EFFICIENT DETECTION & CLASSIFICATION OF DIGITAL HISTOPATHOLOGY IMAGERY

Applications in Medical Diagnostics for Oncology

MS (Research) Thesis

By

AISHWARYA PRIYADARSHINI

2304101003

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

July 2025

EFFICIENT DETECTION & CLASSIFICATION OF DIGITAL HISTOPATHOLOGY IMAGERY

Applications in Medical Diagnostics for Oncology

A THESIS

Submitted in fulfillment of the

requirements for the award of the degree

of

Master of Science (Research)

by

AISHWARYA PRIYADARSHINI

2304101003

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

July 2025

INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled **EFFICIENT DETECTION & CLASSIFICATION OF DIGITAL HISTOPATHOLOGY IMAGERY - Applications in Medical Diagnostics for Oncology** in the fulfillment of the requirements for the award of the degree of **Master of Science (Research)** and submitted in the **Department of Computer Science and Engineering, Indian Institute of Technology Indore**, is an authentic record of my own work carried out during the time period from August 2023 to June 2025 under the supervision of Dr. Surya Prakash, Professor, Indian Institute of Technology Indore, India, and the co-supervison of Dr. Tadepalli Karuna, Professor, Department of Microbiology, All India Institute of Medical Sciences, Bhopal, India.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

Aishwarya 09-06-2025

Signature of the student with date

(AISHWARYA PRIYADARSHINI)

- -

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

09-June-25

Signature of Thesis Supervisor with date

(DR. SURYA PRAKASH)

09/06/2025

Signature of Thesis Co-supervisor with date

(DR. TADEPALLI KARUNA)

- -

Aishwarya Priyadarshini has successfully given her MS (Research) Oral Examination held on 07-July-2025.

07/07/25

Signature of Chairperson (OEB) with date

07-July-2025

Signature of Thesis Supervisor with date

07/07/25

Signature of Thesis Co-Supervisor with date

Soumi 7/7/25

Signature of Convenor, DPGC with date

07/07/25

Signature of Head of Department with date

ACKNOWLEDGEMENTS

I would like to take this opportunity to sincerely thank all those who, in one way or another, contributed to making this journey both intellectually enriching and personally fulfilling.

First and foremost, I wish to express my profound gratitude to my thesis supervisor, **Prof. Surya Prakash**, whose insightful guidance, unwavering support, and steadfast encouragement have been the cornerstone of this research journey. His deep expertise, thoughtful mentorship, and constant motivation have not only shaped the intellectual depth of this thesis but also played a pivotal role in my personal and academic development. I am truly fortunate to have worked under his guidance.

I am also sincerely thankful to my co-supervisor, **Prof. Tadepalli Karuna**, for her valuable feedback, constructive discussions, and academic support, all of which have contributed meaningfully to the refinement of this work.

I extend my sincere thanks to all the faculty members of the Department of Computer Science and Engineering for their time and involvement in evaluating my research progress, and for their constructive feedback throughout each stage. I am especially grateful to the DPGC Convener, **Dr. Aniruddha Singh Kushwaha**, for his consistent coordination and support. I also wish to express my gratitude to the Head of the Department, **Dr. Ranveer Singh**, for fostering a research-friendly environment and ensuring the academic and infrastructural resources necessary for this work.

I am truly grateful to **Prof. Suhas S. Joshi**, Director, Indian Institute of Technology Indore, for providing a dynamic academic environment that enabled me to pursue my research aspirations at this esteemed institute.

Lastly, I owe my heartfelt thanks to my family for their constant love, encouragement, and belief in me. Their unwavering positivity and support have been a continuous source of strength and inspiration throughout this journey.

Aishwarya Priyadarshini

Dedicated to my Mother and Father

Abstract

Lung cancer remains one of the most prevalent and lethal forms of cancer worldwide, necessitating accurate and timely diagnosis to guide effective therapeutic decisions. Among existing diagnostic modalities, *histopathological whole-slide images (WSIs)* stained with Hematoxylin and Eosin (H&E) remain the gold standard for subtype classification, offering rich morphological context. However, the application of computational techniques to analyze WSIs at scale continues to face significant challenges. These include the *absence of detailed, pixel-wise annotations*, *intra-class heterogeneity and inter-class ambiguity*, *staining variations*, and the *computational burden posed by the extremely high resolution of WSIs*, often resulting in inefficient, resource-intensive pipelines. Furthermore, the distribution of disease-relevant patterns within a slide is highly imbalanced, with diagnostically critical regions occupying only a small portion of the tissue.

This thesis addresses these pressing limitations by introducing **two novel, data-efficient, and weakly supervised learning frameworks**—**E-GloConNet** and **AttenEViT-HDMIL**—that collectively advance the state of lung cancer subtype classification from histopathology. These frameworks are designed to operate effectively without reliance on pixel-level annotations, thereby aligning with real-world clinical data availability. Through strategic sampling of representative patches and the use of both geometric and photometric augmentations, we improve the diversity of training data and reduce overfitting, enabling models to generalize across slides and staining conditions.

The first proposed framework, **E-GloConNet**, integrates lightweight convolutional architectures with a global context modeling strategy. It captures broader tissue-level semantics while maintaining efficiency and scalability. By selectively sampling patches and combining global visual priors with local detail, the method significantly reduces the training data requirement, achieving state-of-the-art performance with just a fraction of the data typically used in conventional methods. Building upon these ideas, the second framework, **AttenEViT-HDMIL**, introduces a more structured and biologically motivated approach. It identifies and prioritizes *high-cell-density regions*, which are more likely to contain malignant features, thereby concentrating computational effort on diagnostically salient areas. The architecture integrates a transformer encoder network with *hierarchical attention mechanisms* that capture multi-scale features and long-range dependencies across tissue sections.

This hierarchical structure, combined with a *multi-scale probabilistic feature fusion module*, allows for precise slide-level predictions while preserving contextual awareness and reducing redundancy.

Both frameworks are extensively evaluated on a curated cohort of **1,053 diagnostic WSIs** from *The Cancer Genome Atlas (TCGA).* The results demonstrate that our models consistently achieve high classification performance, with **AUCs exceeding 0.96**, while requiring only **7–10%** of the training data compared to traditional approaches. Importantly, the proposed methods exhibit robustness across variations in slide preparation, staining, and resolution, making them well-suited for deployment in diverse clinical settings. sFurthermore, the integration of *explainability mechanisms*, such as gradient-based visualizations, ensures that the models' decisions remain interpretable to pathologists, thereby fostering trust and facilitating adoption in clinical workflows. These visual maps highlight the cellular and structural features that contribute most significantly to model predictions, aligning well with established histological criteria.

In conclusion, this thesis presents a significant advancement in the field of computational pathology by developing scalable, annotation-efficient, and explainable deep learning methods tailored for histopathological image analysis. The proposed frameworks not only alleviate the dependency on large, meticulously annotated datasets but also offer a promising path toward practical, cost-effective, and clinically relevant AI tools for cancer diagnosis.

LIST OF PUBLICATIONS

1. **Aishwarya Priyadarshini**, Surya Prakash, Tadepalli Karuna, Deba Dulal Biswal, Sagar Khadanga, Debi Prasad Mishra "AttenEViT-HDMIL: Knowledge-guided Hierarchical Attention Transformer for Histopathological Lung Cancer Classification", IEEE Access, 2025. (communicated)

2. **Aishwarya Priyadarshini**, Surya Prakash and Tadepalli Karuna, "E-GloConNet: A Global Context-aware Efficient Network for Lung Cancer Classification in Whole Slide Images", International Journal of Software Science and Computational Intelligence, 2025. (communicated)

TABLE OF CONTENTS

List of Figures

List of Tables

Chapter 1

Introduction

Cancer remains one of the most formidable global health challenges of the 21st century. In 2022, an estimated 20 million new cancer cases were reported worldwide, and this number is projected to rise to over 35 million by 2050, representing a 77% increase driven by aging populations, urbanization, and increased exposure to carcinogenic risk factors such as tobacco, alcohol, obesity, and environmental pollution. Despite advancements in prevention, diagnosis, and therapy, cancer claimed approximately 9.7 million lives globally in 2022, underscoring its persistent burden on healthcare systems and societies [1].

Importantly, the impact of cancer is not uniformly distributed across the globe. Low- and middle-income countries (LMICs) bear a disproportionate share of the burden, accounting for approximately 70% of cancer deaths in 2020 [2]. This disparity is attributed to multiple factors, including limited access to early detection programs, inadequate healthcare infrastructure, and a shortage of specialized medical personnel. Additionally, the economic consequences are severe, often pushing families into financial hardship, particularly in regions where healthcare services are primarily financed through out-of-pocket spending. Among all cancers, lung cancer remains the most lethal, accounting for 1.8 million deaths globally in 2022, or 18.7% of all cancer-related deaths. It was also the most commonly diagnosed cancer in men and the second most common in women, with over 2.48 million new cases reported in the same year [3, 4]. The high mortality associated with lung cancer is largely attributed to late-stage diagnosis, its aggressive progression, and the continued prevalence of smoking, which remains the leading risk factor, responsible for approximately two-thirds of lung cancer deaths [3].

In the United States, projections for 2025 indicate over 2 million new cancer diagnoses and 618,120 deaths, with lung cancer expected to cause 124,730 deaths—more than breast, prostate, and colorectal cancers combined. While mortality rates have declined steadily due to improved detection and therapies, lung cancer continues to be the leading cause of cancer death in both men and women, accounting for roughly 1 in 5 cancer deaths in the country. Notably, lung adenocarcinoma, the most common histological subtype, has shown increasing incidence, particularly in non-smokers, women, and populations exposed to air pollution, especially in East Asia [1, 5]. Furthermore, there exist significant regional and socioeconomic disparities in cancer incidence and outcomes. While high-Human Development Index (HDI) countries are expected to experience the largest absolute increase in cancer cases, low- and middle-HDI regions are projected to see the greatest proportional increases, reflecting critical gaps in access to early detection, screening, and treatment services [2].

These alarming trends highlight the urgent need for innovative, scalable, and resource-efficient cancer detection and classification frameworks, particularly for lung cancer, where early and accurate diagnosis remains pivotal to improving patient outcomes. In this context, the application of advanced artificial intelligence (AI) and deep learning models to histopathological images holds immense promise for improving diagnostic accuracy, reducing pathologist workload, and addressing global disparities in cancer care.

1.1 Lung Cancer: prevalence, subtypes & clinical relevance

Lung cancer is recognized as the most frequently diagnosed cancer worldwide and remains the leading cause of cancer-related mortality, with around 1.8 million deaths reported globally in 2022. It accounts for roughly 12.4% of all new cancer diagnoses, with an estimated 2.5 million new cases each year [2, 4]. In the United States, lung cancer continues to be the deadliest cancer type, with projections indicating over 124,000 deaths in 2025 [6].

This disease is primarily classified into two main histological groups: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). NSCLC comprises approximately 85% of all cases, while SCLC represents about 15%, as represented by the lung cancer epidemiology in Fig. 1.1. Within NSCLC, adenocarcinoma, squamous cell carcinoma, and large cell

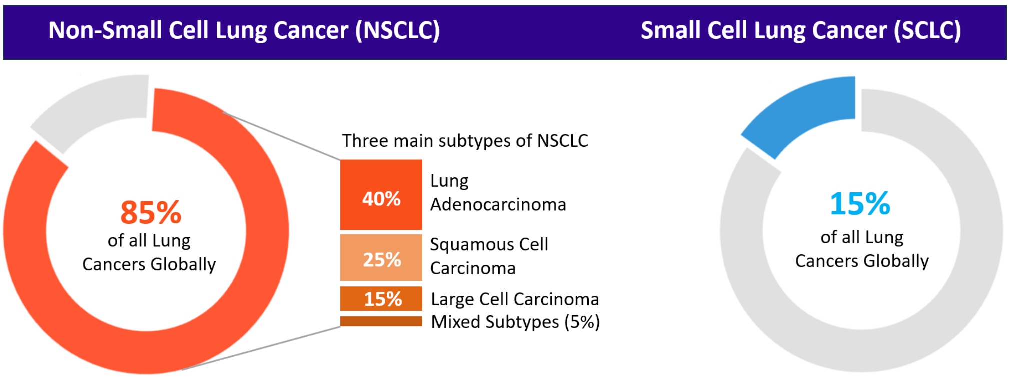


Figure 1.1: Global Lung Cancer Epidemiology

carcinoma are the predominant subtypes. Adenocarcinoma has become the most common form, especially among non-smokers and female patients, and its incidence continues to increase globally [7]. Although the prevalence of squamous cell carcinoma has declined in some populations, it remains a significant subtype, often linked to tobacco use. The differences in tumor biology and clinical behavior between these subtypes have important implications for treatment. NSCLC typically exhibits slower growth and metastasis compared to the highly aggressive and rapidly spreading SCLC. Treatment approaches differ accordingly; surgical removal combined with chemotherapy or targeted therapies is often effective for early-stage NSCLC, while advanced stages generally require systemic therapies, including immunotherapy [8]. In contrast, SCLC, which usually presents at an advanced stage, is mainly managed through chemotherapy and radiotherapy due to its aggressive clinical course.

Accurate diagnosis and subtype differentiation are crucial for determining the most appropriate therapeutic strategy. Molecular profiling plays an increasingly critical role, particularly in NSCLC, where identifying genetic mutations such as EGFR, ALK, and ROS1 informs targeted treatment options that improve patient outcomes [6]. In conclusion, the diverse nature of lung cancer underscores the importance of precise histological and molecular classification to tailor treatment effectively. Given its substantial global impact, ongoing advances in early detection, subtype-specific diagnosis, and personalized therapies are vital for improving survival rates and quality of life.

1.2 Diagnostic modalities for Lung Cancer detection

Multiple diagnostic techniques are employed to detect and evaluate lung cancer, including chest X-ray, computed tomography (CT), positron emission tomography (PET), and magnetic resonance imaging (MRI). These imaging tools help identify suspicious lesions, assess tumor size, and detect metastases. However, they are primarily effective for identifying malignancy rather than distinguishing between different lung cancer subtypes [9, 10].

Chest X-rays are often the initial screening tool due to their accessibility but have limited sensitivity, especially for early-stage tumors. CT scans provide more detailed anatomical information and are widely used for diagnosis and staging [9]. PET scans offer functional imaging to assess metabolic activity and identify areas of high tumor burden, while MRI is mainly used for detecting brain or spinal metastases or evaluating local invasion [11, 12]. While valuable for detecting and staging lung cancer, these imaging methods lack the resolution needed for subtype differentiation, which is critical for personalized treatment planning. This limitation arises because imaging focuses on macroscopic changes, whereas lung cancer subtypes—such as adenocarcinoma, squamous cell carcinoma, and small cell carcinoma—require cellular-level evaluation [9, 12].

Histopathological analysis, performed on biopsy or surgical specimens, remains the gold standard for subtype identification. Through microscopic examination and immunohistochemical staining, pathologists can classify tumors based on cellular morphology and marker expression. This is crucial because each subtype has distinct prognostic implications, molecular drivers, and therapeutic responses. For instance, adenocarcinomas may harbor EGFR mutations or ALK rearrangements, which can be targeted with specific drugs, whereas squamous cell carcinomas often do not respond to the same therapies. Furthermore, the emergence of digital pathology and computational tools has enhanced the diagnostic value of histopathology, enabling more accurate and scalable analysis using artificial intelligence. In summary, although imaging modalities are essential for detecting and staging lung cancer, histopathology remains irreplaceable for accurate subtype classification and treatment planning, making it the cornerstone of modern lung cancer diagnosis [13, 14].

1.3 Challenges in Manual WSI Analysis

Despite being the gold standard, manual examination of whole-slide images (WSIs) for lung cancer subtype classification poses several significant challenges. First, the sheer size and resolution of WSIs—often exceeding 10000×10000 pixels—make detailed inspection and pixel-wise annotation extremely laborious and unsustainable for routine clinical workflows. Identifying specific regions of interest across a slide at this scale becomes nearly infeasible.

Staining variability is another critical hurdle. H&E slides can exhibit considerable inconsistencies due to differences in staining protocols, reagents, and scanner calibrations, which profoundly affect both visual interpretation and digital analysis systems. Studies have reported inter-instrument color variation of up to 8% and inter-run variability reaching 23–28% over time [15]. Other work evaluating multi-site H&E samples demonstrated that laboratory-dependent color shifts cluster distinctly in PCA space, underscoring substantial inter-laboratory differences [16]. Such variability degrades diagnostic reliability and complicates efforts to standardize automated models. Pre-analytical processing variables—such as tissue handling, fixation duration, and slide sectioning—can introduce morphological artifacts, including shrinkage, crushing, and uneven staining [17]. These artifacts disrupt consistent evaluation and contribute to intra- and inter-observer variability, even among expert pathologists, particularly in borderline cases. Inter-pathologist disagreement is a well-documented issue in histopathology. For example, assessments of immunohistochemical staining intensity show significant variability when comparing different stainers or manual versus automated protocols [18]. This inconsistency in labeling compromises the quality of ground-truth data, and by extension, leads to ambiguities in machine learning training and validation.

Manual workflows also struggle with differentiating lung cancer subtypes that share subtle morphological characteristics. Overlapping features such as gland formation, nuclear atypia, and tumor-stroma interactions can lead to misclassification or delays, hampering accurate diagnosis and impacting treatment decisions. Additionally, quantitative measures such as mitotic count, nuclear-to-cytoplasm ratio, and cellular density are challenging to quantify manually and are rarely standardized, limiting their clinical utility [19].

These limitations underscore the urgent need for computationally driven approaches—such as stain normalization, automated segmentation, and attention-aware deep learning—to assist pathologists. Such tools can help mitigate manual burden, increase reproducibility, and improve diagnostic precision across diverse datasets and staining conditions.

1.4 Rise of Computational Pathology

The growing volume and complexity of whole-slide images (WSIs), coupled with the evident shortcomings of manual histopathological assessment, have ignited the emergence of computational pathology—a data-centric discipline integrating high-resolution imaging with advanced machine learning to support tissue interpretation and diagnostic decision-making. This paradigm is particularly transformative in lung cancer, where precise subtype classification and risk stratification are essential to tailor personalized treatment strategies [20, 21]. Computational pathology systems leverage automated high-throughput image analysis to quantify morphological features such as nuclear shape, glandular architecture, spatial patterns, and texture—attributes that are difficult to standardize through human inspection [21, 22]. By converting complex histological images into structured data representations, these systems establish an objective, reproducible foundation for diagnostic and prognostic modeling.

Deep learning—particularly convolutional neural networks (CNNs) and transformer-based architectures—has dramatically advanced the field. These models learn hierarchical tissue representations directly from raw image patches, bypassing the limitations of handcrafted features and achieving performance comparable to expert pathologists in tasks such as tumor detection, subtype differentiation, and grading [23, 24, 25]. They are capable of detecting nuanced morphological differences among lung adenocarcinoma, squamous cell carcinoma, and small-cell carcinoma—patterns often imperceptible to human observers. Moreover, computational pathology can address key technical challenges, such as staining variability, through data normalization and augmentation techniques, and reduce inter-observer variability by delivering consistent, algorithmically reproducible outputs [19]. Beyond diagnosis, AI-driven approaches are being used to infer molecular traits, predict patient outcomes, and integrate multi-modal data for a more holistic view of tumor biology [24, 26].

As the field matures, computational pathology is increasingly embedded within clinical workflows, offering potential benefits in diagnostic accuracy, workflow efficiency, and large-scale screening—particularly in settings where expert pathologists are scarce [21, 27]. Although integration hurdles remain—such as regulatory validation, interoperability, and explainability—growing evidence indicates that AI-enhanced pathology is on track to become a cornerstone of precision oncology.

1.5 Motivation and objectives

The growing burden of lung cancer and the critical role of histopathological subtype classification in guiding treatment have amplified the need for robust diagnostic support systems. As healthcare systems increasingly adopt digital pathology, the availability of large volumes of whole slide images (WSIs) offers unprecedented opportunities for data-driven insights. However, this shift also introduces new challenges in data acquisition, storage, and analysis, particularly due to the ultra-high resolution and complex nature of histopathological imagery.

The digitization of slides produces gigapixel-scale images that demand significant computational resources for storage and processing. Manual examination of these large images is not only time-consuming and labor-intensive, but also difficult to scale in high-throughput clinical environments. The heterogeneity in staining procedures, scanner settings, and sample preparation further complicates analysis, requiring systems that are robust to domain shifts and variability in image characteristics. Additionally, inter-observer variability remains a persistent issue in histopathological interpretation. Even among expert pathologists, diagnostic disagreements can occur, especially when differentiating between morphologically similar lung cancer subtypes. Such variability introduces ambiguity in clinical decision-making and undermines the consistency needed for training reliable AI models. These limitations strongly motivate the development of an automated system for the analysis of digitized histopathology imagery that is scalable, efficient, and capable of producing consistent results. The primary objectives of this work are: (i) to design a deep learning-based framework for automated subtype classification of lung cancer from WSIs, (ii) to ensure robustness against variability in staining and image acquisition, and (iii) to contribute toward reducing diagnostic subjectivity by offering reproducible and quantitative insights that augment pathologist workflows.

By addressing these challenges through a data-driven, AI-enabled approach, this work aims to support the evolution of computational pathology into a reliable adjunct for clinical diagnosis and personalized oncology.

1.6 Thesis Contributions

This thesis introduces two novel deep learning frameworks tailored for the automated classification of lung cancer subtypes from high-resolution histopathological WSIs. These contributions address key challenges in manual analysis, such as staining variability, annotation scarcity, computational efficiency, and the need for spatial contextual awareness. The primary contributions are summarized below.

1.6.1 Stain-Invariant and Foreground-Aware Learning: *E-GloConNet*

We propose **E-GloConNet**, a deep learning framework based on EfficientNetV2-B0 and enhanced with global context attention, designed to perform robust subtype classification under diverse staining and imaging conditions.

- **Stain-Invariant Learning Strategy:** A stain normalization technique is incorporated to mitigate staining variability across labs and scanners, enabling the model to generalize better across domains.

- **Annotation-Efficient Learning:** The model is trained using only slide-level labels, eliminating the dependency on region-level or pixel-level annotations and facilitating scalable learning.

- **Foreground-Aware Patch Filtering:** Low-tissue or background patches are excluded through a pre-filtering step, allowing the model to focus on diagnostically relevant tissue regions and reduce computational burden.

- **Global Context Attention Integration:** We augment EfficientNetV2-B0 with a Global Context Attention (GCA) module to capture broader spatial dependencies beyond local features, improving subtype discrimination performance.

1.6.2 Hierarchical Attention-Based MIL Framework: *AttenEViT-HDMIL*

We introduce **AttenEViT-HDMIL**, a hybrid convolutional-transformer architecture employing hierarchical attention and weakly supervised multi-instance learning (MIL) for enhanced WSI-level classification.

- **Dual-Stream Feature Extraction:** Combines convolutional layers for local morphological patterns with hierarchical transformers to capture global contextual features in WSIs.
- **Weakly Supervised MIL Framework:** Utilizes slide-level labels and patch-based analysis with MIL to learn effectively in the absence of detailed annotations.
- **Biologically Driven Patch Selection:** Applies intelligent tile sampling and ranking based on cell/nuclei density to retain only the most informative patches, improving both interpretability and efficiency.
- **Probabilistic Multi-Scale Feature Fusion:** Introduces a probabilistic fusion strategy to aggregate tile-level classification scores into a unified slide-level embedding, capturing comprehensive spatial and contextual insights.
- **Superior Classification Performance:** Demonstrates state-of-the-art accuracy and robustness in classifying lung cancer subtypes compared to existing benchmarks.

These contributions collectively establish a robust and scalable deep learning pipeline for the automated and accurate classification of lung cancer subtypes from histopathological slides. By addressing key limitations in manual pathology and conventional deep learning pipelines, this work advances computational pathology toward clinically meaningful applications.

1.7 Datasets

This study employs two prominent publicly available histopathology datasets for the classification of lung cancer subtypes, specifically LUAD and LUSC. These datasets were carefully selected based on their clinical relevance, image quality, and extensive annotation availability, enabling robust model development and validation. The details of the datasets, along with their sample counts, annotation and image resolution are presented in Table 1.1.

1.7.1 The Cancer Genome Atlas (TCGA) – Lung Cancer Cohort

The TCGA-Lung Cancer dataset is a comprehensive resource comprising WSIs of hematoxylin and eosin (H&E) stained formalin-fixed paraffin-embedded (FFPE) tissue sections. This cohort includes patients diagnosed with non-small cell lung cancer (NSCLC), predominantly LUAD [28] and LUSC subtypes [29]. The WSIs were generated using Aperio ScanScope XT and CS digital slide scanners, which provide high-resolution images at approximately 0.25 micrometers per pixel, corresponding to a magnification level between 20x and 40x. The dataset encompasses 541 WSIs from 585 LUAD patients and 512] WSIs from 504 LUSC patients, reflecting a diverse and clinically representative population.

TCGA WSIs capture detailed morphological and cellular structures, including tumor architecture, stromal patterns, and nuclear features essential for subtype differentiation. The availability of accompanying clinical and pathological metadata further enhances the dataset's value for supervised learning. This rich dataset serves as a foundation for training deep learning models to learn complex histological patterns critical for lung cancer subtype classification.

1.7.2 Clinical Proteomic Tumor Analysis Consortium (CPTAC) – Lung Cancer Cohort

The CPTAC dataset extends the histopathological resources available for lung cancer research by providing additional WSIs from a distinct patient cohort, with complementary clinical and molecular profiling data. The lung cancer subset includes LUAD and LUSC cases, with 674 WSIs from 374 LUAD patients and 662 WSIs from 363 LUSC patients. The tissue sections are stained with H&E and digitized using Leica Aperio AT2 scanners at a resolution comparable to TCGA, approximately 0.25 micrometers per pixel [30, 31].

Notably, CPTAC provides multiple WSIs per patient, which encompass different tumor regions or tissue blocks, offering a more heterogeneous sampling of tumor morphology.

This multiplicity improves model robustness by exposing the training process to diverse histological variations. Although CPTAC includes multi-omics data such as proteomics and genomics, this study focuses on leveraging the histopathological images and their subtype labels for model development.

Table 1.1 consolidates the critical statistics of the datasets used, including the number of patients, WSIs, and image resolution. These two datasets were selected for their complementary patient cohorts, high image quality, and detailed subtype annotations, forming a solid basis for training and evaluating deep learning models for lung cancer classification.

Table 1.1: Distribution of Lung Cancer histopathology datasets used in this study.

Dataset	**Subtype(s)**	**Patients**	**WSIs**	**Resolution**
TCGA-Lung Cancer	LUAD	585	541	∼0.25 µm/px
	LUSC	404	512	∼0.25 µm/px
CPTAC-Lung Cancer	LUAD	374	674	∼0.25 µm/px
	LSSC	363	662	∼0.25 µm/px

1.8 Evaluation Metrics

In this study, we employ several standard classification metrics to comprehensively assess the performance of our proposed models in distinguishing between the two major subtypes of NSCLC: LUAD and LUSC. These metrics offer insights into both the overall accuracy and the nuanced trade-offs between correctly and incorrectly predicted cases.

- **Accuracy:** This metric represents the proportion of correctly classified samples (both LUAD and LUSC) out of the total number of samples. While simple and intuitive, accuracy may not fully capture model performance when class distributions are imbalanced.

- **Precision:** Precision quantifies the proportion of true positive predictions among all positive predictions. For instance, in the case of LUAD, precision indicates how many of the slides predicted as LUAD are actually LUAD. High precision reduces the risk of false positives, which is critical when misdiagnosing one subtype as another could lead to inappropriate treatment.

- **Recall (Sensitivity):** Recall measures the proportion of actual positives correctly identified by the model. For LUAD, it reflects how many LUAD cases the model successfully detected. High recall ensures that most cancer cases of a particular subtype are caught, which is essential for timely intervention.

- **F1-Score:** The F1-score is the harmonic mean of precision and recall, providing a balanced metric that considers both false positives and false negatives. It is particularly useful in scenarios where an imbalance exists between classes or where both precision and recall are equally important.

- **Area Under the ROC Curve (AUC-ROC):** The AUC score evaluates the model's ability to distinguish between the two subtypes across various classification thresholds. AUC close to 1.0 signifies excellent separability between LUAD and LUSC, while a score of 0.5 implies no discriminatory power.

Together, these metrics offer a robust evaluation framework for assessing the subtype classification performance of our deep learning models on whole slide histopathology images.

1.9 Organization of the Thesis

This thesis is organized into five chapters to systematically address the challenges of lung cancer subtype classification using deep learning. The following chapter provides a comprehensive review of relevant literature in computational pathology and machine learning, establishing the foundation for the proposed approaches. Chapter 3 presents *E-GloConNet*, a deep learning framework leveraging global context attention to improve feature representation from WSIs for lung cancer subtype classification, along with its experimental evaluation. Chapter 4 introduces *AttenEViT-HDMIL*, an enhanced hierarchical attention-based model designed to efficiently process histopathological images within a weakly supervised learning framework, with its results and analyses detailed in the chapter. Chapter 5 concludes the thesis by discussing the interpretability of these models using visualization techniques such as Grad-CAM, summarizing key contributions and clinical implications, and outlining potential future research directions to further advance AI-driven lung cancer diagnosis.

Chapter 2

Literature Survey

The emergence of digital pathology has transformed traditional microscopy by enabling entire tissue specimens to be digitized at microscopic resolution. This shift toward digital pathology has laid the foundation for scalable computational analysis. Concurrently, computer-aided diagnosis (CAD) systems have emerged, blending image processing, machine learning, and artificial intelligence to assist pathologists in diagnostic workflows. These systems offer enhanced diagnostic reproducibility, mitigate inter-observer variability, and improve efficiency in the identification and classification of pathological structures [26].

Deep learning techniques, especially convolutional neural networks (CNNs) and Vision Transformers, have reached expert-level performance in various histopathological tasks, including cancer detection, grading, and subtype classification in high-incidence areas such as breast and lung cancer [32]. For example, using Inception-V3 on TCGA WSIs, Coudray et al. achieved an AUC of 0.97 in distinguishing LUAD from LUSC, validating their model across frozen, FFPE, and biopsy samples [26]. Similarly, Gao et al. utilized CNN-based systems to accurately distinguish lung adenocarcinoma and squamous cell carcinoma using TCGA and ICGC datasets, achieving AUCs ranging from 0.726 to 0.864 [33].

Despite these successes, considerable challenges remain—namely, data heterogeneity, high computational demands, the absence of standardized evaluation protocols, and limited interpretability of deep learning models—which continue to constrain clinical integration [34]. Yet, rapid advancements in algorithmic interpretability, efficient architectures, and cross-domain learning are progressively steering CAD systems toward clinical utility, offering a complementary role to expert pathologists in diagnostic decision-making.

2.1 Strongly Supervised Learning for Lung Cancer Subtype Classification

Strongly supervised learning has been pivotal in advancing histopathological image analysis, particularly for lung cancer subtype classification. As shown in Fig. 2.1(a), these approaches rely on explicit, manually provided annotations at the patch or pixel level, allowing models to learn detailed morphological patterns associated with distinct subtypes. In this setting, each image or image region is paired with a ground-truth label, often provided by expert pathologists.

Coudray et al. employed a strongly supervised approach using an Inception-v3 CNN trained on image patches extracted from WSIs in The Cancer Genome Atlas (TCGA), where each patch inherited the label of its parent slide—LUAD or LUSC [26]. This allowed the model to learn subtype-specific features and resulted in an area under the curve (AUC) of 0.97. The model's performance was validated on additional datasets, including frozen, FFPE, and biopsy samples, demonstrating robustness across specimen types. Similarly, Khosravi et al. fine-tuned pre-trained CNNs (Inception-v1 and Inception-v3) on fully labeled WSIs with known subtypes. Their approach achieved classification accuracies between 75–90% on lung cancer subtype classification tasks, showcasing the potential of transfer learning under strong supervision [35]. Wang et al. proposed a context-aware patch-based CNN approach, where fully annotated patches were used to train a convolutional model. Their method integrated Fully Convolutional Networks (FCNs) with a context-aware block selection strategy, achieving an AUC of 0.856 on the TCGA cohort [23]. This further highlights how contextual information can enhance subtype classification when strong supervision is available.

Despite their high accuracy, strongly supervised methods suffer from scalability challenges. Creating large-scale, high-quality annotations at the pixel or patch level is labor-intensive and requires domain expertise. Additionally, inter-observer variability and tissue heterogeneity can introduce biases and inconsistencies that affect the quality of supervision.

In summary, while strongly supervised learning has proven effective for subtype classification, the heavy reliance on detailed annotations and significant computational resources motivates the exploration of alternative paradigms—such as weakly supervised and self-supervised learning—that can reduce annotation burden while maintaining robust performance.

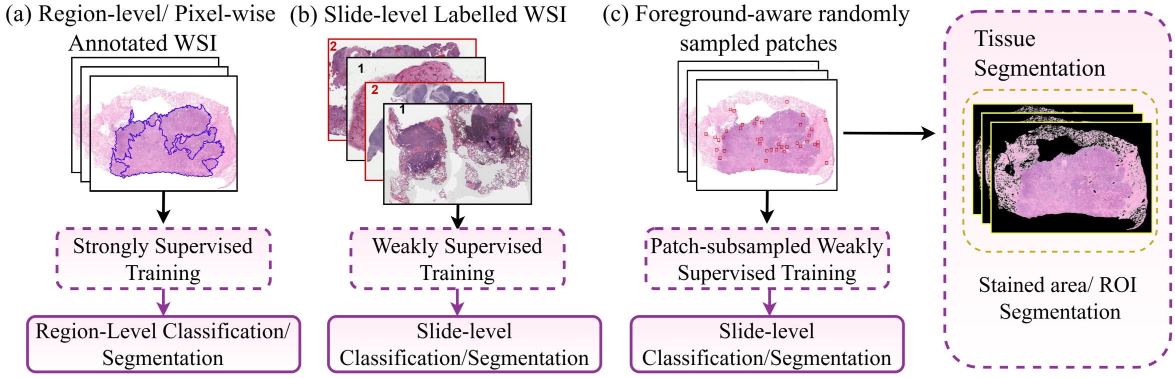

(a) Region-level/ Pixel-wise Annotated WSI
(b) Slide-level Labelled WSI
(c) Foreground-aware randomly sampled patches
2
1
2
1
Strongly Supervised Training
Weakly Supervised Training
Patch-subsampled Weakly Supervised Training
Region-Level Classification/ Segmentation
Slide-level Classification/Segmentation
Slide-level Classification/Segmentation
Tissue Segmentation
Stained area/ ROI Segmentation

Figure 2.1: Overview of supervision strategies in WSI analysis: (a) Strong supervision with pixel-wise annotations, (b) Weak supervision with slide-level labels, and (c) Patch-based weak supervision using foreground-aware sampling.

2.2 Weakly Supervised Learning Approaches

The labor-intensive nature of acquiring exhaustive pixel-level annotations in histopathology has catalyzed the development of weakly supervised learning methodologies. As shown in Fig. 2.1(b), these approaches aim to leverage slide-level labels to train models capable of accurate classification, thereby reducing the dependency on detailed annotations.

Multiple Instance Learning (MIL): Campanella et al. introduced a MIL-based deep learning system that utilizes only slide-level labels for training, circumventing the need for pixel-wise annotations. Their model was evaluated on a large dataset comprising 44,732 WSIs from 15,187 patients, achieving area under the curve (AUC) values above 0.98 across prostate cancer, basal cell carcinoma, and breast cancer metastases to axillary lymph nodes. This study demonstrated the feasibility of training accurate classification models at scale using weak supervision [36].

Patch-based Decision Fusion: Patch-based training (Fig. 2.1(c)) trained under weak supervision enable learning discriminative features from whole slide images using only slide-level labels, effectively bypassing the need for exhaustive pixel- or patch-level annotations. Hou et al. proposed a method that aggregates patch-level predictions using a weakly supervised decision fusion approach to address label scarcity in gigapixel images. While this method aimed to improve classification performance without exhaustive annotations, its precision remained suboptimal for clinical deployment, highlighting the challenges in balancing annotation efficiency and model accuracy [37].

Attention-Enhanced MIL Frameworks: Li et al. developed a dual-stream attention-based MIL model that decouples feature extraction and classification processes. By incorporating self-supervised contrastive learning, the model enhanced patch-level representations, leading to improved classification performance. However, the separation of feature extraction and classification limited joint optimization, indicating areas for further refinement [38]. Zhang et al. introduced DTFD-MIL, a double-tier feature distillation MIL framework that employs pseudo-bags and instance-level weighting to facilitate scalable learning. This approach achieved superior performance on datasets like CAMELYON-16 and TCGA lung cancer. Nevertheless, it faced challenges in capturing global contextual information and exhibited sensitivity to morphological variability [39].

End-to-End Learning Approaches: Cao et al. proposed E2EFP-MIL, an end-to-end MIL framework that achieved AUC values between 0.95 and 0.97. Despite its high performance, the model encountered computational overhead when training on full-resolution WSIs, posing scalability concerns [40]. Zhou et al. developed EWSLF, a framework that generates pseudo-labels using clustering and attention mechanisms to reduce annotation requirements. While this method aimed to alleviate the need for exhaustive annotations, it still relied on extensive patch-level inference, indicating a trade-off between annotation efficiency and computational demands [41].

In summary, WSL methods in histopathology offer a promising avenue to mitigate the challenges associated with exhaustive annotations. However, they continue to grapple with issues such as redundancy and noise from irrelevant patches, limited generalizability due to tissue heterogeneity, and substantial computational requirements. Ongoing research endeavors aim to address these challenges to enhance the clinical applicability of WSL approaches in histopathological analysis.

2.3 Deep Neural Networks in Histopathological Analysis

Deep learning has revolutionized the analysis of histopathological WSI, with its architectures serving as the fundamental building blocks for both diagnostic and prognostic modeling

[34]. Given the ultra-high resolution and complex tissue structures present in WSIs, carefully designed neural architectures are essential to effectively capture both local cellular features and global contextual information.

Convolutional Neural Networks (CNNs): CNNs have been the predominant architecture employed in histopathology, especially for patch-based analysis [42, 23]. Their ability to extract spatially localized features through convolutional filters makes them well-suited for recognizing tissue patterns, cellular morphology, and textural cues. The hierarchical feature maps produced by CNNs allow for progressively abstract representations, which can be leveraged to distinguish subtle differences between cancer subtypes or grades [43]. Moreover, their relative robustness to noise and variability in staining makes them highly practical across datasets acquired from different institutions. CNN-based models have been adopted in both fully supervised and weakly supervised paradigms, often serving as the backbone in complex multi-stage pipelines [37].

Context-Aware and Hierarchical Architectures: While CNNs excel at capturing localized features, histopathological diagnosis frequently requires understanding of broader spatial relationships—such as tissue organization and tumor boundaries—that extend beyond the receptive field of standard convolutional layers. To address this, researchers have proposed architectures that incorporate context-aware modules, multi-scale processing, and hierarchical attention mechanisms [44, 45, 46]. These enhancements enable the model to integrate information across different spatial resolutions and anatomical contexts, thereby improving its interpretability and diagnostic performance. Feature fusion strategies—such as combining low-level morphological details with high-level semantic cues—are also employed to reinforce learning across scales. Notably, models such as U-Net, ResNet with spatial pyramid pooling, and attention-based fusion networks have demonstrated improved accuracy in segmentation and classification tasks [47, 46].

However, these advanced architectures often come with increased computational demands. High-resolution WSIs, which may contain billions of pixels, pose memory and runtime constraints that necessitate the use of powerful GPUs and optimized inference strategies [36]. As a result, a trade-off must be considered between architectural complexity and practical deployment, especially in resource-constrained clinical settings.

In summary, the design of deep neural networks for histopathology is a careful balancing act between extracting detailed local features and capturing long-range dependencies. The evolution from standard CNNs to context-aware and hierarchical models reflects a growing understanding of the structural complexity inherent in pathological slides and underscores the need for computational efficiency alongside model sophistication.

2.4 Transformer-Based Models in Histopathology

The integration of Transformer-based architectures into histopathological image analysis has marked a significant advancement in computational pathology. Unlike traditional CNNs, Transformers utilize self-attention mechanisms to capture long-range dependencies and model global context, which are particularly beneficial for analyzing gigapixel WSIs [48, 49]. This capability allows for a more holistic understanding of tissue architecture and cellular interactions, essential for accurate cancer diagnosis and prognosis.

TransMIL: Correlated Multiple Instance Learning with Transformers: Shao et al. introduced TransMIL, a Transformer-based Multiple Instance Learning (MIL) framework designed to address the limitations of traditional MIL approaches that often assume independent and identically distributed instances [50]. TransMIL incorporates a Pyramid Position Encoding Generator (PPEG) to embed spatial information and leverages Transformer layers to model the correlations among instances within a WSI. This approach achieved notable performance across various cancer datasets, including breast, lung, and kidney cancers, demonstrating the model's versatility and effectiveness in capturing complex histological patterns.

Single-Cell Heterogeneity-Aware Transformer Models: Yu et al. developed a Transformer-guided framework that accounts for single-cell heterogeneity to predict aneuploidy from WSIs [51]. The model performs nuclei segmentation and classification to identify individual cancer cells, which are then clustered into subtypes. By computing the distribution of these subtypes and extracting morphological features, the model captures the heterogeneity within the tumor microenvironment. This approach achieved promising results, with Area Under the Curve (AUC) scores of 0.818 and 0.827 on lung adenocarcinoma (LUAD) and head and neck squamous cell carcinoma (HNSC) test sets, respectively.

DSCA: Dual-Stream Cross-Attention Networks for Multi-Resolution Analysis: Liu et al. proposed the Dual-Stream Cross-Attention (DSCA) network to efficiently exploit WSI pyramids for cancer prognosis [38]. The DSCA model processes WSI patches at two different resolutions through separate streams and employs a cross-attention mechanism to fuse the multi-scale features effectively. This design addresses the semantic gap in multi-resolution feature fusion and reduces computational costs. The DSCA model demonstrated superior performance compared to existing methods, with an average improvement of around 4.6% in the concordance index (C-Index) across multiple datasets.

Despite their advantages, ViTs face computational challenges due to the quadratic complexity ($\mathcal{O}(n^2)$) of the self-attention mechanism. Efficient sampling strategies have been proposed [40] to reduce computational load but may still introduce redundancy or bias. These limitations underscore the need for novel transformer architectures that can handle high-resolution inputs efficiently while preserving global and local morphological details for robust lung cancer subtype classification.

Chapter 3

E-GloConNet: Global Context-aware Network for WSI Analysis

This chapter presents the proposed methodology, E-GloConNet, developed to tackle the critical challenges associated with lung cancer subtype classification using histopathological WSIs. Given the vast size and complexity of WSIs, along with the limited availability of fine-grained annotations, our framework adopts a weakly supervised and data-efficient approach.

The chapter outlines each stage of the proposed pipeline, beginning with a selective sampling strategy designed to capture representative morphological features while minimizing redundancy. It further describes the integration of geometric and photometric augmentations aimed at improving robustness and generalizability. Central to our method is the fusion of efficient feature extraction with a global context modeling module, enabling the system to retain both fine-grained and holistic information from high-resolution tissue samples. Finally, we detail the architectural components, training strategy, and evaluation metrics used to validate the effectiveness of the approach on large-scale TCGA-lung cancer and CPTAC-Lung cancer datasets. Through this chapter, we demonstrate how E-GloConNet addresses limitations of existing methods while achieving strong performance with minimal annotation cost.

3.0.1 WSI Pre-processing

The preprocessing stage is a critical component of our pipeline, aiming to efficiently isolate diagnostically informative regions from whole-slide images (WSIs) stained with hematoxylin and eosin (H&E). Given a WSI $\mathcal{S} \in \mathbb{R}^{H \times W \times 3}$, we begin by applying *Stained Region Extraction (SRE)*, a tissue segmentation strategy that focuses on separating foreground tissue from non-informative background areas. To achieve this, the RGB image is first converted into a grayscale representation, reducing the complexity of the color space while preserving spatial structure. Subsequently, *Otsu's adaptive thresholding method* [52] is employed to compute a global threshold that segments the grayscale image into binary form. This produces a binary tissue mask $M \in \{0,1\}^{H \times W}$, where pixels corresponding to tissue are labeled as 1 (foreground) and the background as 0. This binary mask is then used to filter the original WSI, generating a masked image $\mathcal{S}_{\text{tissue}}$ that retains only the tissue-containing regions in color, effectively removing blank or irrelevant slide portions.

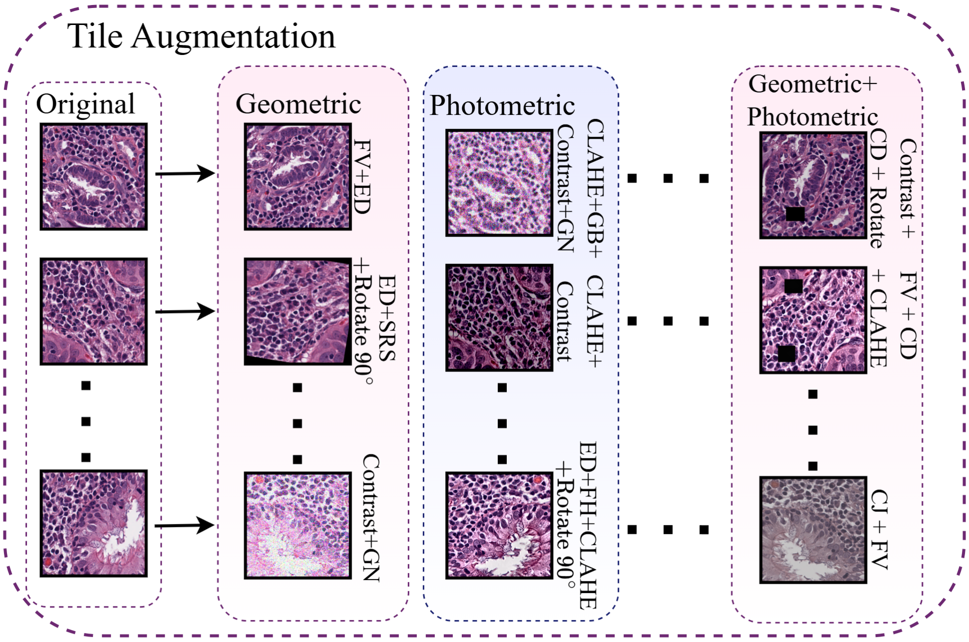
From this processed image, we extract a collection of non-overlapping square tiles, each of size 224 × 224 pixels, resulting in a patch set $\mathcal{T} = \{t_i\}_{i=1}^{N}$ sampled at a resolution of 40× magnification. To enhance the relevance of selected tiles, a *foreground-aware filtering mechanism* is applied. Only patches with a tissue coverage of greater than 80% (as determined by the binary mask) are retained, discarding regions that are either artifact-prone or dominated by whitespace. To further mitigate variability introduced by inconsistent staining procedures across different samples and institutions, we employ *Vahadane's stain normalization technique.* This method transforms each selected patch to a reference stain template while preserving underlying tissue morphology. By standardizing stain appearance, this step ensures greater consistency in color representation, facilitating robust and stain-invariant model training across diverse WSIs [53].

3.0.2 Random Sampling and Patch Augmentation

Given the enormous number of image patches generated from each WSI, training deep neural networks using all available patches is computationally infeasible. To manage this, we adopt a uniform random sampling strategy to maintain computational tractability while preserving morphological diversity. From the foreground-filtered patch set $\mathcal{T}' = \{t'_i\}$, we randomly select a fixed number of patches per slide:

$$\mathcal{T}_{\text{random}} = \{t'_j\}_{j=1}^{k}, \quad k = 1000 \tag{3.1}$$

where k is the number of randomly selected patches per WSI. This sampling approach allows us to retain a broad representation of intra-slide structural variations while keeping the training process scalable and memory-efficient. To enhance the model's ability to generalize across diverse histological appearances, we apply a comprehensive set of data augmentations exclusively during training. These augmentations are biologically inspired and aim to replicate common artifacts and variations encountered in real-world pathology workflows. Each patch in $\mathcal{T}_{\text{random}}$ is independently subjected to a combination of geometric and photometric transformations as presented in Fig. 3.1.



* FV = Flip Vertical, ED = Elastic Deformations, GN = Gaussian Noise,
GB = Gaussian Blur, FH = Flip Horizontal, CD = Coarse Dropout, CJ = Color Jitter

Figure 3.1: Illustration of geometric and photometric tile augmentation techniques applied to histopathological tiles from WSIs.

Geometric Augmentations: Orientation invariance is introduced through random horizontal and vertical flips, each with a probability of 0.7, and random 90° rotations with a probability of 0.5 to mimic variability due to slide scanning orientations. Additionally, affine transformations—including random translations (up to ±5%), scalings (up to ±10%), and in-plane rotations (up to ±15°)—are applied to 40% of the patches to simulate tissue deformation during slide preparation. To introduce realistic non-linear distortions, elastic deformations are applied to 30% of patches, emulating structural changes caused by tissue stretching or sectioning. Coarse dropout is also employed with a probability of 0.3, randomly masking rectangular regions to simulate occlusions caused by artifacts such as air bubbles, dust, or tissue folds.

Photometric Augmentations: To address staining variability and scanner-specific color shifts, color jitter is applied to 60% of the patches, altering brightness (±10%), contrast (±15%), saturation (±10%), and hue (±5°). Additionally, CLAHE (Contrast Limited Adaptive Histogram Equalization) is used in 40% of the patches to enhance local contrast and bring out fine-grained structures like nuclei and cellular boundaries. To simulate imaging noise and minor focus errors, we introduce Gaussian noise in 30% of the patches and apply Gaussian blur with a small kernel to 20% of the samples.

Each augmentation is applied independently using a stochastic augmentation pipeline. This *online augmentation* framework ensures that the same patch can appear in various forms across different epochs, resulting in a dynamically evolving training distribution. Unlike traditional augmentation strategies that generate fixed augmented copies, our approach produces a combinatorially rich training dataset on-the-fly, which significantly improves the model's robustness and its ability to learn discriminative features in weakly supervised settings.

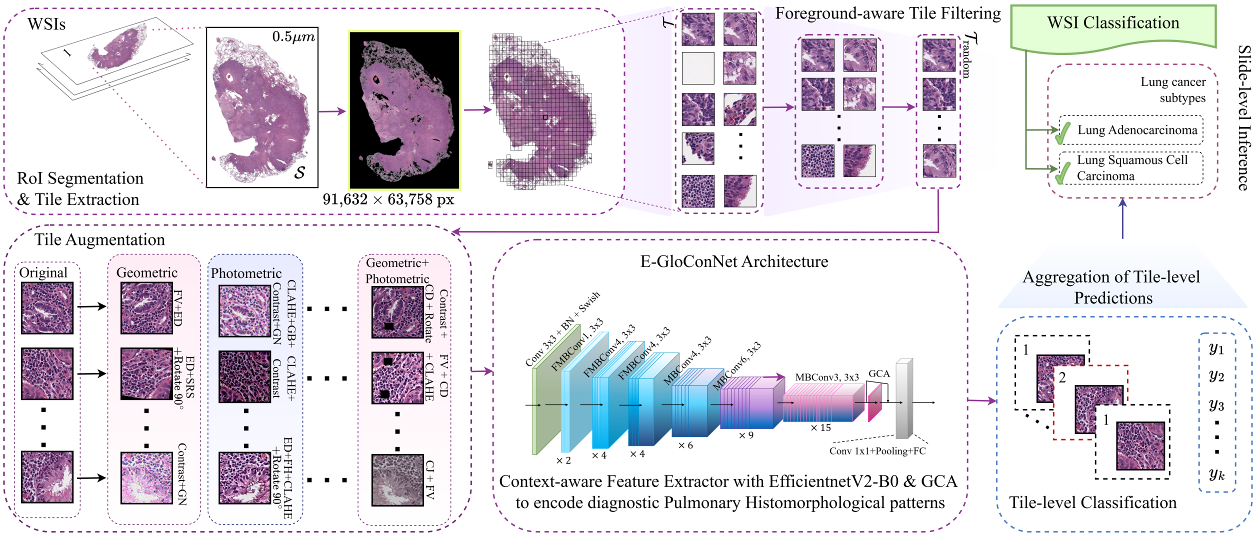
3.1 The Proposed E-GloConNet

Deep convolutional neural networks have become the cornerstone of modern computational pathology, enabling automated analysis of histopathological images with high accuracy. Efficient feature extraction and the ability to capture both local and global contextual information are critical for successful tissue classification. In this work, we build upon the EfficientNetV2-B0 architecture—a state-of-the-art convolutional baseline known for its balance of efficiency and accuracy—and propose novel enhancements tailored for histopathological image analysis. We use E-GloConNet to address the research gaps identified by our study. The important components of the E-GloConNet module is presented in Fig. 3.2.

3.1.1 EfficientNetV2-B0: A Background

EfficientNetV2-B0 is part of the EfficientNetV2 family [54], which builds upon the original EfficientNet series by incorporating faster training, better parameter efficiency, and enhanced accuracy. The backbone design leverages two key components: Mobile Inverted Bottleneck Convolution (MBConv) blocks and Fused-MBConv (FMBConv) blocks. MBConv blocks include depthwise separable convolutions and Squeeze-and-Excitation (SE) attention mechanisms, which help recalibrate feature channels based on global context. The SE module compresses each feature channel via global average pooling, applies a lightweight two-layer MLP to model inter-channel dependencies, and scales the input feature map accordingly. On the other hand, FMBConv blocks fuse the expansion and depthwise convolutional layers into a single standard convolution for improved hardware efficiency, especially during early-stage processing. EfficientNetV2-B0 also employs compound scaling, which uniformly scales network depth, width, and input resolution using a set of optimized coefficients. This makes it particularly suitable for computationally constrained environments while preserving high accuracy, making it a compelling choice for medical imaging tasks such as histopathology. While EfficientNetV2-B0 serves as a strong baseline for feature extraction, it has limitations in modeling spatial relationships—especially in the later layers where global tissue architecture and spatial patterns play a critical role in histopathological classification. The Squeeze-and-Excitation blocks integrated into MBConv layers focus solely on channel recalibration and operate independently of spatial location. This becomes a bottleneck when distinguishing between subtle spatial features like glandular formations, keratin pearls, and tumor-stroma boundaries.

To address this, we propose **E-GloConNet**, a modified architecture that augments EfficientNetV2-B0 with a **Global Context Attention (GCA)** module. Our model processes input patches $\mathbf{x} \in \mathbb{R}^{3\times224\times224}$ through a series of MBConv and FMBConv blocks to extract hierarchical features. We then introduce the GCA module at the final stage of the feature extractor, just before the classification head. The GCA block captures both spatial and channel-wise dependencies in a unified framework. First, it generates a spatial attention map by aggregating contextual information across all spatial positions, effectively enabling long-range dependency modeling. Then, a channel transformation path uses a lightweight MLP to recalibrate features across channels. This dual-path design allows GCA to enhance



* FV = Flip Vertical, ED = Elastic Deformations, GN = Gaussian Noise,
GB = Gaussian Blur, FH = Flip Horizontal, CD = Coarse Dropout, CJ = Color Jitter

Figure 3.2: Schematic overview of the proposed E-GloConNet framework. The preprocessing stage includes RoI segmentation from WSIs, patch extraction, quality filtering, random sampling, and data augmentation. The resulting patches are used to train the E-GloConNet model. During inference, patch-level predictions are aggregated via majority voting to produce the final slide-level classification.

relevant spatial regions while preserving the discriminative power of channel attention. By integrating GCA, E-GloConNet gains the ability to model both local details and global context, which are essential for interpreting high-resolution histopathological images. This improves both classification performance and model interpretability, particularly in complex diagnostic cases. Figure 3.3 illustrates the overall architecture of the proposed model.

3.1.1.1 Global Context Attention (GCA)

As depicted in Fig. 3.3, the GCA module is composed of two synergistic branches: a context modeling path that captures spatial dependencies, and a transformation path that refines the contextual representation using lightweight operations. This module operates on the final feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$, where C, H, and W represent the number of channels, height, and width, respectively. The first branch generates a spatial attention map $\mathbf{A} \in \mathbb{R}^{H\times W}$ to quantify the relative importance of each spatial location. This is achieved via a 1×1 convolution followed by a softmax operation applied across all spatial positions

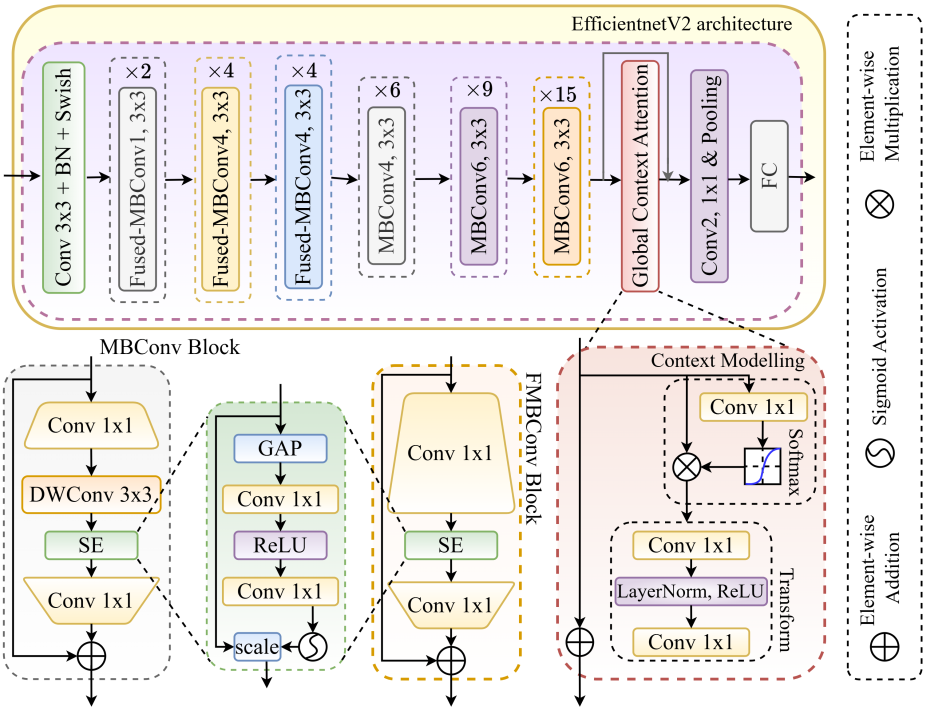


Figure 3.3: Overview of the proposed E-GloConNet framework, showcasing layer-wise EfficientNetV2 baseline integrated with the GCA module prior to the classification head. The figure details key components, including the MBConv block, SE block, Fused-MBConv block, and the GCA module.

$(p, q) \in [1, H] \times [1, W]$:

$$\mathbf{A}(i,j) = \frac{\exp\left(\mathbf{W}_a * \mathbf{F}(i,j)\right)}{\sum_{p,q} \exp\left(\mathbf{W}_a * \mathbf{F}(p,q)\right)} \tag{3.2}$$

where $\mathbf{W}_a$ denotes the learnable convolution kernel, and (i, j) indicates a spatial coordinate in the feature map. This attention map assigns normalized significance to each spatial location. Using the spatial attention map, a global context vector $\mathbf{g} \in \mathbb{R}^C$ is computed by aggregating feature responses across spatial dimensions in a weighted manner:

$$\mathbf{g}(c) = \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{A}(i,j) \cdot \mathbf{F}(c,i,j) \tag{3.3}$$

This vector $\mathbf{g}$ captures a global semantic summary of the feature map. It is then broadcast and fused back into the original feature map via element-wise multiplication to generate a context-aware feature map $\tilde{\mathbf{F}}$. To further enrich these recalibrated features, the transformation branch processes $\tilde{\mathbf{F}}$ through a lightweight stack consisting of two 1×1 convolutions (Conv_1 and Conv_2), interleaved with Layer Normalization (LN) and a ReLU activation function:

$$\tilde{\mathbf{F}}' = \mathrm{Conv}_2\left(\mathrm{ReLU}\left(\mathrm{LN}\left(\mathrm{Conv}_1\left(\tilde{\mathbf{F}}\right)\right)\right)\right) \tag{3.4}$$

The output $\tilde{\mathbf{F}}'$ thus contains spatially and channel-wise enhanced representations, which are subsequently forwarded to the classification head. This dual-branch mechanism enables the network to capture long-range dependencies and integrate high-level semantic context, both of which are crucial for accurate discrimination of complex histopathological patterns.

3.1.2 Patch-Level Classification and Slide-Level Aggregation

Once the features are extracted via the proposed E-GloConNet, each image patch $x_i \in T_{\text{random}}(S)$ is processed independently to determine its histological subtype. For this purpose, the recalibrated and context-enriched feature map $\tilde{\mathbf{F}}'_i$ corresponding to patch x_i is subjected to global spatial pooling to yield a compact one-dimensional feature vector $\mathbf{f}_i \in \mathbb{R}^d$. This vector encapsulates the most salient semantic information of the patch and is passed through a fully connected classification head, followed by a softmax activation to compute class probabilities: $\mathbf{p}_i = (p_{i,1}, p_{i,2})$. The final predicted label for each patch is determined by selecting the class with the highest predicted probability:

$$\hat{y}_i = \arg\max_{c \in \{1,2\}} p_{i,c} \tag{3.5}$$

During training, the model parameters are optimized by minimizing the categorical cross-entropy loss between the predicted probabilities and the ground truth labels for each patch. This enables the network to learn discriminative features that differentiate LUAD from LUSC. For slide-level inference, individual patch predictions $\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_k\}$ associated with a WSI S are aggregated using a majority voting strategy. The final slide-level label $\hat{Y}_S$ corresponds to the most frequently occurring class among the patch predictions:

$$\hat{Y}_S = \mathrm{MajorityVoting}\left(\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_k\}\right) \tag{3.6}$$

In the event of a tie, a random selection mechanism is used to resolve class ambiguity. To further tailor the model to histopathological data, a multi-stage fine-tuning protocol was adopted. Initially, early layers of the network were frozen to retain generic visual representations acquired from ImageNet pretraining [55]. Subsequently, deeper layers were progressively unfrozen and fine-tuned, allowing the model to adapt to the domain-specific characteristics of LUAD and LUSC tissue morphology.

3.2 Experimental Evaluation and Results

In this section, we systematically evaluate the effectiveness of our proposed E-GloConNet framework for lung cancer subtype classification. A series of comprehensive experiments were conducted to validate the model's performance at both the patch and slide levels. We begin by outlining the experimental setup, including dataset details, pre-processing strategies, and training protocols. This is followed by a quantitative comparison with existing baselines and state-of-the-art (SOTA) models to highlight the strengths of our approach. Furthermore, we perform detailed ablation studies to isolate the contributions of key architectural components such as GCA and hierarchical fine-tuning. Together, these experiments demonstrate the robustness and adaptability of our model across various evaluation settings.

3.2.1 Implementation Details and Experimental Setup

This section outlines the procedures followed during model development, including dataset partitioning, training strategy, architectural design, and experimental configuration.

3.2.1.1 Dataset Partitioning and Cross-Validation Strategy

We begin by detailing how the TCGA dataset was partitioned to ensure a fair and unbiased evaluation. To ensure a fair and robust evaluation, we adopt a two-stage evaluation strategy using the TCGA dataset. First, we perform a **90:10 train-test split** followed by **5-fold cross-validation** on the training set to rigorously evaluate generalization and model stability. In each fold, 80% of the data is used for training and 20% for validation. Since only slide-level labels are available, we aggregate patch-level predictions via average pooling to generate the final slide-level classification scores. The detailed distribution of slides used in our experiments is shown in Table 3.1. Each fold maintains a consistent balance between LUAD and LUSC samples, allowing for reliable comparative evaluation across folds.

3.2.1.2 Transfer Learning on CPTAC Dataset

To assess the generalizability of the model beyond the TCGA dataset, we evaluate it on an independent cohort. We employ a transfer learning setup where the model trained on the TCGA dataset is directly tested on the CPTAC cohort without fine-tuning. This simulates a real-world deployment scenario where annotated data in the target domain may

Dataset	#WSIs	Type of Split	No. of Samples
TCGA-Lung Cancer	1053	Train-Test (9:1)	Train: 948 WSIs Test: 105 WSIs
		5-Fold Cross-Validation	Fold 1: Train(757 WSIs), Validation(191 WSIs) Fold 2: Train(756 WSIs), Validation(192 WSIs) Fold 3: Train(757 WSIs), Validation(191 WSIs) Fold 4: Train(758 WSIs), Validation(190 WSIs) Fold 5: Train(757 WSIs), Validation(191 WSIs)

Table 3.1. TCGA-Lung cancer dataset partitioning specifications.

be scarce. By leveraging pretrained knowledge from TCGA, we reduce the need for extensive re-training, minimizing computational and annotation costs. This transfer learning pipeline highlights the robustness of our model across inter-cohort variations in histology and staining protocols. The training, testing and validation sample count for TCGA-lung cancer dataset has been tabulated and presented in Table 3.1.

3.2.1.3 Model Architecture and Training Procedure

We now describe the architectural components and training configuration of the proposed E-GloConNet framework. We utilize EfficientNetV2-B0 as the base baseline, selected for its computational efficiency and strong baseline performance on visual tasks. To enhance global contextual awareness, we integrate GCA modules into key stages of the network, forming our proposed architecture, **E-GloConNet**. The GCA module enables the model to capture long-range spatial dependencies and semantic structure, which are critical for detecting complex histological features such as keratin pearls, glandular morphologies, and tumor-stroma transitions. A lightweight classification head comprising fully connected layers, GeLU activations, dropout, and layer normalization processes the final representation. We adopt a hybrid fine-tuning approach, freezing the early layers of EfficientNetV2-B0 while fine-tuning the last two MBConv blocks and the classification head. The model is trained for **150 epochs** using early stopping to prevent overfitting. We use an adaptive learning rate scheduler with `AdamW` optimizer. The batch size and initial learning rate are optimized using a grid search over the validation set. The remaining details are summarized in Table 3.2.

3.2.1.4 Hardware and Software Configuration

Finally, we summarize the system configuration used for training and experimentation. All experiments are conducted using **NVIDIA V100 GPUs** with 32 GB of VRAM. The training and inference pipelines are implemented using **PyTorch 2.4.0** and CUDA 12.4. WSIs are read and processed using the **OpenSlide v3.4.1** library. Preprocessing includes stain normalization and extraction of non-overlapping patches from high-resolution WSIs at 20× magnification. To ensure reproducibility, all experiments are seeded, and configuration files are version-controlled.

Table 3.2: Hyperparameter summary for the proposed E-GloConNet framework

a. Model Architecture Parameters	
Backbone Network	EfficientNetV2-B0 (ImageNet pretrained)
Fine-tuned Layers	Last 2 MBConv stages + classifier head
Global Context Attention (GCA)	1×1 Conv, Softmax attention, 2-layer MLP
Feature vector dimension (d)	512
Classifier Head	FC → Dropout(0.3) → GeLU → LayerNorm
Normalization Epsilon (ϵ)	1×10^{-5}
Patch-Level Aggregation	Top-κ confidence-based averaging
Final Prediction Layer	Softmax over 2 classes
b. Optimization and Training Hyperparameters	
Batch Size	50
Epochs	150
Loss Function	Categorical Cross-Entropy
Optimizer	Adam
Initial Learning Rate	1×10^{-4}
Weight Decay	4×10^{-5}
Momentum	0.9
Adam β_1, β_2	0.9, 0.999
Adam ϵ	1×10^{-8}
Learning Rate Scheduler	ReduceLROnPlateau
Scheduler Patience	20 epochs
Early Stopping Patience	10 epochs
Cross-Validation Protocol	5-fold (80/10/10 split)
Hardware	NVIDIA V100 (32 GB), CUDA 12.4
Software Stack	PyTorch 2.4.0, Python 3.12.3, OpenSlide 3.4.1

3.2.2 Comparative Analysis of Baseline Methods

A thorough evaluation of nine baseline deep learning architectures was conducted across four benchmark lung cancer histopathology datasets: TCGA-Lung Cancer and CPTAC-Lung Cancer, across key evaluation metrics. The models varied in complexity, ranging from lightweight CNNs such as SqueezeNet [56] and ShuffleNetV2 [57], to more sophisticated architectures like EfficientNetV2B0 and Vision Transformer (ViT) [48]. The results comparing these baselines are presented in Table 3.3.

1. TCGA-Lung Cancer Dataset Among lightweight models, ShuffleNetV2 outperformed SqueezeNet in terms of accuracy (80.45% vs. 78.24%) and recall (0.845 vs. 0.781), but the latter achieved slightly better precision, suggesting stronger specificity. Classical architectures like GoogleNet [58] and InceptionV3 [59] showed steady improvements across all metrics, with InceptionV3 achieving 83.90% accuracy and an F1-score of 0.859. The ViT model introduced transformer-based capabilities, yielding a notable jump in recall (0.894) and F1-score (0.876), indicating better sensitivity to malignant cases. Among CNN variants, DenseNet121 [60], MobileNetV2 [61], and EfficientNetV2B0 exhibited progressively superior performance, culminating in 91.85% accuracy and a 0.912 F1-score for EfficientNetV2B0.

2. CPTAC-Lung Cancer Dataset A similar trend was observed on the CPTAC dataset, albeit with slightly improved results across all models, likely due to higher image consistency and label granularity. MobileNetV2 and EfficientNetV2B0 emerged as strong performers, achieving F1-scores of 0.919 and 0.933, respectively. Transformer-based ViT outperformed traditional CNNs like GoogleNet and InceptionV3 in terms of AUC and F1-score, affirming its capability to capture global spatial patterns in histological features. Notably, ShuffleNetV2, while lightweight, still maintained respectable performance, underlining its viability for resource-constrained environments.

A comprehensive comparison of nine baseline models across four lung cancer histopathology datasets revealed consistent performance trends. Lightweight models like SqueezeNet and ShuffleNetV2 performed reasonably, but lagged in recall and F1-score compared to deeper architectures. Classical CNNs like GoogleNet and InceptionV3 showed moderate improvements, while transformer-based ViT consistently achieved higher recall and F1-scores, reflecting its strength in capturing complex spatial features. Modern CNNs such as MobileNetV2 and EfficientNetV2B0 outperformed all others across datasets, with Efficient-

Table 3.3: Comparative analysis of E-GloConNet with baseline methods on four benchmark lung cancer histopathology datasets.

Baseline	**TCGA-Lung Cancer**					**CPTAC-Lung Cancer**				
	AUC	Acc (%)	Prec.	Rec.	F1	AUC	Acc (%)	Prec.	Rec.	F1
SqueezeNet	0.822	78.24	0.824	0.781	0.802	0.844	80.10	0.839	0.817	0.828
ShuffleNetV2	0.836	80.45	0.812	0.845	0.828	0.859	82.17	0.820	0.847	0.833
GoogleNet	0.851	82.20	0.825	0.861	0.842	0.869	84.75	0.854	0.873	0.863
InceptionV3	0.868	83.90	0.848	0.871	0.859	0.881	86.48	0.863	0.869	0.866
ViT	0.872	85.41	0.859	0.894	0.876	0.893	88.26	0.871	0.903	0.887
ResNet50V2	0.889	86.70	0.890	0.878	0.884	0.911	89.85	0.897	0.889	0.893
DenseNet121	0.898	88.13	0.878	0.902	0.890	0.920	91.02	0.905	0.918	0.911
MobileNetV2	0.905	89.52	0.870	0.915	0.892	0.930	92.25	0.900	0.939	0.919
EfficientNetV2B0	0.933	91.85	0.899	0.926	0.912	0.949	93.73	0.925	0.940	0.933
E-GloConNet	**0.965**	**95.01**	**0.949**	**0.954**	**0.952**	**0.978**	**96.84**	**0.966**	**0.974**	**0.970**

NetV2B0 achieving near-saturation in metrics, particularly on curated datasets. Overall, the results suggest that deeper networks and attention mechanisms significantly enhance classification performance in histopathological image analysis.

3.2.3 Comparison with state-of-the-Art (SOTA) methods

The performance comparison presented in Table 3.4 demonstrates the effectiveness of recent MIL-based models on the TCGA-Lung and CPTAC-Lung datasets, each utilizing different strategies for instance aggregation and feature modeling. Among prior methods, the approach proposed by Yu et al.[51] achieves the strongest results, with AUC values of 0.924 and 0.912 on TCGA and CPTAC respectively, benefiting from a hybrid design that incorporates region selection guided by cellular heterogeneity. However, its reliance on cell-level annotation and pre-processing pipelines imposes additional computational complexity and limits adaptability in fully automated workflows.

Methods like E2EFP-MIL [40] and DTFD-MIL[39] adopt early aggregation mechanisms and dynamic token fusion, showing solid performance particularly in precision and F1-score, but tend to underperform in capturing global contextual dependencies across spatially distributed patches. These models struggle to fully encode the nuanced histomorphological cues — such as inter-patch relationships or region co-dependencies — which are critical in differentiating complex subtypes of lung cancer.

Table 3.4: Performance comparison of SOTA-based methods on four lung cancer datasets.

Method	**TCGA-Lung**					**CPTAC-Lung**				
	AUC	Acc.	Prec.	Rec.	F1	AUC	Acc.	Prec.	Rec.	F1
ClassicMIL	0.845	80.20	0.810	0.790	0.800	0.832	78.50	0.793	0.775	0.784
DSMIL	0.873	82.40	0.832	0.815	0.823	0.854	81.00	0.817	0.800	0.808
TransMIL	0.888	83.80	0.845	0.830	0.837	0.867	82.20	0.830	0.812	0.821
DTFD-MIL	0.901	85.00	0.858	0.842	0.850	0.879	83.50	0.842	0.825	0.833
E2EFP-MIL	0.915	86.60	0.872	0.860	0.866	0.902	85.20	0.860	0.845	0.852
Yu et al.	0.924	87.80	0.884	0.870	0.877	0.912	86.80	0.872	0.855	0.863
E-GloConNet	**0.965**	**95.01**	**0.949**	**0.954**	**0.952**	**0.978**	**96.84**	**0.966**	**0.974**	**0.970**

ClassicMIL [36], though foundational and robust at scale, exhibits noticeably lower performance across both datasets, primarily due to its simplistic aggregation strategy and lack of attention mechanisms. DSMIL [62] and TransMIL [50] improve upon this by incorporating dual-stream learning and transformer-based attention, respectively. While these models offer better spatial reasoning and moderately improved recall, they are often parameter-heavy (e.g., TransMIL) and fail to generalize effectively when diagnostic information is sparse or scattered across the slide.

In contrast, our proposed **E-GloConNet** achieves a significant leap in performance — reaching an AUC of 0.965 on TCGA-Lung and 0.978 on CPTAC-Lung, alongside F1-scores of 0.952 and 0.970. These gains are attributed to our design's emphasis on global context attention, which together enable precise modeling of both fine-grained cellular structures and higher-order spatial organization across tiles. This hierarchical attention framework allows the model to effectively distinguish informative instances from irrelevant ones, enhancing classification even in slides with high heterogeneity. Furthermore, E-GloConNet's lightweight, end-to-end design ensures both computational efficiency and scalability, avoiding the heavy overhead of methods requiring handcrafted selection or excessive pre-training.

By capturing richer histomorphological features — including nuclear arrangement, glandular architecture, and tissue texture — our method not only surpasses the state-of-the-art in key metrics but also demonstrates robustness across diverse datasets with varying staining patterns and resolutions.

3.2.4 Impact of Incorporating Global Context Attention in E-GloConNet

To enrich the feature extraction capabilities of EfficientNetV2-B0, we incorporated a GCA mechanism, culminating in the development of our proposed architecture, E-GloConNet. This integration allows the model to effectively capture long-range spatial dependencies and holistic contextual patterns that are vital for recognizing key histological features—such as keratin pearls in squamous cell carcinoma, glandular morphology in adenocarcinoma, and diffuse boundaries between tumor and stroma. By aggregating contextual signals from the entire spatial domain, the GCA module enables the network to draw meaningful associations between large-scale structural arrangements and localized cellular cues, such as nuclear irregularities, mitotic activity, and cytoplasmic textures. This synergy between global and local features contributes to more accurate and interpretable predictions.

Empirical results highlight the benefit of this enhancement: as detailed in Table 3.3, integrating GCA into the baseline backbone led to a marked improvement in classification performance. The Area Under the Curve (AUC) rose from 0.9353 to 0.9648, while overall accuracy improved by 2.34% (from 92.67% to 95.01%). The F1-score also showed a significant uplift, increasing from 0.9147 to 0.9515, with corresponding gains in precision (+0.0479) and recall (+0.0255). These metrics collectively underscore the framework's improved ability to detect true positive cases while reducing misclassifications—an essential requirement in high-stakes clinical applications. The results affirm the importance of global context modeling in learning biologically relevant features crucial for precise lung cancer subtype classification.

Summary

In this chapter, we presented the experimental evaluation of our proposed E-GloConNet framework, demonstrating its effectiveness in classifying lung cancer subtypes. The model consistently outperformed baselines and state-of-the-art approaches, validating the role of global contextual modeling. Ablation studies further confirmed the individual contributions of key components. These findings establish a solid foundation for the broader discussion of implications and limitations in the next chapter.

Chapter 4

Transformer-Based Histopathological Analysis

AttenEViT-HDMIL: A Hierarchical Attention Transformer for Lung Cancer WSI Analysis

In the previous chapter, we demonstrated the effectiveness of E-GloConNet, which integrated Global Context Attention (GCA) into an EfficientNetV2-B0 backbone to enhance spatial reasoning and contextual awareness in histopathological lung cancer classification. While E-GloConNet delivered strong performance with notable improvements over baseline and state-of-the-art models, it relied primarily on convolutional feature extractors, which may limit the model's capacity to fully capture long-range dependencies and complex tissue morphology in WSIs.

To address these limitations, this chapter introduces **AttenEViT-HDMIL**, a novel transformer-based framework that advances our previous approach by incorporating dual-feature extraction, hierarchical attention, and a high-density multi-instance learning (HD-MIL) strategy. This design not only enhances computational efficiency but also focuses on biologically relevant high-cell-density tumor regions, enabling more accurate and scalable classification. By combining global vision transformer capabilities with domain-specific attention mechanisms, AttenEViT-HDMIL sets a new benchmark for precision histopathology while maintaining interpretability and robustness across staining variations.

4.1 Image Pre-processing

To prepare the histopathological WSIs for transformer-based analysis, we implement a two-stage preprocessing and sampling pipeline designed to retain high-density, diagnostically relevant regions.

4.1.1 Tissue Segmentation and Tiling

To ensure high-quality and informative regions for model training and inference, we first extract tissue regions from each WSI using Otsu-based thresholding [52] for effective background segmentation, as shown in Fig. 4.1(a). The retained gigapixel tissue regions are divided into non-overlapping tiles of size 224 × 224 pixels at 40× magnification. Depending on the tissue content, this results in hundreds to several hundred thousand tiles per slide. Tiles with less than 80% foreground tissue are discarded, which reduces the dataset size by approximately 70–80% while preserving the most informative tissue areas. This filtering ensures computational efficiency without compromising diagnostic utility.

4.1.2 Biomedical Knowledge-Guided Intelligent Sampling

Despite aggressive foreground filtering, not all retained tiles contribute equally to diagnosis. Prior studies have shown that only a small fraction of tiles (less than 20%) are highly relevant for histopathological classification [62]. Regions with dense nuclear content often reflect increased cellular proliferation, a hallmark of malignancy [63]. To prioritize diagnostically important regions, we employ a U-Net-based segmentation model [47], pre-trained on the 2018 Data Science Bowl (DSB) dataset [64], and fine-tuned on TCGA lung cancer data for domain adaptation. For each tile T_i^j, the model predicts a binary segmentation mask M_i^j, where:

$$M_{\text{U-Net}} : T_i^j \rightarrow M_i^j, \quad M_i^j \in \{0,1\}^{H_t \times W_t}. \tag{4.1}$$

Here, pixels classified as nuclei are assigned a value of 1, and background pixels are set to 0. The model is trained using the Dice coefficient loss function, defined as:

$$\mathcal{L}_{\text{dice}}(M_i^{j,\text{true}}, M_i^{j,\text{pred}}) = 1 - \frac{|M_i^{j,\text{true}}| + |M_i^{j,\text{pred}}| + \epsilon}{2 \cdot |M_i^{j,\text{true}} \cap M_i^{j,\text{pred}}| + \epsilon}, \tag{4.2}$$

where $M_i^{j,\text{true}}$ and $M_i^{j,\text{pred}}$ denote the ground truth and predicted masks, respectively, and ϵ is a small constant to avoid division by zero. To enhance boundary delineation, Canny edge detection $C(M_i^j)$ is applied, overlaying detected nuclear contours on the original tiles to aid interpretability. We then quantify the number of nuclei in each tile using connected component labeling:

$$N_i^j = |\text{Unique}(L(M_i^j))| - 1, \tag{4.3}$$

where $L(M_i^j)$ assigns distinct labels to connected components in the segmentation mask. Based on the nuclei count N_i^j, tiles are ranked and the top-k tiles are selected per WSI:

$$\mathcal{T}_{\text{i, int}} = \{T_i^1, T_i^2, \ldots, T_i^k \mid \text{Top-}k\ N_i^j\}. \tag{4.4}$$

This approach significantly enhances the signal-to-noise ratio during training, by focusing on tiles most indicative of tumor morphology. Unless otherwise stated, we use $k = 1000$ tiles per WSI for training. For inference, all filtered (informative) tiles are used to maximize diagnostic coverage.

4.2 Proposed AttenEViT-HDMIL Framework

In this section, we introduce **AttenEViT-HDMIL**, our proposed end-to-end hybrid deep learning framework designed for WSI classification in computational pathology. The framework integrates lightweight convolutional backbones with transformer-based encoders, combining local spatial sensitivity with global contextual understanding. It is optimized for efficiency and precision in processing high-resolution histopathological data, addressing both the scale of the input and the heterogeneity in diagnostic features. The proposed approach introduces hierarchical attention mechanisms and a dual-branch learning structure for effective multi-instance learning, ultimately enabling more accurate and interpretable predictions.

4.2.1 The proposed AttenEViT-HDMIL framework

We present AttenEViT-HDMIL, a convolutional-transformer hybrid framework enhanced with dual-feature extraction and hierarchical attention mechanisms for efficient and accurate histopathological image classification. It integrates EfficientViT with novel improvements to boost feature extraction efficiency and enhance attention mechanisms for superior representation learning. The following sections discuss the core design of the proposed framework, AttenEViT-HDMIL, and its key enhancements over standard models.

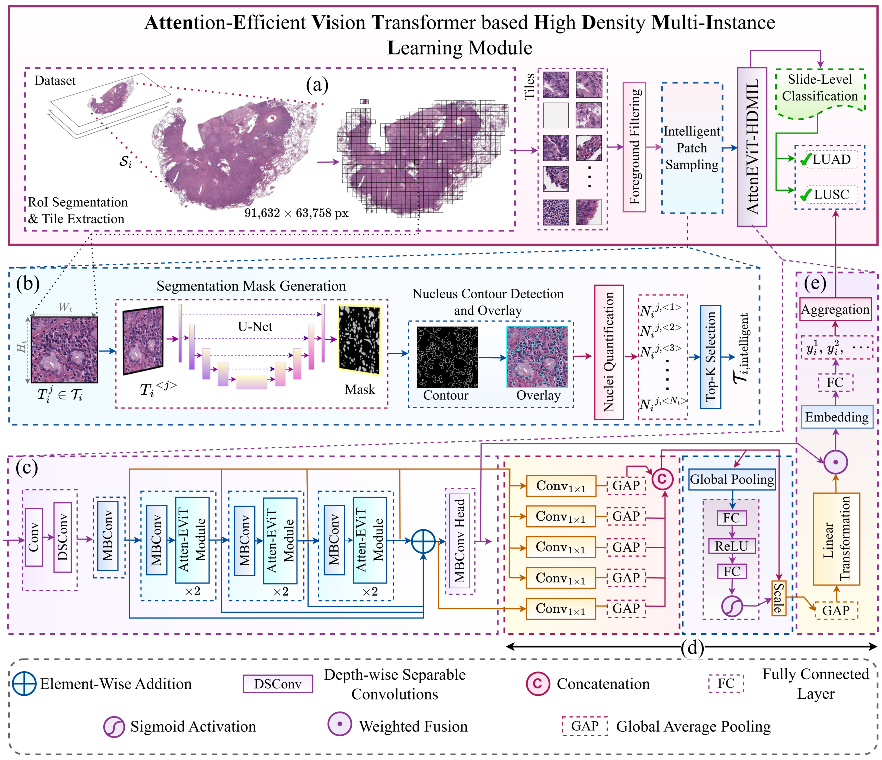


Figure 4.1: Overview of the proposed AttenEViT-HDMIL framework for classifying lung cancer subtypes. In step (a), WSI is segmented into ROIs, retaining 224 × 224 tiles with > 80% foreground. Top-k intelligently sampled tiles in step (b) are processed via AttenEViT-HDMIL in step (c), where transformer-extracted features undergo processing, dimensional-scaling, and fusion with class probabilities and Multi-scale feature integration in step (d). Step (e) aggregates tile-level predictions for final slide-level classification.

EfficientViT serves as the foundational model, combining convolutional layers for local feature extraction with transformer modules for global context understanding. It features an efficient deep feature extractor, multi-scale linear attention encoders, and a lightweight mobile inverted bottleneck convolution (MBConv) classification head. We enhance this framework by incorporating additional mechanisms tailored for hierarchical attention and more effective multi-instance learning to handle the inherent complexities of WSIs and varying levels of biological relevance across regions, respectively. The workflow and the key components of the AttenEViT-HDMIL framework are presented in Fig. 4.1. The input tiles are divided into smaller non-overlapping patches and linearly projected into a high-dimensional embedding space. These embeddings, enriched with positional information and a learnable class token, form the input sequence to the transformer encoder, enabling it to capture spatial and contextual relationships across regions in each tile without additional convolutional overhead.

4.2.1.1 Attention-Efficient transformer encoders

Following the patching and position embedding stage, we sequentially employ the attention-efficient transformer encoders. Six attention-efficient transformer encoders are grouped into three stages, each consisting of two encoders, as demonstrated in Fig. 4.1 (b). The input to the l^{th} transformer encoder (AttenEViTE_l) is the concatenated feature encoding ($\mathcal{E}$) from all preceding AttenEViT encoders as expressed using Eq. (4.5).

$$\mathcal{E}(\text{AttenEViTE}_\ell) = [\mathcal{E}(\text{AttenEViTE}_1), \mathcal{E}(\text{AttenEViTE}_2), \ldots, \mathcal{E}(\text{AttenEViTE}_{\ell-1})] \quad (4.5)$$

The building blocks of an AttenEViTE are shown in Fig. 4.1 (c). It features a convolutional-transformer hybrid architecture, primarily comprising MBConvs [61] and AttenEViT modules. The architecture begins with a standard convolutional stem followed by a depthwise separable convolution (DSConv) layer, consisting of depthwise convolutions (DWConv) and pointwise convolutions (PWConv), facilitating the efficient extraction of local features into a feature map, F_{DSConv}, given by $\text{DSConv}(z_0)$. The MBConv operation, designed for improved quantization and computational efficiency through 1×1 convolutions, produces the feature map, $F_{\text{MBConv}} = \text{MBConv}(F_{\text{DSConv}})$. After the initial feature extraction, the framework introduces significant enhancements to address the complexity of histopathological images, where capturing cellular morphology and alignment is crucial. While EfficientViT provides a strong foundation for local and global feature extraction, its standard configuration does not fully capture the intricate morphological characteristics and spatial patterns of these images. To address this limitation, we propose integrating linear multi-scale attention (LinearMSA) and a global context-aware attention (GCA) mechanism, as shown in Fig. 4.1 (c). The model computes GCA in parallel with the LinearMSA, enabling the capture of correlations between distant regions within a patch. This is essential for understanding complex histopathological patterns, tissue structure, and cellular cohesion. EfficientViT employs lightweight convolutions with varying kernel sizes, strides, and feature dimensions, along with LinearMSA, which differs from vanilla self-attention in standard ViTs. By combining

GCA with LinearMSA, the model achieves fine-grained local detail and broader contextual awareness, fusing these features for a more comprehensive representation. The input image undergoes a linear transformation, converting the flattened image into tokenized representations. These tokens are then processed by GCA to capture global interdependencies and by LinearMSA to improve feature extraction with reduced complexity. Finally, the spatial attention layer refines the fused feature representations from LinearMSA and GCA, achieving improved slide-level predictions. This design effectively balances computational efficiency and predictive performance while ensuring robustness across heterogeneous datasets. To further elaborate on the LinearMSA and GCA mechanisms, we provide their mathematical formulations and detailed working principles below.

a. *Linear multi-scale attention (LinearMSA):* In LinearMSA, the input features Query(Q), Key(K), and Value(V) are projected into matrices. Lightweight convolutions are used to process these matrices, which generate multi-scale tokens. Next, to attain a global receptive field with linear complexity, ReLU-based linear attention is employed. This is accomplished using an alternative similarity function, replacing the traditional softmax-based attention mechanism. The generalized form of softmax attention is expressed in Eq. (4.6) below.

$$\mathcal{A}_i = \sum_{j=1}^{N_p} \frac{Sim(Q_i, K_j)}{\sum_{j=1}^{N_p} Sim(Q_i, K_j)} V_j, \tag{4.6}$$

where $Q = xW_Q$, $K = xW_K$, $V = xW_V$ with $x \in \mathbb{R}^{N_p \times f}$, and $W_Q/W_K/W_V \in \mathbb{R}^{f \times d}$ are learnable linear projection matrices. Here, $\mathcal{A}_i$ denotes the output matrix $\mathcal{A}$ with i^{th} row, and similarity function given by $Sim(\cdot, \cdot)$. The conventional similarity function, Eq. (4.7) is employed to obtain the original softmax attention mechanism.

$$Sim(Q, K) = \exp\left(\frac{QK^T}{\sqrt{d}}\right) \tag{4.7}$$

In contrast, ReLU linear attention [65] achieves the desired properties by substituting $Sim(Q, K)$ by $\text{ReLU}(Q) \times \text{ReLU}(K)^T$ into Eq. (4.6). This substitution reduces the computational complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$, utilizing the associative property of matrix multiplication. The simplification is facilitated by computing only the expressions $\sum_{j=1}^{N_p} \text{ReLU}(K_j)^T V_j$ and $\sum_{j=1}^{N_p} \text{ReLU}(K_j)^T$ once, allowing them to be reused across multiple queries, thereby significantly enhancing efficiency while maintaining performance [66]. Finally, the LinearMSA output is concatenated and projected to form the final feature map for subsequent processing.

b. Global context-aware attention (GCA): The GCA mechanism, captures and integrates global information from the entire tile, complementing the local features obtained through LinearMSA. WSIs and their correspondingly extracted tiles contain vast cellular and tissue detail across extensive spatial dimensions. This vast morphological information necessitates an attention mechanism capable of effectively capturing spatial dependencies and reducing noise, which is beneficial in high-resolution dense images, where local texture variations can obscure critical patterns [67, 68, 69]. To achieve robust global feature representations, we compute the global context vector g using global average pooling (GAP) over the input feature map, F_{MBConv}, efficiently summarizing patch-wise features [67], as expressed in Eq. (4.8).

$$g = \text{GAP}(f_{\text{MBConv}}) = \frac{1}{N_p} \sum_{i=1}^{N_p} f_{\text{MBConv},i} \tag{4.8}$$

The global context vector g consolidates spatial information from patches, enhancing the identification of critical morphological relationships necessary for accurate classification and diagnosis. To effectively modulate local patch features with global information, g is transformed into $g_{\text{transformed}}$ using a shared weight matrix W_g designed to capture global dependencies. The transformation includes a sigmoid activation for non-linearity, represented as $g_{\text{transformed}} = \sigma(W_g \cdot g), \quad g_{\text{transformed}} \in \mathbb{R}^{1\times C}$. Subsequently, each patch embedding $f_{\text{MBConv},i}$ is modulated by the transformed global context vector $g_{\text{transformed}}$ through element-wise multiplication, to enrich feature representation by integrating global contextual information. This has been mathematically shown in Eq. (4.9). Finally, the globally modulated feature map F_{GCA} forms a discriminative representation for downstream tasks, such as classification, as shown in Eq. (4.10).

$$f'^{j}_{\text{MBConv},i} = f^{j}_{\text{MBConv},i} \odot g_{\text{transformed}}, \quad f'^{j}_{\text{MBConv},i} \in \mathbb{R}^{C} \tag{4.9}$$

$$F_{\text{GCA}} = \left[f'^{j,1}_{\text{MBConv},i}, f'^{j,2}_{\text{MBConv},i}, \ldots, f'^{j,N_p}_{\text{MBConv},i} \right] \tag{4.10}$$

The outputs from the LinearMSA and GCA modules, denoted $F_{\text{LinearMSA}}$ and F_{GCA} respectively, are adaptively fused to unify both local and global contextual information. This fusion employs learnable weights α and β to balance their relative contributions, as shown in Eq. (4.11).

$$F_{\text{fused}} \leftarrow \alpha \cdot F_{\text{LinearMSA}} + \beta \cdot F_{\text{GCA}} \tag{4.11}$$

To further enhance the discriminative capacity of these fused features, a Spatial Attention (SA) mechanism is applied. This layer emphasizes spatially relevant regions by computing channel-wise descriptors via average and max pooling, which are then concatenated and refined using a convolutional layer to produce an attention map, $\mathcal{A}_{\text{spatial}}$:

$$\mathcal{A}_{\text{spatial}} = \sigma(\text{Conv}(\text{Concat}([F_{\text{avg}}, F_{\text{max}}]))) \tag{4.12}$$

By selectively amplifying informative regions—particularly those with high cellular density—the SA module improves the model's sensitivity to fine-grained morphological variations. The resulting representation is then passed to an MBConv layer for final refinement and integration into the overall EfficientViT encoder.

4.2.2 Multi-scale feature integration and slide-level classification

To produce a robust and accurate diagnosis from histopathological slides, our framework integrates multi-scale features extracted from different stages of the transformer encoder. These features capture a mix of fine-grained and high-level contextual information across various resolution levels. We combine these representations adaptively, using a learned weighting mechanism that emphasizes the most informative features while suppressing noise. Once we obtain refined predictions for each tile, we aggregate them to infer the slide-level diagnosis. This is achieved by averaging the predicted probabilities across all selected tiles within a slide. Such aggregation ensures that the final prediction reflects a consensus from multiple regions, capturing the overall histopathological pattern present in the tissue. To improve performance in lung cancer subtype classification, we also employ a progressive fine-tuning strategy. Starting from ImageNet-pretrained weights, we gradually adapt the model to domain-specific patterns in the TCGA dataset, allowing it to specialize in distinguishing LUAD from LUSC more effectively.

4.3 Experimental Evaluation and Results

In this section, we rigorously evaluate the effectiveness of our proposed framework through extensive experiments on lung cancer WSIs. We begin by detailing the experimental setup, including dataset specifications, preprocessing strategies, and training protocols. This is followed by a comprehensive performance analysis using multiple evaluation metrics at both tile-level and slide-level. We compare our method against state-of-the-art baselines to highlight its strengths and demonstrate its superiority in accurately classifying histopathological subtypes. Additionally, we conduct ablation studies to assess the contribution of individual components within our model. The results are further supported by qualitative visualizations, offering insights into the interpretability and robustness of our approach.

4.3.1 Implementation Details

We implement the proposed **AttenEViT-HDMIL** framework using PyTorch 2.4.0 [70] and Python 3.12.3. All experiments are conducted on a high-performance computing cluster equipped with 8 NVIDIA DGX A100 Tensor Core GPUs (totaling 320 GB GPU memory) and a dual AMD EPYC 7742 CPU setup (128 cores). WSIs are processed using OpenSlide (v3.4.1). Training, validation, and test splits follow the strategy described in Section 3.2.1.2, with all performance metrics reported on the held-out test set $\mathcal{S}_{\text{Test}}$. The core transformer encoder consists of 6 layers organized into 3 hierarchical stages, with an embedding dimension of 32, patch size of 32×32, and input image resolution fixed at 224×224. Further architectural details, including token initialization, dropout, normalization, and classifier design, are summarized in Table 4.1(a).

For optimization, we use the Adam optimizer to update the model parameters, with hyperparameters carefully selected through cross-validation on the validation set $\mathcal{S}_{\text{Val}}$. The complete set of training and optimization configurations, including batch size, learning rate schedule, early stopping strategy, and regularization, is detailed in Table 4.1(b). A plateau-based learning rate scheduler is employed with a patience of 20 epochs to adaptively refine learning dynamics, while early stopping prevents overfitting by terminating training when no improvement is observed within the same threshold. Additionally, all learnable components—including the attention fusion weights (α, β) and weight matrix W_g—are tuned

Table 4.1: Hyperparameter summary for the proposed AttenEViT-HDMIL framework.

a. Transformer encoder hyperparameters	
Image size (w_z)	224
Patch size (h_p, w_p)	32
Number of patches (n_p)	49
Channel count (c_p)	3 (RGB)
Transformer layers (L)	6 (divided into 3 stages)
Embedding dimension ($\mathcal{D}$)	32
Classifier head dimensions	Fully connected layer, hidden dimension 512
Dropout rate	0.3
Activation	GeLU
Normalization	$\epsilon = 1 \times 10^{-5}$
CLS token	Random initialization (32-D embedding)
b. Optimization hyperparameters	
Batch size	50
Epochs	150
Epochs for early stopping	20
Loss	Categorical cross-entropy
Optimizer	Adam
Optimizer Learning rate	1×10^{-4}
Momentum	0.9
Epsilon (ϵ)	1×10^{-8}
Beta 1 (β_1)	0.9
Beta 2 (β_2)	0.999
Weight decay	4×10^{-5}
Learning rate scheduler	Adaptive (plateau-based)
Scheduler patience	20 epochs

during this phase.

4.3.2 Comparative Analysis with State-of-the-Art MIL Frameworks

Table 4.2 presents a comprehensive performance comparison of various Multiple Instance Learning (MIL)-based models on two challenging lung cancer histopathology datasets: **TCGA-Lung** and **CPTAC-Lung**. The proposed **AttenEViT-HDMIL** framework (denoted as *AttenEViT-HDMIL* in the table) significantly outperforms all existing state-of-the-art (SOTA) methods across all evaluation metrics, including AUC, accuracy, precision, recall, and F1-score.

On the **TCGA-Lung** dataset, AttenEViT-HDMIL achieves an **AUC of 0.9802**, which is notably higher than the closest SOTA method by Yu *et al.* [51], which reports an AUC of 0.9392. Additionally, AttenEViT-HDMIL reports the highest accuracy (95.08%) and F1-score (0.9575), indicating superior capability in correctly identifying both benign and malignant instances with minimal misclassifications. In contrast, other strong MIL variants such as DTFD-MIL and E2EFP-MIL show comparatively lower performance, with F1-scores of 0.8946 and 0.9202, respectively. These improvements can be attributed to the hierarchical attention mechanisms and deep feature fusion strategies employed by our model, which enhance both local and global contextual understanding of high-resolution tissue regions.

Table 4.2: Performance comparison of MIL-based methods on four lung cancer datasets.

Method	**TCGA-Lung**					**CPTAC-Lung**				
	AUC	Acc.	Prec.	Rec.	F1	AUC	Acc.	Prec.	Rec.	F1
ClassicMIL	0.7811	70.02	0.7522	0.7961	0.7732	0.7620	68.50	0.7350	0.7550	0.7449
DSMIL	0.8357	82.03	0.8325	0.8572	0.8445	0.8120	79.30	0.8001	0.8203	0.8101
TransMIL	0.8506	84.07	0.8459	0.8725	0.8583	0.8290	81.20	0.8150	0.8400	0.8273
DTFD-MIL	**0.9153**	**90.23**	0.8878	0.9012	0.8946	0.8910	87.05	0.8652	0.8803	0.8727
E2EFP-MIL	0.8901	88.23	**0.9101**	**0.9305**	**0.9202**	0.8655	85.10	0.8902	0.9153	0.9026
Yu et al.	0.9392	92.25	0.9300	0.9405	0.9346	0.9150	89.80	0.9105	0.9252	0.9178
AttenEViT-HDMIL	**0.9802**	**95.08**	**0.9507**	**0.9641**	**0.9575**	**0.9811**	**95.65**	**0.9555**	**0.9680**	**0.9617**

On the **CPTAC-Lung** dataset, AttenEViT-HDMIL demonstrates consistent and robust generalization, achieving an **AUC of 0.9811** and an F1-score of 0.9617, again outperforming all compared methods. *It is worth noting that the results obtained on the CPTAC-Lung dataset were generated through transfer learning from the TCGA-Lung-trained model, as discussed in Section 3.2.1.2.* Despite domain shifts between the two cohorts, the proposed model maintains its superiority in all metrics. This reflects its resilience and adaptability to distributional variability, which is essential for real-world deployment in computational pathology pipelines.

The consistent improvement across both datasets illustrates that the proposed

AttenEViT-HDMIL not only learns more discriminative representations but also effectively captures hierarchical semantic dependencies between instances. These results confirm the efficacy of integrating multi-head attention, global context modeling, and hybrid token embeddings in MIL-based classification tasks, especially for the complex histological patterns observed in lung cancer subtypes.

4.3.3 Effect of Hierarchical Attention Modules in the Transformer Encoder

To evaluate the contribution of different attention mechanisms within the transformer encoder of the proposed AttenEViT-HDMIL architecture, we conducted a systematic ablation study using various configurations of *Linear Multi-Scale Attention (LinearMSA)*, *Global Context Attention (GCA)*, and *Self-Attention (SA)*. The results, summarized in Table 4.3, examine the individual and combined effectiveness of these components under two sampling strategies: random and intelligent patch selection. Initially, we assessed the model performance using only LinearMSA. With random sampling, this configuration achieved an AUC of 0.9145 and an accuracy of 89.12%. When switching to intelligent sampling—which focuses on morphologically relevant regions—the performance improved to an AUC of 0.9334 and an accuracy of 91.78%. This highlights the advantage of prioritizing diagnostically significant regions in the learning process.

The addition of GCA—designed to enhance global spatial awareness across patches—led to further improvements. When combined with LinearMSA, the model attained an AUC of 0.9267 and accuracy of 89.58% under random sampling, which increased substantially to 0.9689 AUC and 94.87% accuracy with intelligent sampling. These gains suggest that GCA plays a crucial role in modeling contextual dependencies at a broader scale. Finally, the full integration of LinearMSA, GCA, and SA (i.e., the complete AttenEViT-HDMIL configuration) yielded the most significant performance improvements. Under random sampling, the model reached an AUC of 0.9331 and accuracy of 92.28%, while intelligent sampling delivered the best overall results—achieving an AUC of 0.9802 and an accuracy of 95.08%.

These findings collectively demonstrate the value of a hierarchical attention design within the transformer encoder. While LinearMSA effectively captures multi-scale local features, GCA contributes to holistic context aggregation, and SA enhances the model's sensitivity to fine-grained morphological patterns. Furthermore, intelligent sampling consistently leads to better performance across all settings, reaffirming the benefit of focusing on high-yield regions in histopathological images.

Table 4.3: Performance comparison of different configurations of hierarchical attention mechanisms and sampling strategies.

Model Configuration	**Sampling strategy**	**AUC**	**Accuracy (%)**	**Precision**	**Recall**	**F1-Score**
LinearMSA	Random	0.9145	89.12	0.8897	0.9023	0.8959
LinearMSA	Intelligent	0.9334	91.78	0.9155	0.9262	0.9208
LinearMSA + GCA	Random	0.9267	89.58	0.8941	0.9002	0.8971
LinearMSA + GCA	Intelligent	0.9689	94.87	0.9465	0.9543	0.9504
LinearMSA + GCA + SA	Random	0.9331	92.28	0.9380	0.9278	0.9267
AttenEViT-HDMIL	**Intelligent**	**0.9802**	**95.08**	**0.9507**	**0.9641**	**0.9575**

4.3.4 Evaluation of Instance Sampling Strategies in Multiple Instance Learning

To assess the influence of different instance sampling strategies on the performance of the proposed AttenEViT-HDMIL framework, an ablation study was conducted. The results, summarized in Table 4.4, compare three strategies: *Intelligent Sampling* (S_{Int}), *Exhaustive Sampling* (S_{Ex}), and *Random Sampling* (S_{Rd}).

The **Intelligent Sampling** strategy focuses on selecting image regions characterized by high cellular density and pronounced morphological heterogeneity, which are indicative of diagnostically relevant features. This approach achieved the best overall balance across all evaluation metrics, reporting an AUC of 0.9802 and an accuracy of 95.08%, along with the highest precision, recall, and F1-score values. By contrast, the **Exhaustive Sampling** strategy, which involves training on the entire dataset without filtering for relevance, yielded a slightly higher AUC of 0.9832. However, this gain comes at the cost of reduced recall (0.9563 compared to 0.9641 with S_{Int}), likely due to the inclusion of redundant or irrelevant patches that dilute the learning signal and introduce noise. The **Random Sampling** approach, where instances are chosen uniformly without regard to underlying tissue characteristics, resulted in the poorest performance across all metrics. This underscores the necessity of incorporating domain-specific heuristics into sampling to guide the model toward informative

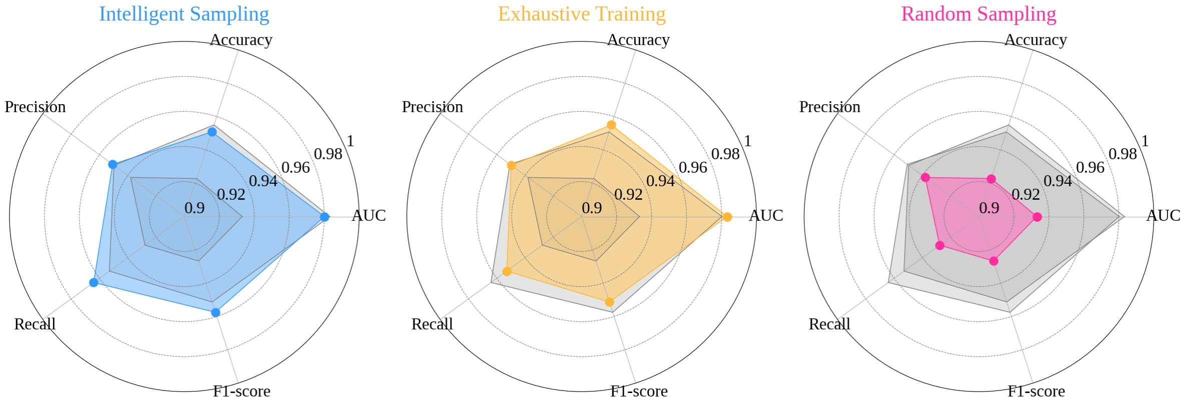


Figure 4.2: Radar plot comparing performance across three sampling strategies: knowledge-guided intelligent sampling, exhaustive training, and random sampling.

content.

To provide a unified evaluation across all metrics, a composite performance score C is computed as:

$$C = \frac{1}{n} \sum_{i=1}^{n} \text{Metric}_i, \quad C \text{ is Composite Score} \tag{4.13}$$

where n is the total number of evaluation metrics considered (e.g., AUC, accuracy, precision, recall, F1-score). According to this metric, the performance hierarchy is clearly established as $S_{\text{Int}} > S_{\text{Ex}} > S_{\text{Rd}}$.

While exhaustive sampling may marginally improve AUC, intelligent sampling offers a more efficient and practical solution, using only 7–10% of the total data while still achieving superior generalization. This makes it particularly advantageous for large-scale histopathological applications, where computational efficiency and diagnostic relevance are critical.

Table 4.4: Comparison of different sampling strategies based on AUC and composite score.

Sampling Strategy	**AUC**	**Composite Score**
Knowledge-Guided Intelligent Sampling	0.9802	$S_{\textbf{Intelligent}} = \mathbf{0.9606}$
Exhaustive Training	**0.9832**	$S_{\text{Exhaustive}} = 0.9582$
Random Sampling	0.9331	$S_{\text{Random}} = 0.9296$

Chapter Summary

In this chapter, we presented a detailed evaluation of the proposed AttenEViT-HDMIL framework across multiple benchmark lung cancer datasets. Through comprehensive comparisons with existing state-of-the-art MIL-based models, we demonstrated the superior performance of our approach in terms of classification accuracy, AUC, and other key evaluation metrics. Ablation studies further highlighted the individual contributions of the hierarchical attention modules—LinearMSA, GCA, and SA—as well as the significant impact of intelligent instance sampling in enhancing model robustness and efficiency. These insights validate the effectiveness of our architectural design and sampling strategy for high-resolution histopathology analysis.

The next chapter concludes this thesis by summarizing the key contributions, discussing broader implications, and outlining potential directions for future research.

Chapter 5

Conclusions

This chapter synthesizes the key contributions, findings, and implications of the deep learning frameworks proposed in this thesis for the histopathological classification of lung cancer subtypes. Building upon the challenges of data variability, annotation scarcity, and the need for scalable yet accurate diagnostic models, two distinct approaches—E-GloConNet and AttenEViT-HDMIL—were developed and rigorously evaluated. While both frameworks address the same clinical objective of distinguishing LUAD and LUSC from WSIs, they do so through fundamentally different architectural paradigms and learning strategies. This discussion critically examines the design choices, performance outcomes, and clinical relevance of each model, while highlighting how their complementary strengths contribute to the broader goal of AI-driven precision oncology. The chapter concludes by outlining potential future directions and opportunities for extending this work to other cancer types and multi-modal diagnostic contexts.

5.1 E-GloConNet: Lightweight Attention-Augmented CNN for Histopathology

E-GloConNet was developed as a lightweight yet effective convolutional neural network enhanced with global context attention (GCA) for histopathological image classification. This framework was designed to address several critical challenges in digital pathology: inter-slide staining variability, limited annotations, computational resource constraints, and the need for spatial interpretability.

A key contribution of E-GloConNet lies in its attention-integrated architecture, where the incorporation of GCA modules into a compact backbone (EfficientNetV2-B0) enables the capture of long-range spatial dependencies—essential for identifying diagnostic structures such as tumour nests, stromal boundaries, and glandular arrangements. Unlike conventional CNNs with local receptive fields, E-GloConNet effectively models non-local interactions through attention, improving both accuracy and interpretability. The training pipeline integrates stain normalization, geometric and photometric augmentations, and a progressive fine-tuning schedule, ensuring robustness across variable clinical conditions. Notably, stain normalization minimizes domain shift by aligning colour profiles across laboratories, allowing the model to focus on morphology rather than colour artifacts. Meanwhile, augmentation strategies introduce structural variability to support generalization across diverse tissue morphologies. Despite its attention-augmented design, E-GloConNet maintains a compact architecture with only 7.34 million trainable parameters—comparable to lightweight baselines like MobileNetV2—while achieving higher classification performance. It offers significant computational efficiency with a throughput of over 200 tiles per second and moderate memory requirements during both training and inference, making it suitable for deployment in clinical workflows with limited resources.

Importantly, E-GloConNet adopts a weakly supervised approach with foreground-aware random sampling, eliminating the need for exhaustive pixel-level annotations. This design choice significantly reduces annotation overhead without compromising diagnostic performance, as the model leverages domain-adaptive fine-tuning to compensate for weak supervision. In summary, E-GloConNet presents a balanced framework that combines architectural efficiency, interpretability, and generalization. Its modular design, low resource footprint, and strong diagnostic performance make it a practical solution for real-world applications in histopathological cancer diagnosis, especially in settings where computational and annotation resources are constrained.

5.2 AttenEViT-HDMIL: Hierarchical Transformer-Based MIL for Efficient Slide-Level Classification

AttenEViT-HDMIL represents a significant advancement in the development of transformer-based models for histopathological image analysis, with a specific focus on accurate and efficient classification of lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). This framework addresses the dual challenges of annotation scarcity and computational scalability by integrating multi-level attention mechanisms with multiple instance learning (MIL), guided by domain-specific morphological cues.

At the heart of AttenEViT-HDMIL lies a hierarchical attention transformer encoder that fuses three complementary attention strategies: Linear Multi-Scale Attention (LinearMSA) for capturing patch-wise relationships efficiently, Global Context Attention (GCA) for modeling spatial dependencies across larger tissue regions, and Semantic Attention (SA) for emphasizing high-level morphological semantics. This layered attention architecture allows the model to progressively aggregate local, regional, and global information, resulting in more discriminative feature representations for cancer subtype classification. To further enhance training efficiency and diagnostic focus, the model employs a nucleus-aware intelligent sampling strategy, selecting a representative subset of the most morphologically informative tiles from each WSI. This targeted sampling reduces training data volume to just 7–10% of the total slide content (i.e., $k = 1000$ tiles per WSI), while preserving critical diagnostic content. Compared to random or exhaustive sampling, this approach significantly improves the quality of input instances used in MIL aggregation, leading to better generalization and reduced computational burden.

The model operates under a weakly supervised learning regime, requiring only slide-level labels rather than detailed region or pixel-level annotations. Despite this minimal supervision, AttenEViT-HDMIL achieves state-of-the-art classification performance—outperforming several baseline and contemporary MIL approaches—demonstrating the strength of its hierarchical design and domain-guided tile selection. Furthermore, the inclusion of Grad-CAM visualizations provides interpretability by revealing attention hotspots aligned with clinically relevant features such as keratinization pearls in LUSC and glandular formations in LUAD. These visual explanations were independently validated by expert pathologists and oncologists, reinforcing the clinical trustworthiness of the model's predictions. With its modular structure, strong diagnostic accuracy (AUC = 0.9802), efficient data usage, and high interpretability, AttenEViT-HDMIL presents a compelling framework for deployment in clinical decision-support systems. Its adaptability across histopathological tasks and potential for integration with multi-modal data sources positions it as a cornerstone for future AI-driven precision oncology platforms.

5.3 Interpretability via Grad-CAM Visualization

Interpretability is a critical aspect of deploying deep learning models in clinical environments, where trust, transparency, and diagnostic alignment are essential. To assess the transparency of AttenEViT-HDMIL's decision-making process, we employed Gradient-weighted

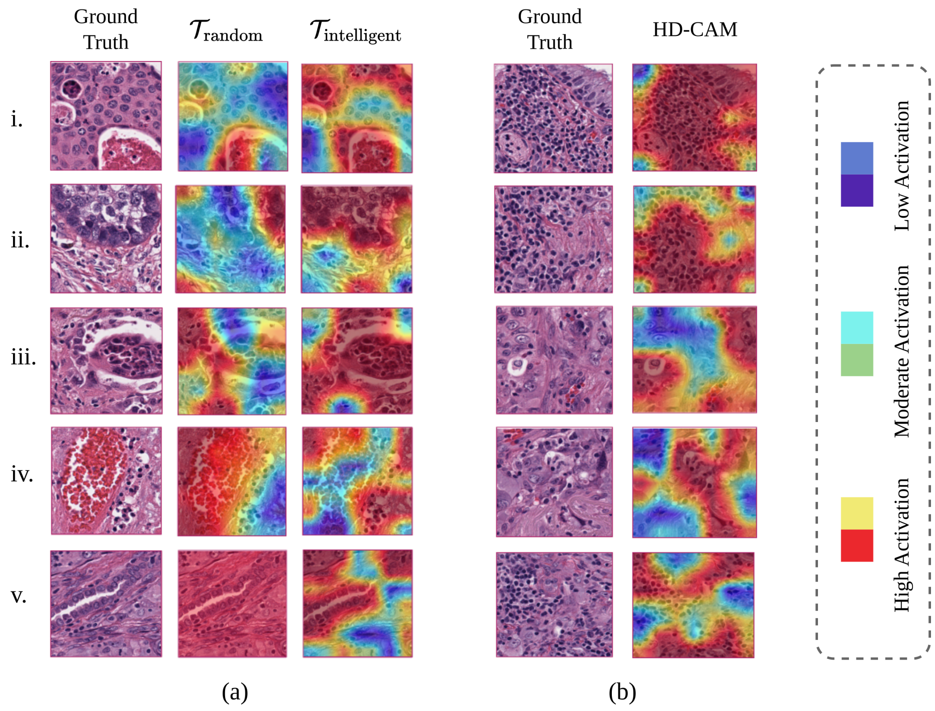


Figure 5.1: Grad-CAM visualizations comparing $\mathcal{T}_{\text{random}}$ and $\mathcal{T}_{\text{intelligent}}$. (a) shows enhanced focus on diagnostically relevant morphological features using intelligent sampling while reducing false activations. (b) HD-CAM focuses attention on high-density tumor regions, improving classification accuracy.

Class Activation Mapping (Grad-CAM) to visualize the attention patterns learned during training. These visualizations reveal the model's focus areas on WSIs, indicating which morphological regions contribute most significantly to its predictions. Figure 5.1 illustrates the comparative attention maps generated under different sampling strategies. Subfigures (a) i, iii, and v show that the intelligent sampling strategy ($\mathcal{T}_{\text{intelligent}}$) consistently guides the model toward diagnostically relevant structures, such as glandular formations characteristic of LUAD and keratin pearls associated with LUSC. In contrast, randomly sampled tiles ($\mathcal{T}_{\text{random}}$) result in dispersed and often irrelevant attention patterns, including stromal or erythrocyte-dominated areas, as shown in subfigures (a) ii and iv. Additionally, subfigure (b) demonstrates the high-density tumor region-focused HD-CAM, which further enhances classification specificity by narrowing attention to malignant cell clusters while minimizing background noise.

5.4 Pathologist review and Clinical validation

The adoption of deep learning models in clinical histopathology critically depends on their interpretability and alignment with diagnostic workflows. To evaluate the clinical robustness and practical utility of the proposed AttenEViT-HDMIL framework, a panel of experienced oncologists and pathologists specializing in pulmonary oncology independently reviewed the Grad-CAM visualizations generated by the model. These visual explanations illustrated the model's focus under both random and nucleus-guided intelligent sampling strategies, highlighting the specific regions of interest (ROIs) that influenced classification decisions. The experts carefully assessed the network's ability to localize high-density tumor areas and distinguish key histological features such as keratinization pearls, glandular architecture, nuclear atypia, and extracellular mucin deposits.

Their analysis confirmed that the model reliably concentrated on morphologically significant regions essential for differentiating LUAD from LUSC, while effectively ignoring irrelevant areas like stromal zones and erythrocyte-dense regions. The pathologists noted that the intelligent sampling strategy significantly improved the model's ability to highlight diagnostically relevant regions, thereby enhancing its explainability and trustworthiness. This close correspondence between the model's focus and expert diagnostic criteria underscores its potential for transparent, interpretable AI-assisted diagnosis.

Moreover, the oncologists emphasized the framework's value in reducing the need for extensive manual annotations and supporting more efficient diagnostic workflows—particularly in scenarios with limited data availability. By targeting tumor-dense regions, the model demonstrates high clinical applicability, offering precise insights into complex morphological patterns that may be difficult to identify through visual inspection alone. These include subtle structural formations such as solid tumor nests, stromal desmoplasia, and nuanced nuclear changes, all critical for accurate subtype classification. Through reliable feature localization and enhanced interpretability, AttenEViT-HDMIL shows considerable promise for real-world integration into digital pathology pipelines.

5.5 Directions for Future Research

This thesis presents two complementary deep learning frameworks—E-GloConNet and AttenEViT-HDMIL—that advance the state of computational pathology for lung cancer subtype classification using WSIs. E-GloConNet emphasizes lightweight, attention-enhanced convolutional architectures that achieve high diagnostic accuracy with efficient computational performance, making it suitable for deployment in resource-constrained settings. On the other hand, AttenEViT-HDMIL introduces a transformer-based hierarchical model that leverages weak supervision and intelligent instance sampling to focus on high-density, morphologically relevant regions, achieving superior performance while dramatically reducing the number of required input tiles.

Both models are tailored to address real-world challenges in digital pathology, including data variability, annotation scarcity, and computational scalability. Their consistent performance across diverse metrics and expert validation highlight their robustness and potential for clinical translation. Furthermore, the interpretability provided by the attention mechanisms and Grad-CAM visualizations enhances trust in their predictions, a critical factor for clinical adoption. Despite the promising outcomes, several directions remain open for further exploration:

- **Extension to Other Cancer Subtypes:** While this work focuses on LUAD and LUSC, extending these frameworks to other histological subtypes such as SCLC and rare NSCLC variants can expand their clinical applicability.

- **Multi-Institutional and Multi-Scanner Validation:** To ensure generalizability, both models should be evaluated using more distinct datasets from multiple institutions with varied staining protocols and scanner types.

- **Adaptive and Learnable Preprocessing:** Incorporating adaptive stain normalization and data augmentation strategies tailored to slide-specific characteristics could improve performance under variable conditions.

- **Integration of Multi-Modal Data:** Fusing histological features with genomic profiles, radiological imaging, and clinical metadata could enhance diagnostic precision and enable personalized treatment recommendations.

- **Real-Time and Edge Deployment:** Optimizing the models further for deployment in edge devices or real-time diagnostic pipelines could accelerate adoption in remote and resource-limited healthcare environments.

- **Weakly Supervised Localization and Explainability:** Enhancing attention maps with fine-grained segmentation capabilities can improve model interpretability, helping clinicians identify actionable regions within WSIs.

In summary, the approaches developed in this thesis lay a strong foundation for practical, interpretable, and scalable AI-based histopathological diagnosis. With further refinements and broader validation, these methods can significantly contribute to the future of precision oncology and digital pathology.

Bibliography

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 74, no. 3, pp. 229–263, 2024.

[2] World Health Organization, "Cancer — Fact Sheet," https://www.who.int/news-room/fact-sheets/detail/cancer, Feb. 2025, accessed on May 1, 2025.

[3] International Agency for Research on Cancer, "Lung cancer," https://www.iarc.who.int/cancer-type/lung-cancer/, 2025, accessed June 7, 2025.

[4] World Cancer Research Fund International, "Lung Cancer Statistics," https://www.wcrf.org/preventing-cancer/cancer-statistics/lung-cancer-statistics/, 2025, accessed June 7, 2025.

[5] R. L. Siegel, A. N. Giaquinto, and A. Jemal, "Cancer statistics, 2024," *CA: a cancer journal for clinicians*, vol. 74, no. 1, pp. 12–49, 2024.

[6] American Cancer Society, "Key Statistics for Lung Cancer," https://www.cancer.org/cancer/types/lung-cancer/about/key-statistics.html, 2025, accessed June 7, 2025.

[7] S. Dubin and D. Griffin, "Lung cancer in non-smokers," *Missouri medicine*, vol. 117, no. 4, p. 375, 2020.

[8] B. Bondhopadhyay, S. Sisodiya, A. Chikara, A. Khan, P. Tanwar, N. Singh, U. Agrawal, R. Mehrotra, S. Hussain *et al.*, "Cancer immunotherapy: A promising dawn in cancer research," *American journal of blood research*, vol. 10, no. 6, p. 375, 2020.

[9] B. Hochhegger, G. R. T. Alves, K. L. Irion, C. C. Fritscher, L. G. Fritscher, N. H. Concatto, and E. Marchiori, "Pet/ct imaging in lung cancer: indications and findings," *Jornal Brasileiro de Pneumologia*, vol. 41, pp. 264–274, 2015.

[10] H. Li, L. Gao, H. Ma, D. Arefan, J. He, J. Wang, and H. Liu, "Radiomics-based features for prediction of histological subtypes in central lung cancer," *Frontiers in Oncology*, vol. 11, p. 658887, 2021.

[11] G. Sommer, M. Koenigkam-Santos, J. Biederer, and M. Puderbach, "Role of mri for detection and characterization of pulmonary nodules," *Der Radiologe*, vol. 54, pp. 470–477, 2014.

[12] P. D. Shreve, Y. Anzai, and R. L. Wahl, "Pitfalls in oncologic diagnosis with fdg pet imaging: physiologic and benign variants," *Radiographics*, vol. 19, no. 1, pp. 61–77, 1999.

[13] R. Nooreldeen and H. Bach, "Current and future development in lung cancer diagnosis," *International journal of molecular sciences*, vol. 22, no. 16, p. 8661, 2021.

[14] M. S. Roh, "Molecular pathology of lung cancer: current status and future directions," *Tuberculosis and respiratory diseases*, vol. 77, no. 2, p. 49, 2014.

[15] C. Dunn, D. Brettle, M. Cockroft, E. Keating, C. Revie, and D. Treanor, "Quantitative assessment of h&e staining for pathology: development and clinical evaluation of a novel system," *Diagnostic Pathology*, vol. 19, no. 1, p. 42, 2024.

[16] F. Prezja, I. Pölönen, S. Äyrämö, P. Ruusuvuori, and T. Kuopio, "H&e multi-laboratory staining variance exploration with machine learning," *Applied Sciences*, vol. 12, no. 15, p. 7511, 2022.

[17] E. A. Chlipala, M. Butters, M. Brous, J. S. Fortin, R. Archuletta, K. Copeland, and B. Bolon, "Impact of preanalytical factors during histology processing on section suitability for digital image analysis," *Toxicologic Pathology*, vol. 49, no. 4, pp. 755–772, 2021.

[18] A. W. Alvarenga, C. M. Coutinho-Camillo, B. R. Rodrigues, R. M. Rocha, L. F. B. Torres, V. R. Martins, I. W. da Cunha, and G. N. Hajj, "A comparison between manual

and automated evaluations of tissue microarray patterns of protein expression," *Journal of Histochemistry & Cytochemistry*, vol. 61, no. 4, pp. 272–282, 2013.

[19] D. Tellez, M. Balkenhol, I. Otte-Höller, R. Van De Loo, R. Vogels, P. Bult, C. Wauters, W. Vreuls, S. Mol, N. Karssemeijer *et al.*, "Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks," *IEEE transactions on medical imaging*, vol. 37, no. 9, pp. 2126–2136, 2018.

[20] M. K. K. Niazi, A. V. Parwani, and M. N. Gurcan, "Digital pathology and artificial intelligence," *The lancet oncology*, vol. 20, no. 5, pp. e253–e261, 2019.

[21] D. Komura and S. Ishikawa, "Machine learning methods for histopathological image analysis," *Computational and structural biotechnology journal*, vol. 16, pp. 34–42, 2018.

[22] S. Doyle, M. Hwang, K. Shah, A. Madabhushi, M. Feldman, and J. Tomaszeweski, "Automated grading of prostate cancer using architectural and textural image features," in *2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro.* IEEE, 2007, pp. 1284–1287.

[23] X. Wang, H. Chen, C. Gan, H. Lin, Q. Dou, E. Tsougenis, Q. Huang, M. Cai, and P.-A. Heng, "Weakly supervised deep learning for whole slide lung cancer image analysis," *IEEE transactions on cybernetics*, vol. 50, no. 9, pp. 3950–3962, 2019.

[24] M. Y. Lu, T. Y. Chen, D. F. Williamson, M. Zhao, M. Shady, J. Lipkova, and F. Mahmood, "Ai-based pathology predicts origins for cancers of unknown primary," *Nature*, vol. 594, no. 7861, pp. 106–110, 2021.

[25] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, no. 1, pp. 221–248, 2017.

[26] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, and A. Tsirigos, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nature medicine*, vol. 24, no. 10, pp. 1559–1567, 2018.

[27] A. Madabhushi and G. Lee, "Image analysis and machine learning in digital pathology: Challenges and opportunities," *Medical image analysis*, vol. 33, pp. 170–175, 2016.

[28] C. G. A. R. Network *et al.*, "Comprehensive molecular profiling of lung adenocarcinoma," *Nature*, vol. 511, no. 7511, p. 543, 2014.

[29] C. G. A. R. Network *et al.*, "Comprehensive genomic characterization of squamous cell lung cancers," *Nature*, vol. 489, no. 7417, p. 519, 2012.

[30] N. C. I. C. P. T. A. C. (CPTAC), "The clinical proteomic tumor analysis consortium lung adenocarcinoma collection (cptac-luad) (version 12)," The Cancer Imaging Archive, Feb 2018, version 12 updated 2023-02-24.

[31] N. C. I. C. P. T. A. C. (CPTAC), "The clinical proteomic tumor analysis consortium lung squamous cell carcinoma collection (cptac-lscc) (version 15)," The Cancer Imaging Archive, Apr 2018, version 15 updated 2024-04-05.

[32] W. Bulten, H. Pinckaers, H. Van Boven, R. Vink, T. De Bel, B. Van Ginneken, J. van der Laak, C. Hulsbergen-van de Kaa, and G. Litjens, "Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study," *The Lancet Oncology*, vol. 21, no. 2, pp. 233–241, 2020.

[33] T. J. Fuchs and J. M. Buhmann, "Computational pathology: challenges and promises for tissue analysis," *Computerized Medical Imaging and Graphics*, vol. 35, no. 7-8, pp. 515–530, 2011.

[34] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[35] P. Khosravi, E. Kazemi, M. Imielinski, O. Elemento, and I. Hajirasouliha, "Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images," *EBioMedicine*, vol. 27, pp. 317–328, 2018.

[36] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature medicine*, vol. 25, no. 8, pp. 1301–1309, 2019.

[37] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2424–2433.

[38] P. Liu, B. Fu, F. Ye, R. Yang, and L. Ji, "DSCA: A dual-stream network with cross-attention on whole-slide image pyramids for cancer prognosis," *Expert Systems with Applications*, vol. 227, p. 120280, 2023.

[39] H. Zhang, Y. Meng, Y. Zhao, Y. Qiao, X. Yang, S. E. Coupland, and Y. Zheng, "DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 802–18 812.

[40] L. Cao, J. Wang, Y. Zhang, Z. Rong, M. Wang, L. Wang, J. Ji, Y. Qian, L. Zhang, H. Wu, J. Song, Z. Liu, W. Wang, S. Li, P. Wang, Z. Xu, J. Zhang, L. Zhao, H. Wang, M. Sun, X. Huang, R. Yin, Y. Lu, Z. Liu, K. Deng, G. Wang, M. Qiu, K. Li, J. Wang, and Y. Hou, "E2EFP-MIL: End-to-end and high-generalizability weakly supervised deep convolutional network for lung cancer classification from whole slide image," *Medical Image Analysis*, vol. 88, p. 102837, 2023.

[41] H. Zhou, H. Chen, B. Yu, S. Pang, X. Cong, and L. Cong, "An end-to-end weakly supervised learning framework for cancer subtype classification using histopathological slides," *Expert Systems with Applications*, vol. 237, p. 121379, 2024.

[42] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II 16*. Springer, 2013, pp. 411–418.

[43] Y. Xu, Z. Jia, Y. Ai, F. Zhang, M. Lai, and E. I.-C. Chang, "Deep convolutional activation features for large scale brain tumor histopathology image classification and segmentation," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 947–951.

[44] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[45] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *International conference on machine learning.* PMLR, 2018, pp. 2127–2136.

[46] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[47] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18.* Springer, 2015, pp. 234–241.

[48] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[49] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[50] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji *et al.*, "Transmil: Transformer based correlated multiple instance learning for whole slide image classification," *Advances in neural information processing systems*, vol. 34, pp. 2136–2147, 2021.

[51] F. Yu, X. Wang, R. Sali, and R. Li, "Single-cell heterogeneity-aware transformer-guided multiple instance learning for cancer aneuploidy prediction from whole slide histopathology images," *IEEE journal of biomedical and health informatics*, vol. 28, no. 1, pp. 134–144, 2023.

[52] N. Otsu *et al.*, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1979.

[53] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, and N. Navab, "Structure-preserving color normalization and

sparse stain separation for histological images," *IEEE transactions on medical imaging*, vol. 35, no. 8, pp. 1962–1971, 2016.

[54] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International conference on machine learning.* PMLR, 2021, pp. 10 096–10 106.

[55] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[56] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[57] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[58] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[59] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[60] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[61] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[62] 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*, 2021, pp. 14 318–14 328.

[63] V. Kumar, A. K. Abbas, J. C. Aster, and A. T. Deyrup, *Robbins & Kumar basic pathology, e-book: Robbins & Kumar basic pathology, e-book.* Elsevier Health Sciences, 2022.

[64] J. C. Caicedo, A. Goodman, K. W. Karhohs, B. A. Cimini, J. Ackerman, M. Haghighi, C. Heng, T. Becker, M. Doan, C. McQuin, M. Rohban, S. Singh, and A. E. Carpenter, "Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl," *Nature Methods*, vol. 16(12), no. 12, pp. 1247–1253, 2019.

[65] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are RNNs: fast autoregressive transformers with linear attention," in 37^{th} *International Conference on Machine Learning*, 2020, pp. 5156–5165.

[66] H. Cai, J. Li, M. Hu, C. Gan, and S. Han, "EfficientViT: Lightweight multi-scale attention for high-resolution dense prediction," in *2023 IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 256–17 267.

[67] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention LSTM networks for 3D action recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1647–1656.

[68] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, "Asymmetric non-local neural networks for semantic segmentation," in *2019 IEEE/CVF International Conference on Computer Vision*, 2019, pp. 593–602.

[69] A. Hatamizadeh, H. Yin, G. Heinrich, J. Kautz, and P. Molchanov, "Global context vision transformers," in 40_{th} *International Conference on Machine Learning*, 2023, pp. 12 633–12 646.

[70] A. Paszke, "Pytorch: An imperative style, high-performance deep learning library," *arXiv preprint arXiv:1912.01703*, 2019.