual workload. However, challenges such as variations in staining, large image sizes, and limited dataset diversity still affect model generalization. Building on these developments, this chapter explores previous research related to pathology image classification using AI and identifies the most effective approaches for improving diagnostic accuracy and efficiency.

3.2 ML Approaches

ML is a branch of AI that learns from data to make predictions. ML techniques have been successfully used by several pathology image classification studies for accurate tumor diagnosis.

Early studies on ML represented the intial stage of automation in DP, A recent study [29] evaluated AutoML algorithms for classifying colon polyps and showed their applicability in DP. Using Google's Vertex AI, the model was tested on the Alberta Precision Laboratories teaching slide set, with publicly accessible WSIs from Leeds

University, and the Cancer Imaging Archive to increase dataset variability, achieving recall and precision approaching 100%, highlighting the potential of ML to transform diagnostic pathology. Limitations included the small sample size, limited dataset diversity, and the misclassification of normal tissue as hyperplastic polyps, suggesting the need for more normal images with different-shaped lumens would help with proper identification.

Another study [8] explored feature aggregation to enhance the accuracy and computational efficiency of WSI classification. The study addressed the limitations of conventional Multiple Instance Learning (MIL) approaches, which require analysing thousands of patches per slide, making the process slow and computationally demanding. The study mainly relied on a ML–based feature aggregation approach, where DL models ResNet-50 and MobileNetV3-small were only used for feature extraction before applying the Fisher Vector encoding for slide-level classification. This method focused on selecting diagnostically significant patches and combining their extracted features into a single, concise representation that summarizes the entire slide. The datasets included Camelyon17 for breast cancer metastasis detection and TCGA Lung for EGFR mutation prediction, achieving 80% and 85% accuracy with ResNet-50, while MobileNetV3-small showed lower results with 74% and 80% but reduced computational cost. Despite the outcomes, the technical complexity of Fisher Vector implementation was limiting the study, including the dependency on accurate patch selection and the lack of interpretability compared to attention-based model limitations. Overall, the study indicated that feature aggregation methods can achieve good accuracy while reducing computational cost, making them more practical for large-scale WSI analysis.

3.3 Deep Learning (CNN-Based)

The ML approaches described previously rely on a multi-step process where feature extraction and classification are separate tasks, which is a complex and less optimal approach. This guided the shift towards end-to-end DL models that combine the feature learning and classification into a single step for a more optimal system.

CNNs are the most used DL approach in CP. Their ability to extract features automatically and analyze complex patterns allows them to be easily used in analyzing pathological images.

For simple CNN approaches according to [21], an Explainable and lightweight lung Cancer Net (XLLC-Net) a CNN meticulously crafted for the classification of lung cancer from histopathological images, as shown in Figure 1, using LC25000 dataset. The model consists of four Convolutional layers and contains approximately 3 million parameters, with integration of Explainable AI techniques (XAI) such as Saliency Maps and GRAD-Cam to enhance interpretability. Furthermore, pretrained models (AlexNet, ResNet50, VGG16, and VGG19) were used in comparison with the proposed model which demonstrates the highest overall performance (0.99 to 1.00) across all cancer types and metrics. Despite its promising performance, the study's evaluation was limited to the LC25000 dataset, which may restrict its generalizability to more diverse pathological data.

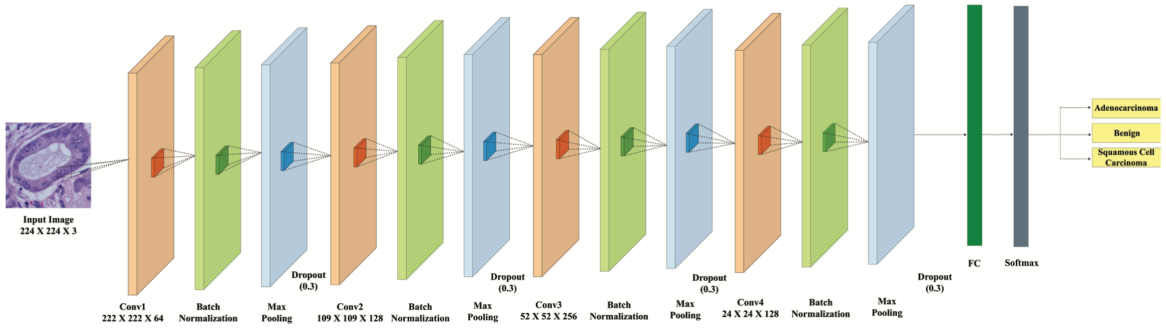


Figure 1. Architecture of XLLC-Net model [21]

Similarly, Doğan and Yılmaz [3] compared traditional manual analysis with an AI automation approach for the classification of colorectal cancer histopathology images. The study developed two models: a manual analysis pipeline using Histogram of Oriented Gradients (HOG) for feature extraction with a Random Forest (RF) classifier, and an AI-based approach using a custom-built Convolutional Neural Network (CNN). These models were evaluated on open-source datasets containing H&E-stained images sorted into nine tissue classes. The results demonstrated the superior performance of the

AI model, which achieved 91% accuracy in binary (normal vs. tumor) classification compared to the manual method's 75%, and for the more complex nine-class task, the AI model reached 97% accuracy, far surpassing the manual model's 44% [3]. Despite these results, both studies showed that the models depend heavily on the dataset size and quality, potential class imbalance, and the compact architecture may restrict the ability to capture highly complex features.

A more enhanced approach with multi-stage pipeline was conducted in [30], where a two-stage AI pipeline was developed for the classification of lung adenocarcinoma WSIs, aiming to distinguish tumor from normal tissue and to predict epidermal growth factor receptor (EGFR) mutational status, as shown in Figure 2. The study employed two CNNs based (CNN) on the InceptionResNet-V2 architecture, pretrained on ImageNet and fine-tuned on histopathology data, with interpretability enhanced through High-Dimensional Grad-CAM visualizations. The first model is a Tumor Classifier CNN fine-tuned to discriminate healthy from cancerous tissue, where the second model is the EGFR Mutation Classifier another CNN that receives only the tiles identified as cancerous by the first model and predicts EGFR-mutated versus wild-type status. its reliance on limited datasets from three medical centers and binary classification tasks. The findings demonstrated that the two-stage model can reliably identify tumor regions and moderately predict EGFR mutations, supporting its potential role in clinical workflows. The study acknowledged limitations including the relatively small dataset size, weak supervision at the tile level, and the restriction to binary mutation classification, which may limit broader clinical applicability.

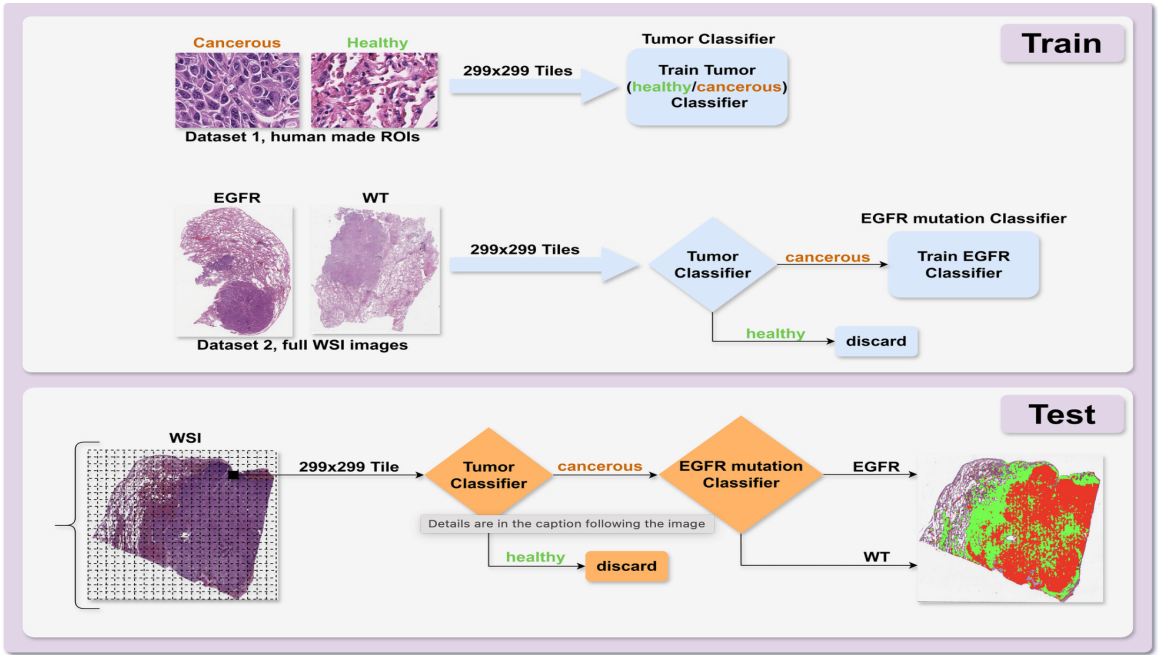


Figure 2. Architecture of InceptionResNet-V2 model [30]

For more complex CNN approaches, a DL framework ResMAPNet proposed in [31] integrating multidimensional attention (CLIA) and pinwheel convolution (PSAConv) to improve robustness, accuracy, generalization, feature extraction of brain tumor classification, as well as efficiency compared to state-of-the-art models. The model was evaluated using the Figshare and the Kaggle publicly available brain tumor datasets. The model achieved 99.51% accuracy in three-class classification and 98.01% accuracy in four-class classification, improving the baseline by 4.41–4.45% and showing strong generalizability with better efficiency. However, despite these promising results, the study's validation was limited to publicly available datasets and specific brain tumor types, which restricts its clinical applicability. This highlights a gap in developing more adaptable and generalizable frameworks capable of maintaining high accuracy across diverse medical imaging conditions.

Similarly, DL-based framework for AI-assisted medical imaging introduced in [1], incorporating the Adaptive Multi-Resolution Imaging Network (AMRI-Net) to enhance diagnostic accuracy across various imaging techniques, and the Explainable Domain-Adaptive Learning (EDAL) strategy to improve domain generalizability and clinical trust in AI-driven diagnoses. The framework was trained on multi-modal datasets including ISIC, HAM10000, OCT2017 and the Brain MRI dataset, and evaluated on the NIH ChestX-ray14 and CAMELYON16 datasets. The results showed classification accuracies up to 94.95% and F1-Scores of 94.85%, which highlights improvement in transparency and performance, as shown in Figure 3. However, despite its strong generalizability across modalities, the study primarily relied on benchmark datasets rather than clinical real-world data, limiting its assessment of true domain adaptation under heterogeneous imaging conditions—an area that remains an open research gap for ensuring consistent reliability in real diagnostic workflows.

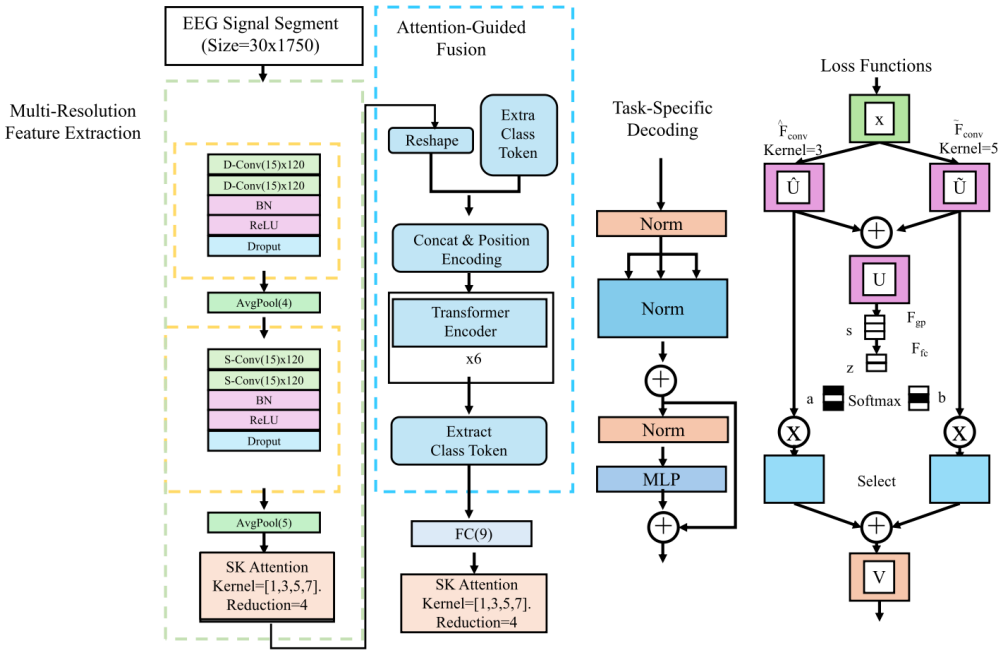


Figure 3. Architecture of AMRI-Net model [1]

In another study, Carreras et al. [27] used the ResNet-18 architecture to predict follicular lymphoma and reactive lymphoid tissue using hematoxylin and eosin (H&E) histological images to classify a large series of 221 cases. Explainable AI techniques including Grad-CAM, image LIME, and occlusion sensitivity were employed to enhance interpretability, and a patient-level independent validation set confirmed high classification performance.

A Spatially Aware Histology Convolutional Neural Network (SA-HCNN) model proposed in [23] for analyzing gigapixel pathology images. The key innovation in this study is the Pyramid Tiling with Overlap (PTO) technique, which provides the model with spatial awareness by using multiple resolution views of the same tissue region. This enables the model to understand not only the target area but also the surrounding tissue context. The SA-HCNN model is built on the EfficientNetV2 architecture and leverages PTO to effectively classify pathologies. Additionally, the model integrates an attention mechanism in its final classifier block to enhance its focus on the most relevant regions of the image. This approach allows for accurate pathology classification while reducing the computational complexity associated with analyzing large gigapixel images. The Results from comparisons with EfficientNetV2 and SegFormer show that SA-HCNN significantly outperformed them in tissue classification, achieving F-scores between 0.97 and 0.99. The model also demonstrated superior performance on other datasets, including BACH for breast cancer, as shown in Figure 4.

Despite their high performance, these complex CNN models have shown common limitations, including large model and dataset sizes, high computational requirements, sensitivity to dataset quality and class imbalance, and concerns regarding generalizability, particularly for models limited to a single center or task-specific scope [1], [23], [27], [31] .

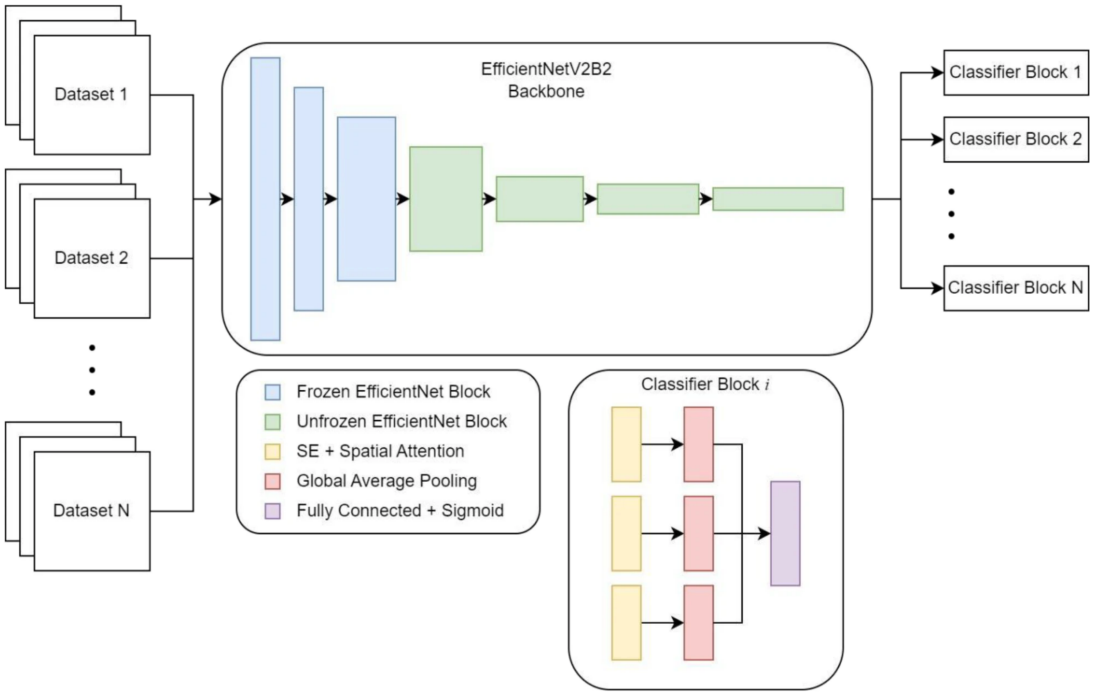


Figure 4. Architecture of SA-HCNN model [23]

3.4 Transformer and Attention-Based Models

While CNNs excel at capturing local features in histopathological images, recent research has explored transformer- and attention-based architectures to improve classification performance on complex pathology tasks.

A semi-supervised DL framework (GasMIL) presented in [32] as a diagnostic approach for gastric biopsy specimens, combined with OLGA/OLGIM for individual gastric cancer risk classification. The model used semi-supervised MIL, multi-scale patch embeddings, and transformer-based aggregation to analyze WSIs. The dataset included 2725 WSIs of 545 patients. GasMIL achieved AUC 0.884 (IM) and AUC 0.877 (atrophy), and in the observer study achieved an 80% sensitivity, 85% specificity, a weighted kappa value of 0.61, and an AUC of 0.953, surpassing the performance of all ten pathologists in atrophy detection.

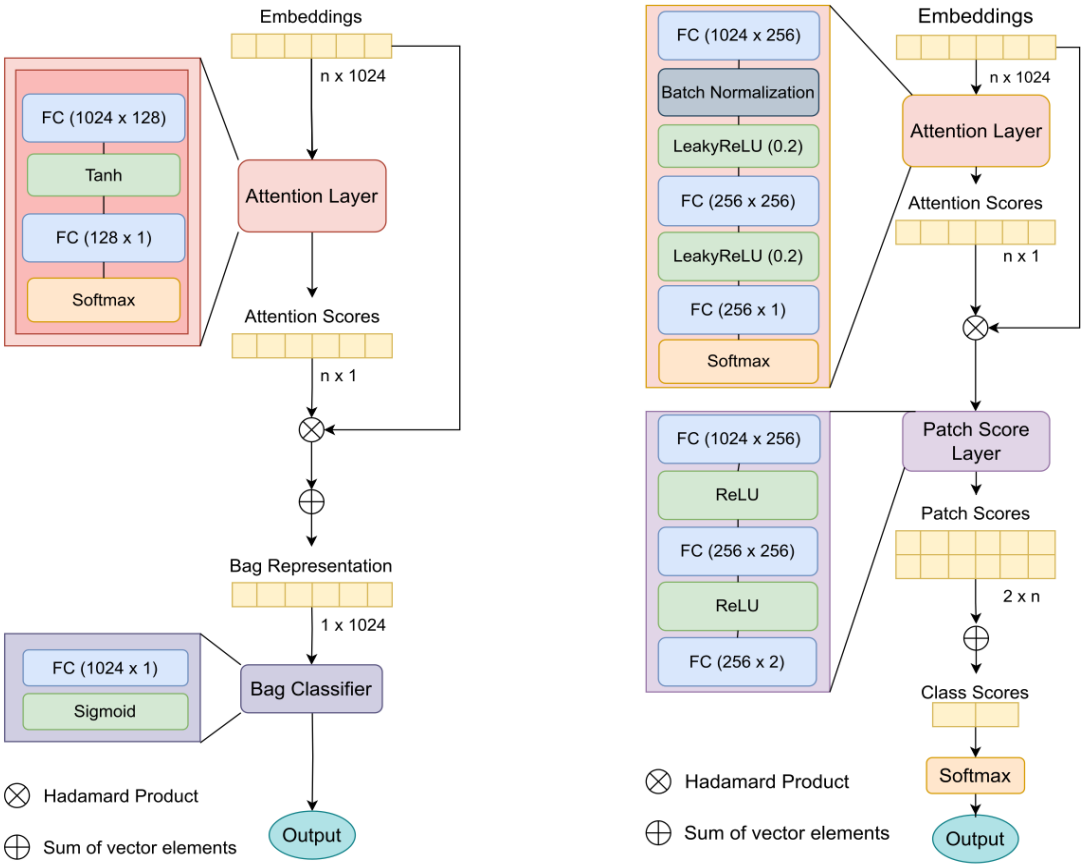
To further explore transformer integration, a framework was proposed in [25] utilizing a lightweight convolutional neural network (CNN) designed in two stages, a CNN-based tile encoder (CTE) to extract features from selected tiles, and a transformer-based feature aggregator (TFA) to combine local features of all sampled tiles into global representation for WSI classification. The model was evaluated on H&E slides from TCGA datasets on three cancer subtypes breast, lung, and kidney, which achieves higher classification accuracy and AUC compared to max-pooling.

In another study, the use of a Wave-Vision Transformer (Wave-ViT) for the pathological classification and staging of esophageal cancer was presented in [4]. The model integrated transformer-based representations with CNN feature extraction to capture both local and contextual information and was trained on the Hyper Kvasir dataset covering esophagitis, Barrett's esophagus, and adenocarcinoma. Wave-ViT achieved over 89% accuracy and showed notable sensitivity for early-stage cancer detection, surpassing conventional CNN and MLP baselines. Further comparative analysis in [26] examined three attention-based MIL architectures for WSI analysis in tumor detection and TP53 mutation detection, also explored interpretability via ROI heatmaps and proposed modified AdMIL for better patch-level insights. The TCGA-LUSC dataset was used for tumor detection, while the TCGA-BRCA was used for TP53 mutation detection. The models, Attention MIL (AMIL), Additive MIL (AdMIL), as shown in Figure 4, and modified AdMIL, were evaluated using the KimiaNet feature extractor. KimiaNet is specialized network based on the DenseNet-121 topology that, unlike general-purpose models, was trained on histopathology images from the TCGA repository. In Tumor detection, Attention MIL (AMIL) obtained the best performance with an AUROC of 0.971±0.015. Performance was lower for the more challenging TP53 Mutation Detection (20×), with AMIL achieving 0.704±0.040, AdMIL 0.624±0.033, and modified AdMIL 0.642±0.029.

A deep MIL model proposed in [33], dubbed dual-stream multiple instance learning network (DSMIL) that learns a patch and an image classifier using a two-stream architecture. It first applies standard max-pooling to identify the highest scoring patch, then computes an attention score for each patch by measuring its distance to the critical

patch. DSMIL was designed to address limitations in previous MIL approaches, such as max-pooling when only small portion of patches contain tumor tissue the models misclassify these and lead to weak performance. DSMIL was evaluated on two public WSI datasets Camelyon16 and TCGA lung cancer and compared it to a set of MIL baselines such as Max-pooling, Mean-pooling, and ABMIL. The model achieved higher accuracy and AUC across datasets. For example, on TCGA dataset, DSMIL achieved 94.7% accuracy and 0.98 AUC, outperforming Max-pooling (77% accuracy, 0.82 AUC).

Despite these results, transformer- and attention-based models face challenges including developing and verifying models separately for different gastric problems, which increased memory use and missed relationships between image representations. The black-box nature limited interpretability, and small sample sizes reduced reliability [11]. Some models faced sampling bias when the tumors were very small, overfitting when adding more transformer decoder layers, and limited generalizability due to single source data [4], [25]. Others struggled with small datasets, weak label annotations, the unreliable attention mechanisms, and the incomplete validation due to not fine-tuning the models [26].



(a) Attention MIL Architecture. **(b)** Additive MIL Architecture.

Figure 5. Architecture of AMIL and AdMIL[26]

3.5 Emerging Frameworks (Self-Supervised and Foundation Models):

Beyond transformer architectures, newer studies explore self-supervised and large-scale multi-omic frameworks, where CNNs serve as one component within broader, multi-modal DL systems.

Jaume et al. [5] proposed Madeleine As shown in Figure 6, a self-supervised learning framework that leverages multiple stains rather than relying solely on H&E slides. By treating H&E and IHC stains as complementary views, the model applied a global-local alignment strategy to learn more generalizable slide representations.

Madeleine was pretrained on over 16,000 WSIs from breast cancer and kidney transplant cohorts and tested on 21 downstream tasks across multiple centers, where it consistently outperformed single-stain approaches. The study demonstrated robust cross-institution generalization and strong performance in morphology and molecular prediction tasks.

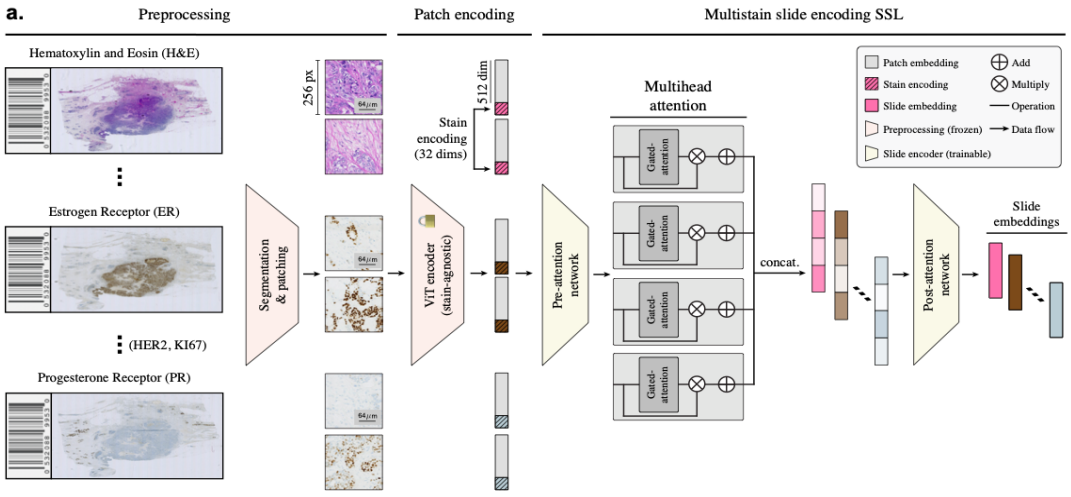


Figure 6. Architecture of Madeleine model [5]

To address the challenge of acquiring annotations and develop a new method to better capture histopathology-oriented features, another study [34] proposed CTransPath, a hybrid model that combines a CNN with a multi-scale Swin Transformer architecture. This model uses a self-supervised learning strategy called Semantically-Relevant Contrastive Learning (SRCL), which relies only on unlabeled data to generate informative representations and generalizes well to various downstream tasks. This model was pretrained on large unlabeled histopathological datasets of WSIs from TCGA and PAIP, containing around 15 million unlabeled patches in total. To evaluate the model, accuracy (ACC), AUC score, and F1 score were used as performance measures for classification tasks. The results showed that in patch-level classification on the CRC dataset, the model achieved an accuracy of 98.20%. and for WSI classification, it scored an ACC of 0.922 and an AUC of 0.942 on CAMELYON16, and AUC scores of 0.973 and 0.991 on TCGA-NSCLC and TCGA-RCC, respectively. These findings clearly show

that SRCL-based representations are more robust and transferable than other SSL methods.

For a multi-omic DL approach, a comprehensive pan-cancer study [29] evaluated the potential of DL for molecular profiling of multi-omic biomarkers directly from H&E-stained WSI. The Cancer Genome Atlas (TCGA) dataset was used. For evaluation, the CNN architecture consisting of an encoder (feature extractor), a decoder, and a profiling module for classification. A total of 12,093 unique models were trained. The results show that 50% of the models achieved an AUC of 0.644 or higher, and the AUC for 25% of the models is at least 0.719 and exceeds 0.834 for the top 5%. The highest performing models were from the standard clinical biomarkers (AUC 0.742±0.120). The lowest average performance was seen in the prediction of SNVs in driver genes (AUC 0.636±0.117).

Similarly, another study [35] proposed a novel slide-based histopathology analysis framework that creates a WSI representation map called HipoMap, which can be applied to any slide-based problems, coupled with CNNs. As shown in Figure 7, the HipoMap converts a WSI of various shapes and sizes to structured image-type representation. A generalized task-independent framework for slide-based analysis is essential, since most state-of-the-art slide-based analysis methods are task-dependent, allowing for efficient training of the model without pixel-wise (ROI) annotations. The HipoMap framework itself is a representation method analyzed by a separate CNN and the Google-Brain (GB) and CAncer-Texture Network (CAT-Net) models were used as patch-wise pretrained CNN models. The GNUH (Gyeongsang National University Hospital) lung cancer biopsies dataset was used for cancer classification (tumor vs non-tumor) and subtype classification. The TCGA-LUAD dataset was used for survival analysis and survival prediction. It was also further analyzed on TCGA-STAD and TCGA-COAD. For cancer classification, HipoMap produced the best AUC of 0.966±0.026 and 0.945±0.034 with the pretrained models of CAT-Net and GB respectively. For survival analysis, HipoMap achieved the highest c-index of 0.787±0.013, and 0.763±0.016 with pretrained models of CAT-Net and GB respectively,

which was 4.7% improved compared to second highest benchmark of RNNSA. For survival prediction, the RMSE and R2 of HipoMap were 2.77±0.36 and 0.978±0.032.

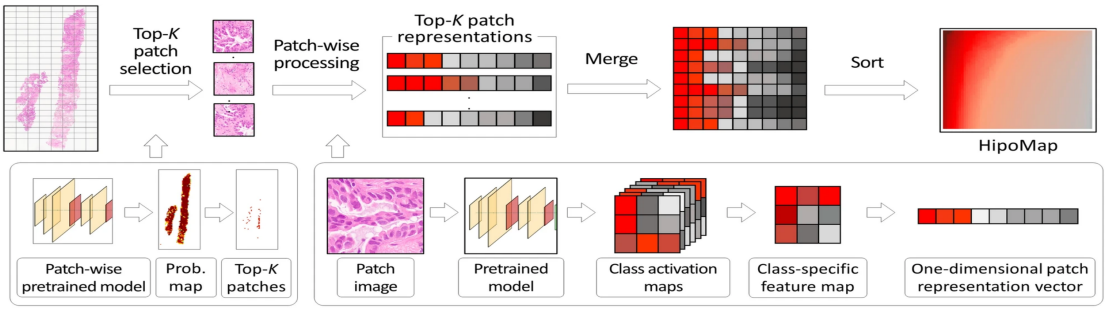


Figure 7. Architecture of HipoMap model [33]

Still, several limitations were noted across these studies. For Madeleine [5], high computational costs and requiring access to multistain cohorts that are not always available were major challenges. In Arslan et al. [28], limitations included restricting the scope of biomarker acquisition to maintain a manageable scope for the research, data heterogeneity since all biomarkers were tested under different sample sizes and prevalence conditions, the nature of the TCGA dataset of having site-specific fingerprints inherent in digital slides that could introduce bias, and the unreliable AUC values in the minority class due to very few examples. For HipoMap [35], even though it was meant to be robust to noises and outlier patches, the method was highly reliant on the hyperparameter K and showed possibility for false alarms on normal tissue of non-tumor patients who had severe chronic inflammation.

3.6 Hybrid and Explainable AI Methods

While many studies focus on advancing single-model architectures, some studies explore hybrid approaches that combine different models, such as a CNN with a classical ML classifier or multiple CNNs to further enhance performance.

Before the adoption of end-to-end DL, earlier frameworks relied on handcrafted features combined with ML classifiers. in one such frameworks [33], textural and pathologist-annotated features alongside (SVM) and a DL CNN based on Alexnet and

LeNet architectures. for the classification task of osteosarcoma WSIs into non-tumor, necrotic tumor, and viable tumor. This study used archival tumor samples from 50 patients (1995–2015), consisting of 942 WSIs. The results showed both SVM and DL models achieved high accuracy in distinguishing tissue classes, confirming the suitability of deep neural networks in histopathology image interpretation [36].

While this study demonstrated the strong preformance of hybrid ML and DL methods, it also highlighted a critical dependency on manual pathologist identified features, highlighting the need for automatic feature extraction approaches. This is addressed by more advanced hybrid models. For example, Du et al [2] produced a study combining DL and pathomics to classify the invasiveness of lung adenocarcinoma (LUAD) presenting as ground-glass nodules (GGNs). The study aimed to provide a supportive AI tool for diagnosis. The approach used three CNNs (ResNet18, ResNet50, and ResNet101) for tumor-region recognition, followed by machine-learning classifiers applied to pathomic features extracted from the best performing model. The dataset included 289 surgically resected LUAD cases with manually annotated whole-slide images, from which over 2.5 million patches were processed with stain normalization and augmentation. ResNet18 achieved the strongest tumor localization performance with 90.4% accuracy and an AUC of 0.956. For invasiveness prediction, RF performed best with test accuracy of 81.4% and AUC of 0.807, outperforming other classifiers. AI assistance also improved diagnostic accuracy for junior and intermediate pathologists, raising their AUCs by 0.212 and 0.113 respectively.

A related study [12] develop an AI-based workflow for the classification of salivary gland tumors (SGTs), aiming to distinguish benign from malignant cases, identify four common malignant subtypes, and predict tumor grade. Motivated by the diagnostic challenges of SGTs, which are rare and histologically diverse, the study compared ML models with DL CNNs. RF classifiers achieved strong performance with F1-scores of 0.95 for benign vs. malignant and malignant subtyping, and 0.87 for grading, with AUROC values up to 0.97, while DL models performed less effectively. The dataset included 320 H&E-stained WSIs from a single tertiary center, with internal splits and an external validation cohort confirming generalizability for benign vs.

malignant and grading tasks. However, limitations included relatively small cohort size, reliance on manual annotation, restriction to certain CNN architectures, failure to detect minor invasive foci, the exclusion of rarer malignant subtypes, lack of external validation for some tasks, and weaker DL performance due to modest data size [2], [12].

In addition to hybrid studies, an explainable hybrid framework [7] that combines DL feature extraction with rule-based ML. The study introduced the SVIS-RULEX framework as shown in Figure 8**,** combining MobileNetV2 model for feature extraction and Decision Tree and RuleFit classifiers for interpretable rule generation. The framework was trained on five publicly available Kaggle medical imaging datasets covering a range of diagnostic tasks such as COVID-19 detection from chest X-ray, breast tumor classification, brain tumor identification, lung and colon cancer recognition, and glaucoma detection. In total, more than 70,000 samples were used, with data augmentation techniques applied to improve generalization. Among all tested models, MobileNetV2 achieved the highest performance, reaching 94% accuracy and an F1-score of 0.93 across datasets. The Decision Tree and RuleFit classifiers also performed well, with accuracy above 85%, showing a clear explanation. Overall, the SVIS-RULEX framework proved that combining DL with explainable ML can yield in accurate and generalizable results across multiple medical imaging domains. However, the study limitations included the oversimplification of rule-based explanations, a potential mismatch between visual heatmaps and clinical interpretation, and the high computational cost associated with combining multiple explainability methods.

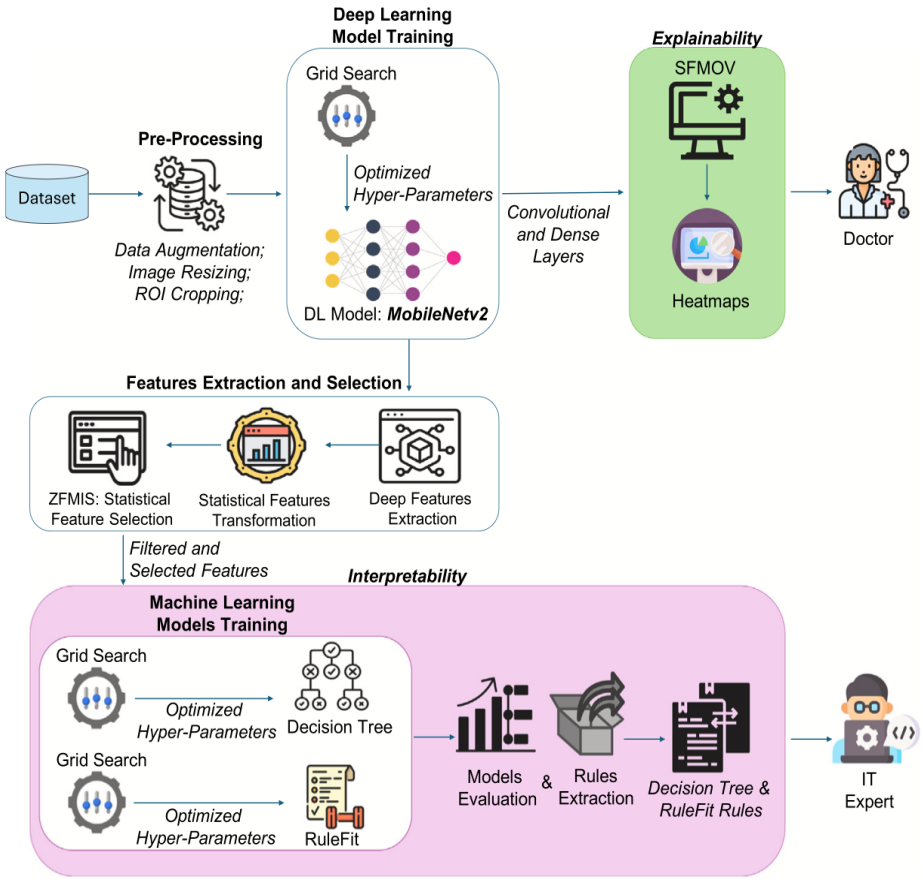


Figure 8. Architecture of SVIS-RULEX model [7]

Multiple CNNs were combined in an ensemble framework [22] to enhance the overall performance of a system. This study used an ensemble model consisting of three CNNs: InceptionV3, InceptionResNetV2 and VGG16, which was based on averaging techniques and by combining their respective predictions. Model was trained on a publicly available dataset from KVASIR, following the training of the model, they employed Shapley additive explanations (SHAP) that explains more broadly and mathematically giving precise value for the contribution of each feature. Additionally, Local interpretable model-agnostic explanations (LIME) a tool that explains why the model made a particular decision by focusing on the most important parts of the image that impact the result. Results showed a positive and encouraging advancement in the exploration of explainable AI (XAI) approaches, and a significant improvement in the

overall accuracy compared to the individual models, specifically in the context of gastrointestinal cancer detection within the healthcare domain. The study was limited to KVASIR dataset, and while ensemble models improve performance, they also increase computational cost, which can limit their practicality in clinical usability.

The latest progress in DL is the use of large multimodal language models for pathology image classification. In recent work [37], PathChat+, as shown in Figure 9, was introduced as a Multimodal Large Language Model (MLLM) for human pathology, trained on over 1 million samples and nearly 5.5 million question-answer turns. To extend its diagnostic reasoning capabilities, they developed SlideSeek, a multi-agent AI system leveraging PathChat+ to autonomously evaluate whole-slide images (WSIs). The model was evaluated on public datasets, including PathMMU, UniToPatho, BRACS, and HiCervix, achieving approximately 80% accuracy in primary diagnosis and 93% accuracy in differential diagnosis, and further assessed using their curated datasets, PathQABench and DDxBench for reasoning and interpretability. Outperforming models such as GPT-4o, Claude 3.5, and Gemini 2.0. Despite these promising results, the evaluation was primarily retrospective and has not yet been validated in real clinical settings. Additionally, SlideSeek still requires improvement to handle more complex cases with multiple tissue sections and to incorporate multimodal information to further enhance clinical utility.

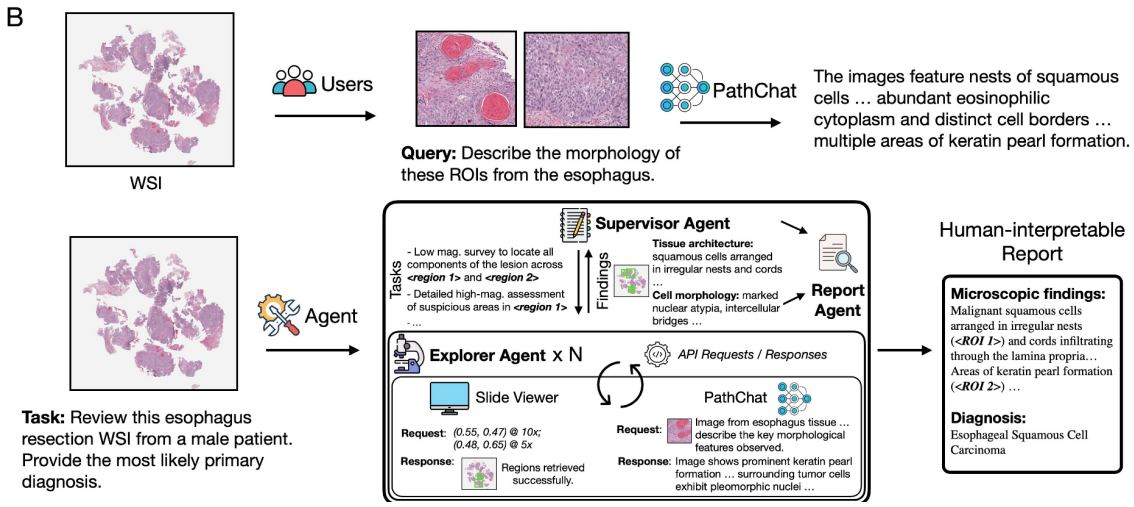


Figure 9. Architecture of PathChat+ model [37]

3.7 Summary and Research Gap

This literature review shows the great advancements and evolution of pathology image classification, starting with the more traditional ML approaches that use handcrafted features, to automatic feature extraction and improved overall performance in DL frameworks, and lastly taking advantage of more than one model in hybrid approaches as shown in Table 3.

The overall evolution shows great potential in achieving more automated, explainable, and generalizable diagnostic tools, which is highly promising for future applications in DP. The main strengths that were noted by many studies are the ability to apply full automation in the diagnostic workflow, improve diagnostic accuracy, adapt to various cancer types, and gradually enhance model explainability and scalability.

Across all approaches, recurring challenges remain, such as limited data diversity, high complexity and computational cost, along with limited interpretability. These limitations highlight the gap between performance and real clinical applicability, showing that not all research progress leads to practical use. Additionally, there is still a lack of

systematic, reproducible benchmarking of established DL models on large-scale public datasets.

The aim of this project is to address part of this gap. Our research will develop a robust classification pipeline for binary (Tumor vs. Normal) classification. By conducting a comparative analysis of several DL models on a suitable dataset, this project will provide a clear benchmark of their performance, allowing a reproducible pipeline and providing an empirical analysis of overall best model performance for this task in CP.
By conducting a comparative analysis of several feature extractors models with one classifier on a suitable dataset, this project will provide a clear benchmark of their performance, allowing a reproducible pipeline and providing an empirical analysis of overall best model performance for this task in CP**.**

Table 3: Summary of Literature Review

Approach Type	**Reference Number**	**Architecture**	**Specific Model**	**Datasets**	**Key Findings**
ML	[29]	AutoML (Traditional ML)	Google Vertex AI	Alberta Precision Labs, Leeds University, TCIA	Recall & Precision: ~ 100%
	[8]	Feature Aggregation	Fisher Vector	Camelyon17, TCGA	Accuracy: 80%, 85%
DL	[21]	CNN (simple)	XLLC-Net	LC25000	Performance: 0.99 - 1.00
	[3]		Custom CNN	Open-source H&E datasets	Accuracy: 97%
	[30]	CNN (multi-stage)	InceptionResNet-V2	A dataset gathered from three medical centers	Reliable tumor identification & moderate EGFR prediction
	[31]	CNN (Complex)	ResMAPNet	Figshare, Kaggle	Accuracy: 99.51%
	[1]		AMRI-Net	ISIC, HAM10000, OCT2017, etc.	Accuracy: 94.95% F1 score: 94.85%,

	[27]		ResNet	eosin (H&E) histological images	High classification performance
	[23]		SA-HCNN	A dataset of 221 cases	F-score: 0.97-0.99
	[32]	Transformer / MIL	GasMIL	A dataset from 13 different institutions	Sensitivity: 80% Specificity: 85% AUC: 0.953
	[25]	Transformer / CNN	Custom	TCGA	high classification accuracy and AUC
	[4]	Transformer-based	Wave-ViT	HyperKvasir	Accuracy: >89%
	[26]	Attention-based MIL	AMIL, AdMIL	TCGA-LUSC, TCGA-BRCA	(Tumor Detection) AUROC: 0.971
	[32]		DSMIL	TCGA-LUSC, Camelyon16	Accuracy: 94.7% AUC: 0.98
	[5]	Self-Supervised	Madeleine	A dataset of 16,000 WSIs	Strong performance
	[34]	Self-Supervised / Hybrid	CTransPath	TCGA, PAIP	Patch Acc: 98.20%; WSI AUC: 0.942
	[28]	Large-Scale DL + Multi-Omic (CNN)	Custom	TCGA	AUC: 0.742±0.120
	[35]	Large-Scale DL (CNN)	HipoMap	GNUH, TCGA-LUAD, TCGA-STAD, TCGA-COAD	(Cancer Classification) AUC: 0.966±0.026
	[37]	Foundation Model (MLLM)	PathChat+	PathMMU, UniToPatho, BRACS, HiCervix,	Accuracy: 93%
Hybrid	[36]	Hybrid (Early CNN+ML)	AlexNet/LeNet + SVM	A dataset of 942 WSIs	High accuracy
	[2]	Advanced Hybrid (CNN+ML)	ResNet + Random Forest	A dataset of 289 cases	Accuracy: 81.4% AUC: 0.807
	[12]		Custom CNN vs. Random Forest	A dataset of 320 H&E-stained WSIs	F1-score: 0.95

	[7]	Explainable hybrid	MobileNetV2 + RuleFit	Kaggle	Accuracy: 94% F1-score: 0.93
	[22]	Multiple CNNs (Ensemble)	InceptionV3, InceptionResNetV2, VGG16	KVASIR	Improvement in overall accuracy

Chapter 4: Methodology

4.1 Overview of the Proposed WSI Classification Model

For our approach in this research, we provide a pipeline for classifying WSIs from the TCGA-LUSC dataset using AMIL vs. MIL baseline Max-pooling, by comparing three different pretrained models (KimiaNet, CTransPath, and MobileNetV2) as they will be used as frozen feature extractors. The AMIL architecture will be detailed in Section 4.5, while the baseline MIL model will be covered in Chapter 5. The pipeline begins by dividing each WSI into smaller patches in the preprocessing step to handle its large size. Then, each of the feature extractors takes the patches as input, then convert these patches into numerical embeddings that represent their visual characteristics. These embeddings are then passed to both lightweight MIL classifiers. Finally, each model produces a slide level prediction (Normal or Tumor). To ensure a fair comparison between feature extractors, each of them will have the same rule and follow the same steps. The aim of comparing different feature extractors is to examine how model choice affects the overall WSI classification performance, and to determine which one provides the most effective patch embeddings.

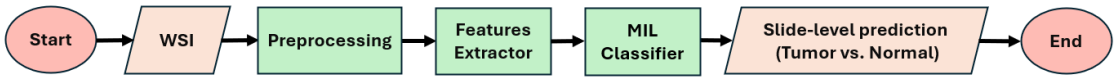


Figure 10: Proposed Classification Pipeline

This pipeline that is shown in Figure 10 was chosen because it provides a practical and efficient approach to WSI analysis. Full slide training is often computationally expensive and requires large GPU resources, which makes it difficult to implement in small scale research. By using pretrained models in frozen mode at the slide level and applying the same AMIL, the workflow becomes more lightweight while still maintaining strong performance. Furthermore, using the same AMIL architecture across all feature extractors ensures a fair comparison. As part of the pipeline, each WSI will undergo by preprocessing step as shown in Figure 11, where (a) represents thumbnail of

WSI, (b) shows the tissue segmentation mask of WSI, (c) represents the patch grid generated from the tissue region of WSI, and (d) shows patches after filtering to get informative regions of WSI.

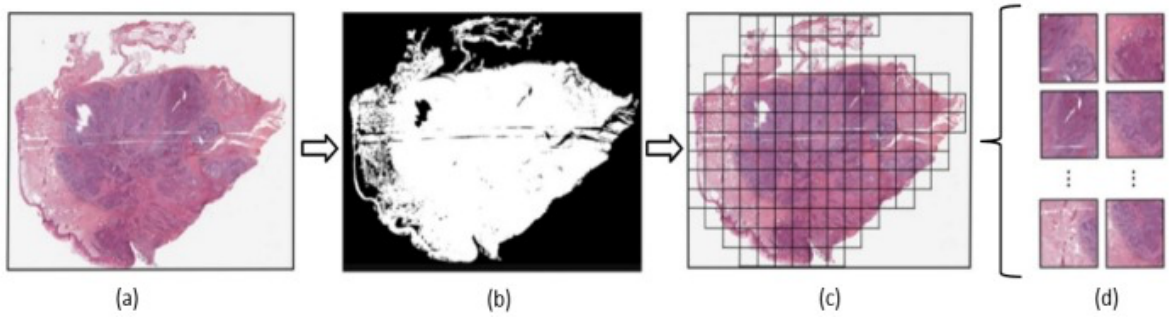


Figure 11. WSI preprocessing framework [8]

Following the preprocessing step, the comparative classification pipeline, shown in Figure 12, begins with the preprocessed patches being processed by one of three frozen feature extractors to generate embedding vectors: KimiaNet produces 1×1024 vector, CTransPath produces 1×789 vector, and MobileNetV2 produces 1×1280 vector. The embeddings from each feature extractor will then be processed separately and passed to both the AMIL and MIL classifiers, where each classifier will output a single slide-level prediction (probability of Tumor vs. Normal).

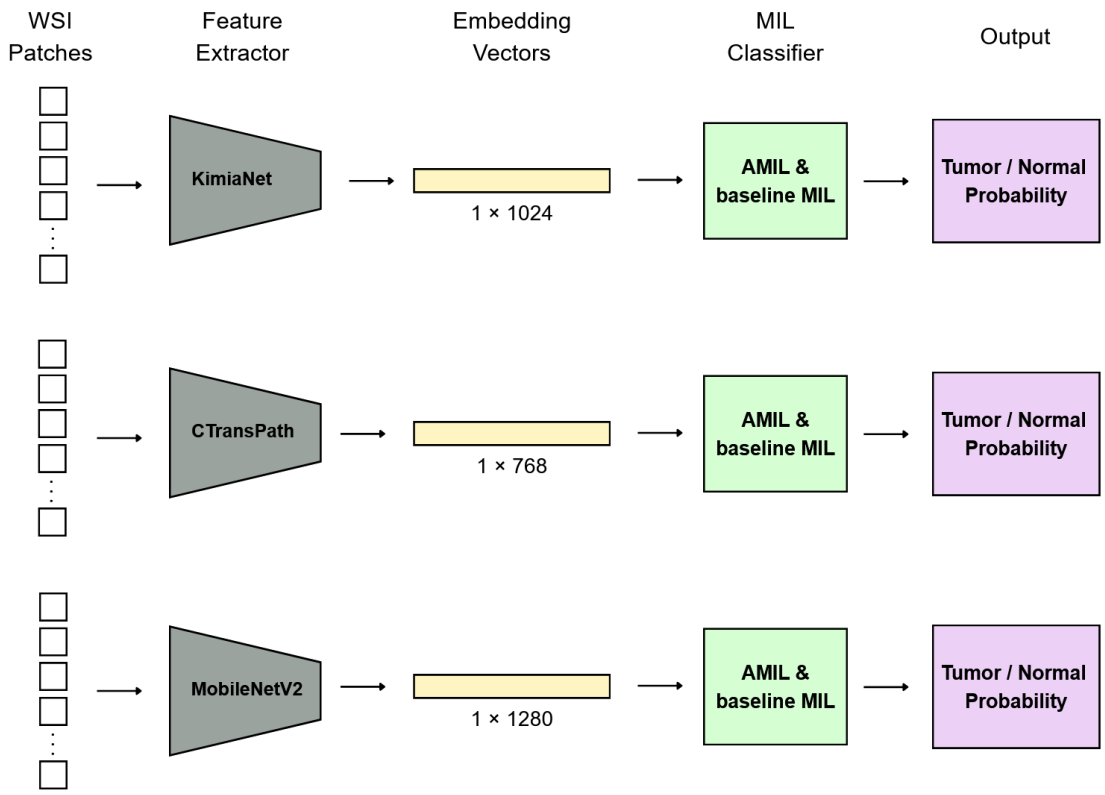


Figure 12. Overview of the comparative classification pipeline

4.2 Dataset: TCGA-LUSC

For our comparative analysis, we will be using the TCGA-LUSC public cancer dataset, which is composed of WSIs from Lung Squamous Cell Carcinoma. The dataset was selected due to its extensive use in digital pathology research and the availability of publicly accessible WSIs with slide level diagnostic labels suitable for MIL frameworks. The dataset is composed of 504 cases, containing 512 FFPE slides and approximately 1100 flash-frozen slides. For our purpose, a balanced subset of 694 flash-frozen slides (347 negative and 347 positive) will be used. Since WSIs in this dataset are huge, with resolutions up to 100,000 × 100,000 pixels, they do not allow for direct feeding into the model. This will be addressed by dividing each WSI into multiple patches that serve as instances of within a “bag” (the slide), This paradigm is known as Multiple Instance Learning (MIL) [26].

4.3 Preprocessing

A preprocessing pipeline is essential to transform the raw data from the dataset into a format suitable for model training. As a first step, tissue will be segmented from the background by detecting pixel regions that correspond to tissue based on their color and intensity. This ensures that only patches containing meaningful tissue information are kept for further processing.

After tissue detection, each WSI in the TCGA-LUSC dataset will be divided into patches of fixed size 512×512 pixels, using a non-overlapping stride of 512 at a $\times 5$ magnification. This magnification is chosen for its lower number of tiles per slide reduces storage consumption, and because at this level, patches show tissue-level information, which is more adequate for tumor detection tasks. These patches will serve as the basic instances for the MIL paradigm described in Section 4.5.

For patches consisting exclusively of background or artifacts, a patch filtering step is applied by detecting tissue regions, removing blank regions or artifacts, and discarding any patch that does not contain enough tissue. In addition, to improve model generalization, the dataset will be expanded by applying data augmentation to the patches during the training phase only. These augmentation methods include random HED stain perturbation, Gaussian noise addition, rotations, horizontal and vertical flips.

After patch preparation, the dataset will undergo standard operations to ensure consistent input quality. These operations include resizing all images to an input resolution of 512×512 pixels for KimiaNet and CTransPath, while 224×224 pixels for MobileNetv2, normalizing pixel intensities using standard normalization parameters to account for variations in scanner and staining quality, and performing a balanced split into training, validation, and test sets.

4.4 Feature Extraction

Pretrained models are neural networks that have been previously trained on large datasets and have learned general visual representations that can be transferred to new

tasks. In this project, we will compare the performance of different model architectures as feature extractors for the selected classification pipeline. The main role of the feature extractor is to convert the image patches into numerical vectors (embeddings) that can be used as the input for the MIL classifiers. Each of the following models will be utilized in a frozen state, meaning their parameters will remain fixed during feature extraction. Keeping the model parameters fixed during feature extraction ensures that all feature extractors operate under the same conditions, which allows for a fair and consistent comparison across models. Using pretrained networks in a frozen mode also helps prevent overfitting, since the pretrained weights that are already optimized on large and diverse datasets remain unchanged. In addition, freezing the parameters reduces computational cost associated with fine-tuning large models.

4.4.1 KimiaNet:

KimiaNet is a deep neural network that employs the topology of the DenseNet with four dense blocks, specifically utilizing the DenseNet-121 architecture, with preceding convolutional and pooling layer as shown in Figure 13. Unlike models trained on natural images, KimiaNet was fine-tuned and trained with histopathology images from the TCGA repository. While the repository covers 32 primary diagnoses, the model was trained on a subset of 30 primary diagnoses, including Lung Squamous Cell Carcinoma (LUSC) dataset [38].

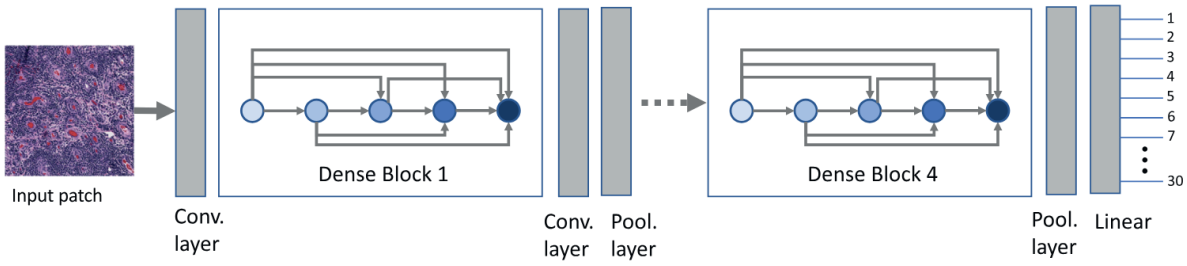


Figure 13: DenseNet architecture of KimiaNet. [38]

In this project, KimiaNet serves as a pathology-specific feature extractor. This model was chosen because it is optimized for the target domain and has demonstrated strong performance in previous studies using AMIL for classification. The way KimiaNet

operates in the proposed pipeline as a feature extractor is by converting each 512×512 pixel patch into an embedding vector with a length of 1024, derived from the network's final pooling layer.

4.4.2 CTransPath:

CTransPath is a pretrained hybrid CNN-Transformer network designed specifically for histopathology images, which makes it suitable for WSI-level analysis. It has been shown to function effectively as a feature extractor since its designs combine the local feature extraction capability of CNNs with the global structural patterns of Transformers. Its architecture begins with a narrow convolutional module composed of three sequential layers with kernel sizes of (3, 3×3, and 1×1). This initial CNN stage captures detailed local tissue patterns within each patch. The representation is then passed to the hierarchical Swin Transformer encoder, where window-based self-attention (W-SA) is applied. W-SA is computed as:

$$W - MA(Q, K.V) = Softmax\left(\frac{QK^T}{\sqrt{d}} + B\right)V \quad (4.1)$$

Where Q, K, and V are the query, key, and value projections of each window, d denotes the feature dimensionality, and B represents the relative positional bias.

To further enable information exchange beyond individual windows, the Swin Transformer incorporates shifted-window self-attention (SW-SA), where windows are shifted before recomputing attention [24]. This shifting mechanism allows the model to capture cross-window relationships that standard window-based attention cannot model. Together, the W-SA and SW-SA operations allow the network to relate distant regions within the patch, capturing broader contextual information that the CNN layers alone cannot. CTransPath provides stable feature representations suitable for WSI-level analysis by converting each preprocessed patch into a compact embedding. As illustrated in Figure 14(A), the final global pooling layer outputs a 768-dimensional vector for every patch [34].

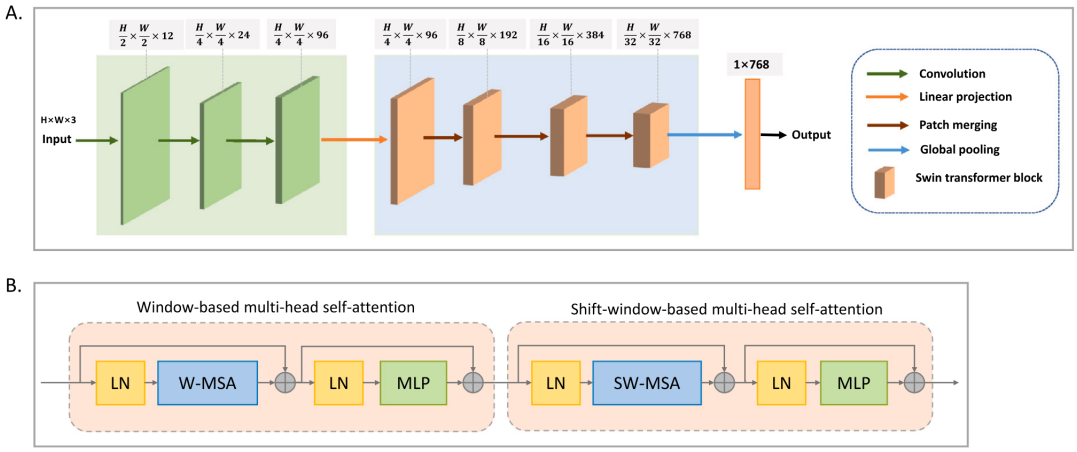


Figure 14: The architecture of the CTransPath hybrid CNN–Transformer backbone. (A) illustration of the convolutional module block (B) illustration of a Swin transformer block [34]

In this project, CTransPath will be used in frozen mode as a feature extractor. Although the model was originally pretrained on patches of size $1{,}024 \times 1{,}024$, we standardize the input to 512×512 pixels in our pipeline. Once the feature representations are generated through its convolutional and transformer components, the resulting 768-dimensional embedding vectors will be used as the input to MIL classifiers for the slide-level prediction stage.

4.4.3 MobileNetV2:

MobileNetV2 is a lightweight CNN architecture designed for efficient feature extraction, relying on depthwise–separable convolutions, and inverted residual blocks to reduce the number of parameters and computations while preserving high representational power. Because of its effectiveness in capturing complex visual patterns and deep informative features. MobileNetV2 is suitable for use as a frozen backbone for patch-level embedding extraction in WSI pipeline. After forwarding each patch through the model, the high-dimensional feature maps are then converted into a compact fixed-size vector using Global Average Pooling (GAP), which averages each feature map into a single representative value.

To formally describe the extraction process, each input patch is passed through the frozen MobileNetV2 backbone to obtain a deep embedding vector $\mathbf{F}_{\text{deep}}(\boldsymbol{I}_x) \in \mathbb{R}^d$.

These embeddings represent numerical features used by MIL classifiers, constructing the deep feature matrix:

$$\mathbf{F}_{\text{deep}} = \left[\mathbf{F}_{\text{deep}}(\boldsymbol{I}_1), \mathbf{F}_{\text{deep}}(\boldsymbol{I}_2), \dots, \mathbf{F}_{\text{deep}}(\boldsymbol{I}_m)\right]^T \tag{4.2}$$

The extracted features are then used for statistical transformation to compute interpretable, and ZFMIS statistical feature selection technique ranks and filter features based on their interpretive value [7].

In this research, MobileNetV2 will be used as a general-purpose feature extractor alongside pathology-specific models (KimiaNet and CTransPath) with the classification pipeline. It was selected due to its efficiency, low computational cost, and proven ability to extract deep discriminative features, making it suitable for large-scale WSI pipelines. Unlike the pathology-specific models which handle larger inputs, standard MobileNetV2 architectures require an input resolution of 224×224 pixels. Therefore, the standardized 512×512 patches will be down sampled specifically for this branch. We will pass each resized patch through MobileNetV2 to obtain an embedding vector of size 1280 from the final layer. These embeddings will serve as numerical input instances for the MIL classifiers.

The combination of KimiaNet, CTransPath, and MobileNetV2 allows a balanced comparison between domain-specific, hybrid, and lightweight CNN-based extractors, summarized in Table 4. This setup supports a fair comparison between different model architectures and provides insights into how architectural design impacts both accuracy and computational efficiency.

Table 4: Summary of Feature Extractors architecture type, pretraining dataset, input resolution, and embedding size.

Feature Extractor	Architecture Type	Pretraining Dataset	Original Task	Input Size (Pixels)	Output Embedding Size
KimiaNet	(DenseNet-121)	Histopathology (TCGA)	Tumor Primary Diagnosis Classification (30 classes)	512×512	1024
CTransPath	Hybrid (CNN + Transformer)	Histopathology (TCGA + PAIP)	Semantically-Relevant Contrastive Learning (SRCL)	512×512	768
MobileNetV2	Lightweight CNN	Natural Images (ImageNet)	General Object Classification (1000 classes)	224×224	1280

Table 4 details each model's architecture type, the dataset and specific task used for its original pretraining, and the input resolution and output embedding size. These models are compared to evaluate the impact of different architectural designs and pretraining domains, whether they are specific to pathology or based on general images, on performance when functioning as frozen feature extractors within the proposed classification pipeline.

4.5 Attention-Based Multiple Instance Learning (AMIL)

The Attention Multiple Instance Learning (AMIL) architecture was chosen as one of the MIL classifiers. It was chosen for its strong performance and efficiency. It achieved an AUROC of 0.97 on the TCGA-LUSC dataset, which was one of the highest in the literature review as shown in Table 3. In addition to that, it is considered lightweight, making it a great choice for students' hardware. The model also applies an attention mechanism to the MIL framework, which provides some interpretability that could help

with the DL black-box problem in pathology. The AMIL uses a weakly supervised approach, where only the overall slide label is known, but the individual patch labels are unknown [26].

The model is defined by a 3-part function g:

$$g(x) = (p \circ a \circ f)(x) \tag{4.3}$$

Where f is the feature extractor, a is the Attention Module, and p is the Predictor. The feature extractor f returns the respective embedding vector h_i for each instance x_i of the bag x with n instances, where $h_i \in h = [h_1, \dots, h_n]$:

$$h_i = f(x_i) \tag{4.4}$$

The Attention Module a returns an attention score α_i, for each embedding vector h_i, which determines the importance of each image patch (embedding h_i). This score is computed in two sequential steps, the first one being a neural network layer that uses the tanh activation function ψ_a to calculate an initial weight for the patch:

$$\psi_a(h_i) = w^T \tanh\ (Vh_i^T) \tag{4.5}$$

Where w is a weight vector and V is a weight matrix. Both are trainable parameters.

Second is a softmax function that normalizes these initial weights to give the final normalized score α_i that tells the model where to focus its attention for a more accurate classification:

$$a_i(h_i) = \text{softmax}_i\ \big(\psi_a(h)\big)h_i \tag{4.6}$$

And finally, the predictor function or bag classifier p that uses the attention scores α_i to compute a weighted sum of all instance embeddings, and passes this aggregate vector through a final classifier defined as ψ_p, which is a multi-layer perceptron (MLP), to produce the final classification:

$$p(h) = \psi_p \left(\sum_{i=1}^{n} (\alpha_i \cdot h_i) \right) \tag{4.7}$$

This final output $p(h)$ is the model's prediction of whether the whole slide contains a tumor or is normal.

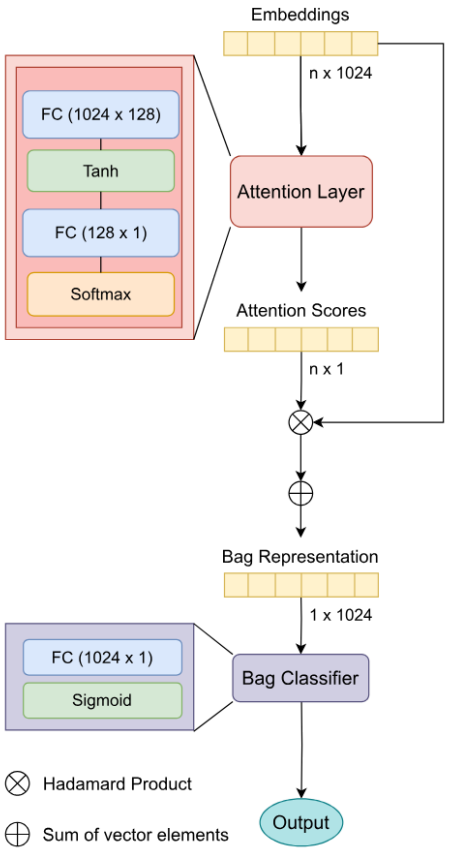


Figure 15. Architecture of AMIL [26]

This mathematical framework is shown more clearly in Figure 15. Note that while the diagram illustrates using the baseline embedding size of 1024 (from KimiaNet), the input layer of the head is adjusted to match the output dimension of the specific feature extractor will be tested in the experiment (e.g., 768 for CTransPath, 1280 for MobileNetV2).

The process starts when the list of embeddings created by the feature extractor enters the attention layer, which contains the tanh activation function that generates the

attention scores. The scores are then normalized by the softmax function to produce the final attention scores, that are used as weights to perform a sum of vector elements combining them into a single bag representation vector that summarizes the most important information from the entire slide. Finally, this bag representation is passed to the bag classifier, a simple fully connected layer with a sigmoid activation function to output a single probability score representing the final binary classification of the WSI as either tumor or normal.

4.6 Performance measures

As part of the experimental comparison, the performance of the different feature extractors within the proposed classification pipeline will be evaluated using a set of standard classification metrics shown in Table 2. All metrics will be computed at the slide level by comparing the model's prediction with the actual labels from the dataset. These measures provide a clear evaluation of how effectively each feature extractor, combined with the MIL classifiers, can distinguish between normal and tumor slides.

4.7 Summary

This chapter presents the methodology that will be used to develop the proposed WSI classification pipeline. The pipeline will begin by dividing each WSI from the TCGA-LUSC dataset into smaller image patches to handle the extremely large resolution of WSIs. The extracted patches will then undergo a preprocessing procedure that will include resizing to a unified input resolution, pixel-intensity normalization, and splitting into training, validation, and test sets. Next, three pretrained models (KimiaNet, MobileNetV2, and CTransPath) will be employed in frozen mode as feature extractors to convert the patches into numerical embeddings that represent their visual characteristics. These embeddings will then serve as input to the AMIL and the baseline MIL, which will learn how to aggregate the patch-level features and produce the final slide-level prediction. Using different models as feature extractors in this pipeline will allow us to investigate and determine which model provides the most effective patch embeddings for the same AMIL classifier. Finally, the performance of all feature extractors will be

evaluated using standard metrics such as Accuracy, AUC, Precision, Recall, F1-score, and Specificity. These measures will provide a clear assessment of how effectively each feature extractor, in combination with the MIL classifiers, can distinguish between normal and tumor slides.

Chapter 5: Experimental Design

5.1 Introduction

This chapter outlines the experimental design used to evaluate the performance of the proposed WSI classification pipeline described in Chapter 4. The main goal of these experiments is to assess how different feature extractors affect the overall classification performance of the lightweight AMIL and MIL models when applied to the TCGA-LUSC dataset.

Specifically, the experiments aim to validate whether pretrained frozen feature extractors, including domain-specific KimiaNet, hybrid CTransPath, and lightweight MobileNetV2, can provide informative embeddings patch representations for slide-level classification while maintaining computational efficiency.

5.2 Dataset Split

The TCGA-LUSC dataset will be divided into three subsets: training, validation, and testing. The split will be performed at the slide level to prevent data leakage, ensuring that slides from the same patient appear only in one subset. 70% of data will be used for training, 15% for validation to monitor the model's performance on unseen data, and 15% for testing to measure model's accuracy, to maintain a balanced distribution, providing enough data for model learning while ensuring reliable validation and testing.

5.3 Experimental Setup

5.3.1 Experimental Conditions

The experimental setup is designed to evaluate how different feature extraction backbones and MIL classifiers affect slide-level classification performance. Two main variables will be tested: the choice of feature extractor and the type of MIL model used, while other experimental factors will remain fixed to ensure a fair and consistent comparison.

A. Experimental Factors:

1. Feature Extractor Backbone:

Three pretrained feature extractors models (KimiaNet, CTransPath, and MobileNetV2) will be evaluated as frozen feature extractors. Each model will generate patch-level embeddings that are later processed by the MIL classifiers. As described in Chapter 4, these backbones represent different architectural designs and comparing them helps analyze how these design differences affect the quality of the extracted features which affects the overall classifier accuracy.

2. MIL Model Type:

The second experimental factor focuses on the type of MIL model used in the classification pipeline. To test the effect of adding attention mechanisms, two models will be compared, a baseline MIL model and an Attention-based MIL (AMIL) model. The baseline MIL combines all patch features using simple max pooling, where only the most highly activated feature (the patch with the strongest response) is selected to represent the entire slide. In contrast, the AMIL model learns to give more weight to the most important patches that contribute to the final slide-level prediction, as mentioned in [33]. This setup helps identify whether the attention mechanism improves the model's ability to focus on diagnostic regions and leads to better classification performance.

B. Control Conditions:

To ensure a fair and controlled comparison, several factors will remain constant across all experiments:

- Preprocessing: All WSIs will follow the same procedure described in Section 4.3.
- Patch Parameters: Patch size (512×512) and magnification level ($\times 5$) will remain consistent for all experiments.
- Frozen Backbones: All feature extractors (KimiaNet, CTransPath, MobileNetV2) will be used in frozen mode with no fine-tuning.
- MIL Classifiers Architectures: both AMIL and the Max-pooling MIL baseline (as described previously), along with their training hyperparameters will remain fixed.

- Dataset Split: The same 70 / 15 / 15 train-validation-test split defined in Section 5.2 will be used across all runs.
- Hardware Environment: All experiments will be performed on Google Colab using an NVIDIA T4 GPU environment to maintain consistent computational resources.

This setup ensures that any difference in performance comes from the feature extractor or the MIL classifiers being tested, not from variations in the data, preprocessing, or hardware.

5.3.2 Training Configuration

In this research, both AMIL and the Max-pooling MIL classifier will be trained using the same training configuration to ensure a fair and controlled comparison. Since each feature extractor outputs embeddings with different dimensionalities (1024 for KimiaNet, 768 for CTransPath, and 1280 for MobileNetV2), the input of both the Max-pooling MIL and AMIL classifiers is adjusted to match the corresponding backbone, while the hidden layer size, number of layers, activation functions, and output layer are kept identical across experiments. The models will be optimized using the Adam optimizer with a fixed learning rate of 0.0001 and a weight decay of 1e-5. Each WSI is treated as a separate bag; therefore, the batch size will be set to one bag per iteration. Training will be conducted for 30 epochs, and Binary Cross-Entropy (BCE) will be used as the loss function for the slide-level classification task. Early stopping will be applied based on the validation loss to prevent overfitting. A similar training configuration is commonly used in previous MIL-based WSI studies, which supports the suitability of these hyperparameter choices for our experiments. All feature extractors will remain frozen during training, and the same set of patch embeddings will be used for both classifiers. This ensures that any observed differences in performance arise solely from the Max-pooling MIL and AMIL architecture rather than variations in training conditions.

5.3.3 Evaluation Protocol and Comparison Strategy

To evaluate the performance of the three pretrained models, we will use the standard metrics listed in Table 2, including Accuracy, AUC, Precision, Recall, F1-score,

and Specificity. These metrics will be used to assess and compare the models' ability to classify normal and tumor slides. Slide-level predictions will be generated by aggregating all patch embeddings through the AMIL pipeline, resulting in a single label or decision for each WSI. Evaluation metrics will then be computed at the slide level, meaning they will be based on the classification of entire slides rather than individual patches. The comparison across backbones will focus on how each pretrained model (KimiaNet, CTransPath, and MobileNetV2) affects the performance of the AMIL and Max-pooling MIL classifiers when used as a feature extractor. As described in Section 4.1, each backbone will extract patch-level embeddings, which will be aggregated by the AMIL model to produce slide-level predictions. Under identical training and evaluation conditions, the performance of each backbone will be compared using the mentioned metrics. This comparison will highlight how variations in the representation quality of each feature extractor impact the overall classification performance. To further evaluate the contribution of the attention mechanism, the AMIL model will be compared with a Max-pooling MIL. Both models will process identical patch embeddings from each feature extractor, ensuring that any observed performance difference directly reflects the impact of the attention-based aggregation rather than differences in feature representation.

5.4 Summary of the Experimental Plan

This experimental plan evaluates how different pretrained feature extractors and MIL variants affect the performance of the proposed WSI classification pipeline. Three frozen backbones (KimiaNet, CTransPath, and MobileNetV2) will be tested within the same AMIL framework, alongside a Max-pooling MIL for comparison. All experiments will follow identical preprocessing, training, and evaluation settings to ensure a fair comparison and isolate the effect of the tested variables. This setup directly addresses the research question by examining how different lightweight model architectures and feature extractors influence slide-level classification accuracy and computational efficiency in WSI classification.

Chapter 6: Conclusion

Pathology image classification plays a vital role in medical diagnosis, enabling pathologists to identify disease patterns and support clinical decision-making with enhanced diagnostic efficiency and accuracy. This research explores the use of DL in DP, focusing on the classification of WSIs into tumor or normal tissue. The study reviews recent advancements in AI-driven pathology and highlights ongoing challenges. By reviewing previous research and identification of suitable publicly available datasets, this work establishes a solid foundation for developing reliable and scalable AI-based solutions that enhance diagnostic accuracy and support pathologists in clinical practice.

The research findings show that CNNs, transformers and attention-based architectures are among the most effective DL models for pathology image classification, achieving high accuracy across several datasets such as TCGA, Camelyon17, and LC25000. Even with these advances, several challenges remain, including variability in staining and imaging conditions, limited dataset diversity, the lack of interpretability in DL models, the large image sizes and the heavy computational load of these models. These limitations highlight the need for balanced approaches that combine model accuracy with interpretability and generalization along with computational efficiency.

To address these limitations, this research proposes a pipeline for classifying WSIs from the TCGA-LUSC dataset using the AMIL model benchmarked against a Max-Pooling MIL baseline. Both approaches are lightweight and provide better efficiency, while AMIL also offers some interpretability. We will employ three distinct pretrained models, KimiaNet, CTransPath, and MobileNetV2, as frozen feature extractors to ensure a clear comparative analysis. By examining how these different architectural choices affect classification performance, this project evaluates the effectiveness of lightweight models with different feature extractors in classifying WSIs as Tumor or Normal. This research provides a reproducible and empirically validated pipeline that contributes a practical solution for computational pathology to the growing field of AI in healthcare within Saudi Arabia, supporting national efforts toward medical innovation and digital transformation. It also supports the global progress in CP by promoting

diagnostic systems that are more accurate, efficient, and accessible.

In the next phase of this research, we will focus on implementation of the proposed models, performance evaluation, and comparison between different feature extractors backbone to improve classification performance. These next steps will validate our approach to identifying the most effective combination for WSI classification pipeline.

References

[1] L. He, L. Luan, and D. Hu, “Deep learning-based image classification for AI-assisted integration of pathology and radiology in medical imaging,” *Front. Med.*, vol. 12, p. 1574514, June 2025, doi: 10.3389/fmed.2025.1574514.

[2] H. Du, X. Wang, K. Wang, Q. Ai, J. Shen, R. Zhu, and J. Wu, “Identifying invasiveness to aid lung adenocarcinoma diagnosis using deep learning and pathomics,” *Sci. Rep.*, vol. 15, no. 1, p. 4913, Feb. 2025, doi: 10.1038/s41598-025-87094-5.

[3] R. S. Doğan and B. Yılmaz, “Histopathology image classification: highlighting the gap between manual analysis and AI automation,” *Front. Oncol.*, vol. 13, Jan. 2024, doi: 10.3389/fonc.2023.1325271.

[4] W. Wei, X.-L. Zhang, H.-Z. Wang, L.-L. Wang, J.-L. Wen, X. Han, and Q. Liu, “Application of deep learning models in the pathological classification and staging of esophageal cancer: A focus on Wave-Vision Transformer,” *World J. Gastroenterol.*, vol. 31, no. 19, May 2025, doi: 10.3748/wjg.v31.i19.104897.

[5] G. Jaume *et al.*, “Multistain Pretraining for Slide Representation Learning in Pathology,” Aug. 05, 2024, *arXiv*: arXiv:2408.02859. doi: 10.48550/arXiv.2408.02859.

[6] A. A. Balasubramanian *et al.*, “Ensemble Deep Learning-Based Image Classification for Breast Cancer Subtype and Invasiveness Diagnosis from Whole Slide Image Histopathology,” *Cancers*, vol. 16, no. 12, p. 2222, June 2024, doi: 10.3390/cancers16122222.

[7] N. Ullah, F. Guzmán-Aroca, F. Martínez-Álvarez, I. De Falco, and G. Sannino, “A novel explainable AI framework for medical image classification integrating statistical, visual, and rule-based methods,” *Med. Image Anal.*, vol. 105, p. 103665, Oct. 2025, doi: 10.1016/j.media.2025.103665.

[8] R. K. Gupta, D. Dharani, S. Shanker, and A. Sethi, “Efficient Whole Slide Image Classification through Fisher Vector Representation,” 2024, *arXiv*. doi: 10.48550/ARXIV.2411.08530.

[9] A. Waqas, M. M. Bui, E. F. Glassy, I. El Naqa, P. Borkowski, A. A. Borkowski, and G. Rasool, “Revolutionizing Digital Pathology With the Power of Generative Artificial Intelligence and Foundation Models,” *Lab. Invest.*, vol. 103, no. 11, p. 100255, Nov. 2023, doi: 10.1016/j.labinv.2023.100255.

[10] M. Unger and J. N. Kather, “Deep learning in cancer genomics and histopathology,” *Genome Med.*, vol. 16, no. 1, p. 44, Mar. 2024, doi: 10.1186/s13073-024-01315-6.

[11] T. Sakamoto *et al.*, “A narrative review of digital pathology and artificial intelligence: focusing on lung cancer,” *Transl. Lung Cancer Res.*, vol. 9, no. 5, pp. 2255–2276, Oct. 2020, doi: 10.21037/tlcr-20-591.

[12] I. Alsanie, A. Shephard, N. Azarmehr, P. Vargas, M. Pring, N. M. Rajpoot, and S. A. Khurram, “Exploring the feasibility of AI-based analysis of histopathological variability in salivary gland tumours,” *Sci. Rep.*, vol. 15, no. 1, p. 29171, Aug. 2025, doi: 10.1038/s41598-025-15249-5.

[13] S. Masjoodi, M. H. Anbardar, M. Shokripour, and N. Omidifar, “Whole Slide Imaging (WSI) in Pathology: Emerging Trends and Future Applications in Clinical Diagnostics, Medical Education, and Pathology,” *Iran. J. Pathol.*, vol. 20, no. 3, pp. 257–265, July 2025, doi: 10.30699/ijp.2025.2044210.3367.

[14] L. Pantanowitz *et al.*, “Review of the current state of whole slide imaging in pathology,” *J. Pathol. Inform.*, vol. 2, no. 1, p. 36, Jan. 2011, doi: 10.4103/2153-3539.83746.

[15] A. Janowczyk and A. Madabhushi, “Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases,” *J. Pathol. Inform.*, vol. 7, no. 1, p. 29, Jan. 2016, doi: 10.4103/2153-3539.186902.
[16] T. Kiyuna *et al.*, “Evaluating Cellularity Estimation Methods: Comparing AI Counting with Pathologists’ Visual Estimates,” *Diagnostics*, vol. 14, no. 11, p. 1115, May 2024, doi: 10.3390/diagnostics14111115.
[17] M. G. Hanna and O. Ardon, “Digital pathology systems enabling quality patient care,” *Genes. Chromosomes Cancer*, vol. 62, no. 11, pp. 685–697, Nov. 2023, doi: 10.1002/gcc.23192.
[18] B. Ceachi, F. Muresan, M. Trascau, and A. M. Florea, “Efficient Tissue Detection in Whole-Slide Images Using Classical and Hybrid Methods: Benchmark on TCGA Cancer Cohorts,” *Cancers*, vol. 17, no. 17, p. 2918, Sept. 2025, doi: 10.3390/cancers17172918.
[19] W. Lingle *et al.*, “The Cancer Genome Atlas Breast Invasive Carcinoma Collection (TCGA-BRCA).” The Cancer Imaging Archive, 2016. doi: 10.7937/K9/TCIA.2016.AB2NAZRP.
[20] S. Kirk *et al.*, “The Cancer Genome Atlas Lung Squamous Cell Carcinoma Collection (TCGA-LUSC).” The Cancer Imaging Archive, 2016. doi: 10.7937/K9/TCIA.2016.TYGKKFMQ.
[21] J. R. Jim, Md. E. Rayed, M. F. Mridha, and K. Nur, “XLLC-Net: A lightweight and explainable CNN for accurate lung cancer classification using histopathological images,” *PLOS One*, vol. 20, no. 5, p. e0322488, May 2025, doi: 10.1371/journal.pone.0322488.
[22] M. M. Auzine, M. Heenaye-Mamode Khan, S. Baichoo, N. Gooda Sahib, P. Bissoonauth-Daiboo, X. Gao, and Z. Heetun, “Development of an ensemble CNN model with explainable AI for the classification of gastrointestinal cancer,” *PLOS ONE*, vol. 19, no. 6, p. e0305628, June 2024, doi: 10.1371/journal.pone.0305628.
[23] C. Greeley, L. Holder, E. E. Nilsson, and M. K. Skinner, “Scalable deep learning artificial intelligence histopathology slide analysis and validation,” *Sci. Rep.*, vol. 14, no. 1, p. 26748, Nov. 2024, doi: 10.1038/s41598-024-76807-x.
[24] Z. Liu *et al.*, “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada: IEEE, Oct. 2021, pp. 9992–10002. doi: 10.1109/ICCV48922.2021.00986.
[25] Z. Fan, C. Zhang, L. Sun, W. Thorstad, H. Gay, X. Wang, and H. Li, “A computation-efficient network with feature aggregation for cancer subtype classification on histopathological images,” *Eng. Appl. Artif. Intell.*, vol. 160, p. 111913, Nov. 2025, doi: 10.1016/j.engappai.2025.111913.
[26] M. Afonso, P. M. S. Bhawsar, M. Saha, J. S. Almeida, and A. L. Oliveira, “Multiple Instance Learning for WSI: A comparative analysis of attention-based approaches,” *J. Pathol. Inform.*, vol. 15, p. 100403, Dec. 2024, doi: 10.1016/j.jpi.2024.100403.
[27] J. Carreras *et al.*, “Histological Image Classification Between Follicular Lymphoma and Reactive Lymphoid Tissue Using Deep Learning and Explainable Artificial Intelligence (XAI),” *Cancers*, vol. 17, no. 15, p. 2428, July 2025, doi: 10.3390/cancers17152428.
[28] S. Arslan *et al.*, “A systematic pan-cancer study on deep learning-based prediction of multi-omic biomarkers from routine pathology images,” *Commun. Med.*, vol. 4, no. 1, p. 48, Mar. 2024, doi: 10.1038/s43856-024-00471-5.
[29] D. Beyer, E. Delancey, and L. McLeod, “Automating Colon Polyp Classification in Digital Pathology by Evaluation of a ‘Machine Learning as a Service’ AI Model: Algorithm Development and Validation Study,” *JMIR Form. Res.*, vol. 9, p. e67457, July 2025, doi: 10.2196/67457.

[30] M. Zanoletti *et al.*, “EGFR Mutation Detection in Whole Slide Images of Non-Small Cell Lung Cancers Using a Two-Stage Deep Transfer Learning Approach,” *Cancer Med.*, vol. 14, no. 18, p. e71249, Sept. 2025, doi: 10.1002/cam4.71249.
[31] J. Zhang, R. Lv, W. Chen, G. Du, Q. Fu, and H. Jiang, “A novel residual network based on multidimensional attention and pinwheel convolution for brain tumor classification,” *Sci. Rep.*, vol. 15, no. 1, p. 31066, Aug. 2025, doi: 10.1038/s41598-025-16564-7.
[32] S. Fang *et al.*, “Diagnosing and grading gastric atrophy and intestinal metaplasia using semi-supervised deep learning on pathological images: development and validation study,” *Gastric Cancer*, vol. 27, no. 2, pp. 343–354, Mar. 2024, doi: 10.1007/s10120-023-01451-9.
[33] B. Li, Y. Li, and K. W. Eliceiri, “Dual-stream Multiple Instance Learning Network for Whole Slide Image Classification with Self-supervised Contrastive Learning,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA: IEEE, June 2021, pp. 14313–14323. doi: 10.1109/CVPR46437.2021.01409.
[34] X. Wang *et al.*, “Transformer-based unsupervised contrastive learning for histopathological image classification,” *Med. Image Anal.*, vol. 81, p. 102559, Oct. 2022, doi: 10.1016/j.media.2022.102559.
[35] S. Kosaraju, J. Park, H. Lee, J. W. Yang, and M. Kang, “Deep learning-based framework for slide-based histopathological image analysis,” *Sci. Rep.*, vol. 12, no. 1, p. 19075, Nov. 2022, doi: 10.1038/s41598-022-23166-0.
[36] H. B. Arunachalam *et al.*, “Viable and necrotic tumor assessment from whole slide images of osteosarcoma using machine-learning and deep-learning models,” *PLOS ONE*, vol. 14, no. 4, p. e0210706, Apr. 2019, doi: 10.1371/journal.pone.0210706.
[37] C. Chen *et al.*, “Evidence-based diagnostic reasoning with multi-agent copilot for human pathology,” 2025, *arXiv*. doi: 10.48550/ARXIV.2506.20964.
[38] A. Riasatian *et al.*, “Fine-Tuning and training of densenet for histopathology image representation using TCGA diagnostic slides,” *Med. Image Anal.*, vol. 70, p. 102032, May 2021, doi: 10.1016/j.media.2021.102032.