U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Automatic Report Generation from Histopathological Images

Athos Mekanna Moraes

Dissertação de Mestrado apresentada à
Faculdade de Ciências da Universidade do Porto em
Bioinformática e Biologia Computacional

2024

Automatic Report Generation from Histopathological Images

Athos Mekanna Moraes
Mestrado em Bioinformática e Biologia Computacional
Departamento de Ciência de Computadores
2024

Orientador
Pedro G. Ferreira, Professor Auxiliar
Faculdade de Ciências da Universidade do Porto

Coorientador
António Polónia, Médico Especialista
Instituto de Patologia e Imunologia Molecular da Universidade do Porto

Agradecimentos

O reconhecimento de que nossas realizações não são apenas o produto do nosso empenho individual, mas que também representam a soma de um coletivo de contribuições, é o que me leva a dedicar este espaço ao agradecimento de todos aqueles que, de alguma forma, depositaram sua fração no somatório que compõe o presente trabalho.

Ao meu orientador, Pedro Ferreira, que com grande atenção e entusiasmo introduziu-me ao universo académico, agradeço imensamente por seu suporte, conhecimentos transmitidos e pelo voto de confiança sobre o meu trabalho.

Ao António Polónio, agradeço pela aceitação em participar deste projeto.

Ao colega e veterano Rogério Ribeiro, obrigado pela paciência e por todos os valiosos conselhos académicos partilhados.

Ao Donald, que me introduziu à Ciência de Dados, agradeço pelos conhecimentos partilhados ao longo de nossa colaboração.

Aos amigos de longa data, Caetano, Camilo, Carol, Cecília, Erick, Gabriel, Isadora, Lucas, Pedro e Rodrigo, agradeço pela revisão e pelos anos de conversas nada convencionais.

Aos meus irmãos, Erick e Natasha, e aos meus pais, William e Silvana, que jamais deixaram de me apoiar em todos os meus projetos, sou eternamente grato.

Por fim, à minha esposa, Daniela, por sua infinita paciência, compreensão e suporte ao meu percurso académico, nenhuma palavra é o bastante para agradecê-la. Sem você, este trabalho não seria possível.

Abstract

For over a decade, Deep Learning (DL) has been successfully applied in histopathology, addressing challenges such as tumor detection, cellular segmentation, and biomarker quantification, among others. Following recent advancements, Foundation Models (FM) based on Transformer architectures have significantly gained popularity across academia, industry and public domains. This surge has led to unprecedented growth in the DL field, with new articles and ideas introducing fresh perspectives to longstanding challenges in Computer Vision (CV) and Natural Language Processing (NLP). As a result, interdisciplinary areas like Computational Pathology (CPATH) and Clinical NLP have seen significant advancements, particularly through Multimodal Deep Learning applications such as image captioning, visual question-answering (Q&A) and cross-modal retrieval (CMR).

Despite these advancements, many potential interconnections between strategies in these applications remain unexplored. Techniques such as Prompt Engineering (PE) and Retrieval-Augmented Generation (RAG), which are known to enhance responses in Large Language Models (LLMs), have not yet been fully integrated with CPATH algorithms into common systems. Furthermore, working with FMs entails significant computational demands during both training and inference phases. This high resource requirement complicates experimentation and evaluation of new algorithms, often necessitating extensive use of cloud-based GPUs. Additionally, the inherent challenges in handling gigapixel Whole Slide Images (WSIs) typically require the adoption of resizing and segmentation strategies, approaches that inevitably lead to the loss of crucial local and global information from the inherent structural arrangement of the tissues.

In an effort to overcome these challenges, our research intends to bridge the gap between CPATH and Clinical NLP by leveraging recent innovations in both fields to build a unified solution that integrates both domains. For that, we have developed an automatic pathology report generation system for WSIs guided by three main premises: first, to develop methodologies that can be replicated on systems with limited computational resources; second, to ensure the solution is easily maintainable and subject to future improvements; and third, to enhance accessibility for end users, thereby facilitating practical applications in both clinical and educational environments.

Focused on lung tissue, our system workflow operates in three main stages: Data Inference (INF), Aggregation (AGG), and Generation (GEN). In the first stage, the system predicts demographic data such as age and smoker status, and utilizes cross-modal retrieval to gather diagnosis information from similar cases. Then, in AGG, we generate a concise caption containing all the infomarion from the previous modules. Finally, in the GEN stage, the system incorporates all this information into a LLM enhanced with PE and RAG to generate a histopathology report enriched with aditional information from academic literature. Moreover, we have integrated our system into a user-friendly graphical interface, making it accessible to non-specialists in the computational field. This design facilitates the practical evaluation of our solution, allowing users to easily interact with and test the system's capabilities, ultimately enabling them to apply it in their own contexts.

The results we obtained, while not yet suitable for deployment in real clinical settings, demonstrate the practical feasibility of utilizing the aforementioned technologies in environments constrained by limited computational resources in order to test and generate valuable solutions for the histopathology field.

Resumo

Durante mais de uma década, o Deep Learning (DL) tem sido aplicado com sucesso na histopatologia, abordando desafios como a detecção de tumores, segmentação celular e quantificação de biomarcadores, entre outros. Com avanços recentes, os Foundation Models (FM) baseados em arquiteturas Transformer ganharam significativa popularidade na academia, indústria e domínios públicos em geral. Essa ascensão levou a um crescimento sem precedentes no campo do DL, com novos traabalhos e ideias trazendo novas perspectivas para os desafios de longa data em Visão Computacional (CV) e Natural Language Processing (NLP). Como resultado, áreas interdisciplinares como a Computational Pathology (CPATH) e Clinical NLP viram avanços significativos, especialmente através de aplicações multimodalidade de DL, como geração de caption para imagens, visual question-answering (Q&A) e cross-modal retrieval (CMR).

Apesar desses avanços, muitas interconexões potenciais entre estas diferentes estratégias permanecem inexploradas. Técnicas como Prompt Engineering (PE) e Retrieval Augmented Generation (RAG), conhecidas por melhorar respostas em Large Language Models (LLMs), ainda não foram totalmente integradas com algoritmos de CPATH em sistemas em comum. Além disso, trabalhar com FMs implica o uso de significativos recursos computacionais durante as fases de treinamento e inferência. Essa alta exigência de recursos dificulta a experimentação e avaliação de novos algoritmos, muitas vezes necessitando do uso extensivo de GPUs baseadas em cloud. Adicionalmente, os desafios inerentes ao manuseio das gigapixel Whole Slide Images (WSIs) tipicamente requerem a adoção de estratégias de redimensionamento e segmentação, abordagens que inevitavelmente levam à perda de informações locais e globais inerentes à disposição estrutural dos tecidos.

Na tentativa de superar esses desafios, nossa pesquisa pretende cobrir a lacuna existente entre a CPATH e o NLP Clínico, aproveitando as inovações recentes em ambos os campos para construir uma solução unificada que integre ambos os domínios. Para isso, desenvolvemos um sistema automático de geração de relatórios patológicos para WSIs, sistema este guiado por três principais premissas: em primeiro lugar, desenvolver metodologias que possam ser reproduzidas em sistemas com recursos computacionais limitados; em segundo lugar, garantir que este sistema seja passíveil de fácil manutenção e melhorias contínuas;

e em terceiro lugar, facilitar ao usuário final o acesso a este sistema para fins de aplicação prática em contextos clínicos e educacionais.

Focado em tecido pulmonar, o fluxo de trabalho do nosso sistema opera em três etapas principais: Inferência de Dados (INF), Agregação (AGG) e Geração (GEN). Na primeira etapa, o sistema prevê dados demográficos como idade e status de fumante, além de utilizar CMR para reunir informações de diagnóstico de casos semelhantes. A seguir, em AGG, geramos uma caption concisa contendo todas as informações geradas no estágio anterior. Por fim, na etapa de GEN, o sistema incorpora todas essas informações em um LLM aprimorado com PE e RAG para gerar um relatório de histopatologia enriquecido com informações adicionais da literatura acadêmica. Além disso, integramos nosso sistema em uma interface gráfica user-friendly, tornando-a acessível a não especialistas na área computacional. Esse design facilita a avaliação prática da nossa solução, permitindo que os usuários interajam facilmente com o sistema, testem suas capacidades e consigam aplicá-lo em seus contextos.

Os resultados obtidos, embora ainda não adequados para implantação em ambientes clínicos e educacioanais reais, demonstram a viabilidade prática de utilizar as mencionadas tecnologias em ambientes limitados por recursos computacionais para testar e gerar soluções de valor para o campo da histopatologia.

Contents

List of Figures

List of Tables

List of Abbreviations

AGG	Aggregation Stage
AI	Artificial Intelligence
AP	Age Prediction
CMR	Cross-Modal Retrieval
CNN	Convolutional Neural Network
CPATH	Computational Pathology
CV	Computer Vision
DL	Deep Learning
DR	Diagnosis Retrieval
FM	Foundation Model
GEN	Generation Stage
HIPT	Hierarchical Image Pyramid Transformer
INF	Inference Stage
LLM	Large Language Model
MLP	Multilayer Perceptron
NLP	Natural Language Processing
PE	Prompt Engineering
Q&E	Question-and-Answer
RAG	Retrieval-Augmented Generation
SC	Smoker Classification
SIM	Similarity Stage

ViT	Vision Transformer
VLM	Vision Language Model
WSI	Whole Slide Image

Chapter 1

Introduction

Histopathology is a specialized branch of Pathology focused on studying diseases in biological tissues primarily through the visual analysis and interpretation of glass slides containing frozen sections of study samples. Its historical roots trace back to the early 19th century, beginning with Johannes Muller who first coined the term histopathology [Dewi and Purwanto, 2023]. As a discipline reliant on microscope use, its development has paralleled technological advancements related to these instruments.

The advent of telepathology and the introduction of Whole Slide scanning devices between the 1980s and 1990s ([Jahn et al., 2020], [Farahani et al., 2015]) marked the emergence of a new area known as Computational Pathology, which is responsible for the study and application of techniques that enable the digitization of histology slides. Subsequent advancements in scanning technology have enabled the production of high-resolution slides in large-scale, paving the way for the application of Computational Pathology (CPATH) methods to handling Whole Slide Images (WSIs), aiming the analysis and extraction of useful information.

By 2013, influenced by the successful application of Convolutional Neural Networks (CNNs) in Computer Vision, the Computational Pathology research community began utilizing these architectures to analyze WSIs. Over the last decade, since the introduction of these first CNN systems, the academic production of Deep Learning (DL) algorithms in CPATH has grown significantly and has been the latest driving force for advances in the field of pathology [Cooper et al., 2023], demonstrating performance comparable to trained pathologists in various tasks [van der Laak et al., 2021].

Today, the histopathology workflow encompasses multiple challenges, from sample collection and laboratory processing to the final microscopic slide analysis. Despite a growing shortage of trained pathologists, the demand and complexity of their work continue to increase. In fact, due to the increasing number of cases to be analyzed, combined with the need for more precise diagnoses to identify the most ideal treatment for each patient, the workload of these professionals has become bigger than ever [Cooper et al., 2023]. Additionally, the variability in diagnosis among different pathologists on the same samples remains alarmingly high [Kartasalo et al., 2021].

Given these challenges, one might expect the last decade's advancements in CPATH to be invaluable for histopathology, not only in reducing pathologist's workloads but also in enhancing diagnostic accuracy and consistency. Yet, today, this field remains largely sustained by manual processes and traditional microscope analysis of glass slides. This situation is due to different factors that have been explored in the literature through various works such as [Cooper et al., 2023], [Abels et al., 2019], and [Jahn et al., 2020].

Despite these challenges, the pace of academic production in CPATH remains undiminished. Recent advances in Computer Vision and Deep Learning, particularly with the rise of transformer architectures and generative AI, have propelled innovative approaches in the field. These include the application of Vision Transformers (ViT) to handle the complex structure of WSIs [Chen et al., 2022] and multimodal integrations between CPATH and NLP in clinical contexts (Clinical NLP), which have led to novel systems for generating histology images from textual descriptions [Osorio et al., 2023], n-shot learning systems for histopathology classification [Ikezogwo et al., 2023], and caption generation systems for WSIs [Sengupta and Brown, 2023].

With this backdrop of deep learning's impact on histopathology and the emergence of new techniques alongside persistent implementation barriers, we now turn to the objectives of this work, formulated in response to these evolving challenges.

1.1 Research Objectives

As previously mentioned, the application of DL in histopathology has been significantly advanced by the emergence of Transformer-based FM models and Generative AI. These technologies have introduced promising multimodal applications between CAPTH and Clinical NLP. However, despite their recent development, many NLP techniques have not been fully explored within the context of CPATH, particularly in enhancing the utility in histopathological analysis and reporting.

For instance, PE techniques have been applied in the context of LLMs to refine textual responses and mitigate model hallucinations [Nori et al., 2023]. The integration of such strategies within histopathology, however, remains largely unexplored. Similarly, RAG techniques, which utilize external databases to enhance LLMs without extensive fine-tuning, have already been investigated in clinical settings [Gao et al., 2024, Unlu et al., 2024], but, to our knowledge, their integration in histopathology and combination with CPATH algorithms has not been directly addressed.

In this research, we aim to explore the integration of these recent strategies, specifically PE and RAG, into the CPATH framework. The objective is to develop a unified system that connects various CPATH tasks (including the inference of patient demographic attributes such as age and smoking status, and information retrieval) with a pipeline of generative AI, aiming to further using the outcome of the CPATH models as a fuel to the automatic generation of pathology reports from WSIs.

In this research, we aim to explore the integration of recent strategies, specifically PE and RAG, into the CPATH framework. The objective is to develop a unified system that connects various CPATH tasks — including the inference of patient demographic attributes such as age and smoking status, and information retrieval — with a pipeline of generative AI, aiming to leverage the outcomes of these CPATH tasks to fuel the automatic generation of pathology reports from WSIs.

Furthermore, considering the high computational demands of FMs in the text and image domains, this project also aims to explore the feasibility and effectiveness of utilizing such models on personal computers with limited resources. This approach seeks to determine whether it is possible to build valuable solutions with recent models without relying on costly cloud GPU services.

Finally, considering the existing barriers between pathologists, their practice, and the integration of algorithms into their daily routines, we also aim, as a proof of concept, to embed our solution within a user-friendly interface. This initiative is designed to facilitate access, testing, and potential implementation of this system in the daily activities of these professionals, thereby minimizing the need for technical programming skills.

The following section will outline the contributions this research seeks to make, based on the objectives previously discussed.

1.2 Contributions

Our interdisciplinary work integrates multiple areas of Computer Science, including CV, DL, NLP, and CPATH, and apply various techniques to the domain-specific area of histopathology. Thus, we believe that various fields can benefit from our work.

In the sections that follow, we outline our contributions by categorizing them according to their technical and practical aspects, and discuss their implications for both theoretical frameworks and clinical applications.

1.2.1 Technical Contributions

The technical contributions of this thesis support advances in the Computer Science community, particularly in CV, DL, and NLP. They are:

1. **Multimodal Learning Application** : The report generation system developed in this work utilizes DL architectures to manage learning tasks involving multiple data types, such as images and texts. We not only employed a pretrained model but also customized our own CLIP-like model, training it from scratch, to directly handle WSIs and retrieve textual descriptions. Thus, we believe that this integration can contribute to enhancing the understanding of these model's functionalities, particularly in the computational biology domain, where there is limited literature on such applications.

2. **Advancement in Computational Pathology** : This research explores emerging technologies within the domain of LLMs, such as PE and RAG, to enhance report generation from WSIs. This integration is, to the best of our knowledge, unprecedented in CPATH literature. Consequently, it has the potential to advance theoretical knowledge and refine existent techniques that use LLMs to handle medical datasets.

3. **Management of Limited Computational Resources** : This study also investigates the implementation of FMs under computational constraints, providing insights into efficient resource management for deploying computer-intensive techniques.

1.2.2 Practical Contributions

The practical contributions of this work aim to enhance pathology practices in clinical and educational settings:

1. **Enhancing Clinical Efficiency** : The automated report generation system developed in this project is designed to reduce the workload of pathologists by automating aspects of the diagnostic process, which can lead to faster and more consistent diagnostic outcomes, thereby improving patient care.

2. **Educational Tool Development** : Additionally, this system has the potential be used as an educational tool for trainee pathologists by providing a practical platform to enhance their learning process, facilitating access to refined domain-specific information.

3. **User Interface for Accessibility** : The integration of this system within a user-friendly graphical interface aims to reduce the complexity typically associated with advanced computational tools, increasing accessibility for pathologists without extensive technical backgrounds and promoting wider adoption.

Overall, this thesis has potential contributions to both the academic community by advancing the integration of multimodal learning systems and to the clinical field by offering practical tools that address key challenges in pathology.

1.3 Document Outline

This dissertation is structured to present the development and evaluation of our work in building an automatic report generation system for histopathology, focused on lung tissue. The chapter 1 covers the research objectives, contributions, and the motivation for integrating elements from Computer Vision, Deep Learning, Natural Language Processing, and Computational Pathology.

The chapter 2 reviews research projects conducted prior to this thesis, specifically focusing on age prediction from histology images and smoker classification. Insights and methodologies from these projects have significantly contributed to the report generation solution presented in this dissertation.

Then, in chapter 3, we present a literature review, multiple applications of DL in histopathology withing a cohort of the specific tools and techniques that were important for the development of our current work

The chapter 4 four examines the current workflow in histopathology, identifying its limitations and challenges. It explains the potential of CPATH and the motivation for developing our automatic report generation system.

Next, we dedicated chapter 5 to the exploratory analysis of our dataset. We start by defining Whole Slides and some of its properties. Then, we present information regarding data collection and preparation processes, and provide an overview of our datasetm focusing on Lung tissue. Next, the chapter addresses challenges such as data imbalance and co-occurrence of pathologies and discusses initial attempts to use a caption generation system for histopathological images.

In the methodology chapter (chapter 6), the system workflow is described, covering the inference, aggregation and generation stages. Each module's system architecture, data preparation, training and evaluation are discussed, along with the challenges and limitations encountered.

The results chapter (chapter 7) presents the outcomes of the research, including performance metrics for various models used in age prediction, smoker classification, and diagnosis retrieval. It includes parameter optimization and final model analysis, as well as the evaluation of the report generation module.

In chapter 8, we provide an overview of the app developed to embed our solution. This section discusses how the application integrates the various components of our system, facilitating user interaction and demonstrating a practical application of our research.

The discussion chapter (chapter 9) interprets the research findings in the context of existing literature and research objectives. It discusses the implications of the results, potential limitations of the study, and future research directions.

The chapter 10, presents our conclusion and perspectives for future works.

Finally, the appendices provide additional technical details and specific prompts used in the research. Appendix A covers topics such as caption generation metrics and the contrastive loss used in the CLIP model, while Appendix B includes details regarding all the prompts used to instantiate the agents in the GEN stage. We also included an Appendix C that provides concrete examples of the reports generated by our system.

Chapter 2

Publications During the Thesis

Within the scope of our research, we have co-authored three papers that contributed to both the methodological and theoretical foundations of the present work.

The first research project applied a multimodal approach to age prediction using methylation, gene expression, and histology images. This work led to the publication of two articles: Moraes et al. [2023] and Ribeiro et al. [2024]. The second project, on the other hand, investigated the impacts of smoking across various omics, including analyses of different tissues through WSIs, and culminated in one article is currently under review for publication [Ramirez et al., 2024].

The knowledge acquired and the results obtained from these projects have been instrumental in developing the independent modules for the report generation solution presented in this dissertation. In the following sections, we will detail each project and the valuable insights gained, which have been extensively applied in this work.

It is important to note that while both studies were conducted within the multi-omics field, our discussion will specifically concentrate on the histology aspect. This focus is due to our significant involvement in this area and its critical importance to the current work.

2.1 Age Prediction From Histology Images

2.1.1 Motivation

In our studies on aging, we aimed to address the challenge of determining an individual's biological age, a complex concept influenced by various factors such as genetics, environment and lifestyle. The identification of reliable biomarkers for biological age has been a long-standing research question, and we sought to contribute to this field by developing computational models that predicted biological age using multi-modal datasets. Specifically, we focused on lung tissue samples and explored three different data modalities: gene expression, methylation status, and histological images. Our goals were to compare the predictive power of these data types, assess whether their integration improved age predictions and address the technical challenges associated with incomplete and imbalanced age distribution in the dataset. By doing so, we aimed to gain a better understanding of the aging process and contribute to the development of strategies for healthy aging.

Next, we will individually discuss each of the two publications in the field of age prediction, their successes, limitations, and some results.

2.1.2 First Contact with Computational Pathology

In this first study [Moraes et al., 2023], we used a training set of 750 lung tissue samples or WSIs and a small test sample of 45 subjects, all obtained from the GTEx portal. Once downloaded, we segmented them into tiles of size $256x256$ at the maximum magnification level of $40x$. To split the data into training, validation and test sets, we built a custom stratification function.

The training of the models involved the use of various pre-trained convolutional neural networks (CNNs), such as VGG16, VGG19 and Xception, as backbones in our architecture, coupled with some top layers for the final regression.

We trained these CNNs with data augmentation and a sample weights module that prioritized less represented ages. After testing 173 different trials using the same validation set, we found that undersampling and data augmentation techniques such as rotations, reflections, and translations improved model performance. However, oversampling by repeating images did not yield any benefits and led to worse results.

Our analysis revealed that the primary factor hindering model performance was the inadequate representation of the neighborhood, particularly isolated donors from age groups with few or no neighbors. The lack of representativity in the neighborhood had a more significant impact on model performance than the number of samples per individual. We found that ensuring adequate representation of the neighborhood with a balanced distribution of donors across age groups is essential for obtaining reliable and high-performing image-based models. Therefore, relying solely on data augmentation and oversampling strategies in less represented age groups is not sufficient to achieve better performance in histological datasets.

We also hypothesized that another factor with a significant impact on the models' performance was associated with the training process using tiles. Since it was impossible to use WSIs as input for the neural networks, we had to segment them into small 256x256 pixel pieces. This process led to a significant loss of useful information related to the neighboring regions of each tile and the chosen resolution level. The first problem is due to the Field of View dimension, and the second to the Magnification level. For a visualization of these two concepts applied to WSIs, see Figure 5.1 and Figure 5.2.

2.1.3 Second attempt in solving the problem

In this second study, [Ribeiro et al., 2024], all the strategies we adopted were guided by the lessons learned from the previous work. The training and test sets remained almost unchanged, with only a slight variation in the test set in common between the different omics to obtain greater representation of subjects under 45 years of age.

The first significant change we made from the previous study to this one was the use of the HIPT neural network [Chen et al., 2022], developed to preserve the information present in the hierarchical structure of WSIs, generating embedding spaces rich in semantic information from histology images. With this, we aimed to mitigate the previously mentioned problem of information loss that occurs when training models only with isolated or decontextualized tiles.

Another important change we made was the implementation of the LDS method described in Yang et al. [2021]. The idea behind this method is to convolve a kernel (Gaussian, Laplacian, or triangular) over the original label space to produce a new label space where conventional weight calculation techniques, as used in classification problems, can be applied. In this way, we sought to circumvent the empty neighborhood problem we discussed earlier, where certain ages do not have a representative sample neighborhood, and weight calculation techniques, such as the inverse of the frequency of each class, are not necessarily effective.

Finally, another crucial step we took here was the use of k-fold cross-validation, with k=5, for selecting the best model. This was something we could not do in the first study due to the difficulty of training models with hundreds of tiles, but it was possible here since we were working with aggregated HIPT features. Thus, we were able to select a much more robust and generalizable model than the one selected in the previous study.

For a theoretical background on how HIPT and LDS work, refer to Appendix A.1 and A.2 respectively.

2.1.4 Results

In our first age prediction study, we encountered many obstacles that were better addressed in the second study. In Figure 2.1 (a), we can see the predictions of our best model generated in the first study. The values presented in the figure indicate low performance in various indicators such as R2, RMSE, MAE and MED. Additionally, we can see that this model has significant difficulty accurately predicting the ages of younger subjects, reflecting the impact of the imbalance we faced.

In the second study, we were able to mitigate several problems. Figure 2.1 (b) shows the results of our best histology model. Although both graphs are not directly comparable due to the slight variation in the test set we mentioned, we can see that the values of the different metrics are already at more acceptable levels, with a MAE of 5.57, MED of 4.3, and R2 of 0.70. Furthermore, we can see that the model performed more accurately for subjects under 45 years of age, so the slope of the curve is not as pronounced as in the previous model.

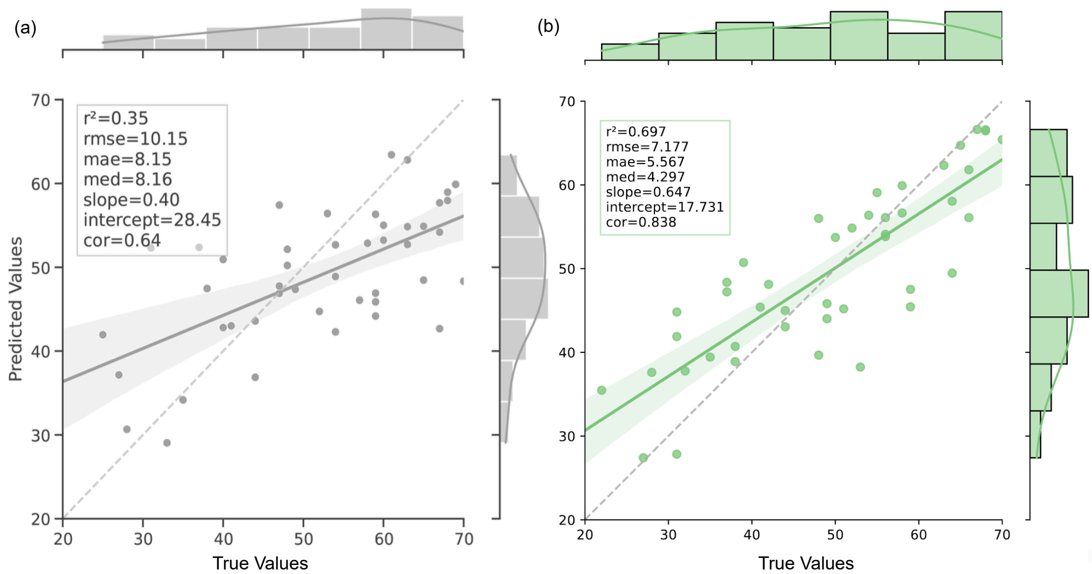


Figure 2.1: In the figure, the predicted versus true age values from the histology models of our two published age prediction articles are shown. Panel (a) displays results from the first article [Moraes et al., 2023], while panel (b) shows results from the second article [Ribeiro et al., 2024]. Although the results are not directly comparable due to slight variations in the test sets, the predictions in panel (b) are noticeably more robust.

2.2 Smoker Classification from Histology Images

2.2.1 Motivation

In this publication [Ramirez et al., 2024], we aimed to investigate the impact of cigarette smoking on human tissues at the molecular and histological level. Although smoking is a major cause of preventable mortality worldwide, and while it has been associated with accelerated aging and increased disease risk, the molecular mechanisms driving these effects in most tissues still remain unexplored. To address this knowledge gap, in this work, we used data from the GTEx project to systematically analyze the transcriptomic, epigenetic and histological changes induced by cigarette smoking across multiple human tissues.

2.2.2 Workflow

To investigate the impacts of smoking on human tissues, we focused our analysis on four distinct tissue types: Lung, Esophagus, Pancreas, and Thyroid.

As in our age prediction studies, we procured WSIs for these tissues and segmented them into non-overlapping tiles of size $512x512$ pixels. We divided the samples into training, testing, and validation sets, ensuring that the test sets remained consistent across tissues to facilitate comparative analyses.

2.2.2.1 Classifier Training

The objective was to classify each image as either 'Smoker' or 'Non-Smoker.' We identified class imbalance as a significant factor affecting model performance and conducted multiple trials to optimize the class weights in conjunction with cross-entropy loss.

We segregated the dataset into train and test subsets, with the test set consisting of a common set of donors (n = 85) across the four tissues, ensuring robust model comparisons. For each tissue, 80% of the training subjects were used for model training, and the remaining 20% served for validation. We employed the Xception convolutional neural network, modifying the top layers to include a Max Pooling layer, a Dropout layer, and a sigmoid activation function to estimate the smoking probability for each subject.

After approximately 100 trials per tissue to fine-tune parameters and hyperparameters, we trained an optimized model for each tissue using both test and validation sets and assessed their performance using ROC-AUC curves on the common test set.

2.2.2.2 Ex-Smoker Classification

This phase involved selecting ex-smokers and applying the optimal model from each tissue to classify these subjects. The goal was to statistically determine whether the tissue characteristics of ex-smokers were more similar to smokers or non-smokers, thus providing deeper insights into both the models and the sample.

2.2.3 Results

In Figure 2.2 (a), the ROC-AUC curves for each tissue model are displayed, assessing the effectiveness of our classification models. The Lung tissue model demonstrated superior performance with an AUC of 0.85, followed closely by the thyroid and pancreas models with AUC values of 0.74 and 0.73, respectively, and the esophagus model at 0.69.

In Figure 2.2 (b), the results of our analysis on ex-smokers using the optimal models for each tissue are presented. Generally, the models predominantly classify ex-smokers as non-smokers. However, an exception is noted in the case of pancreatic tissue, where a significant proportion of ex-smokers is classified as smokers.

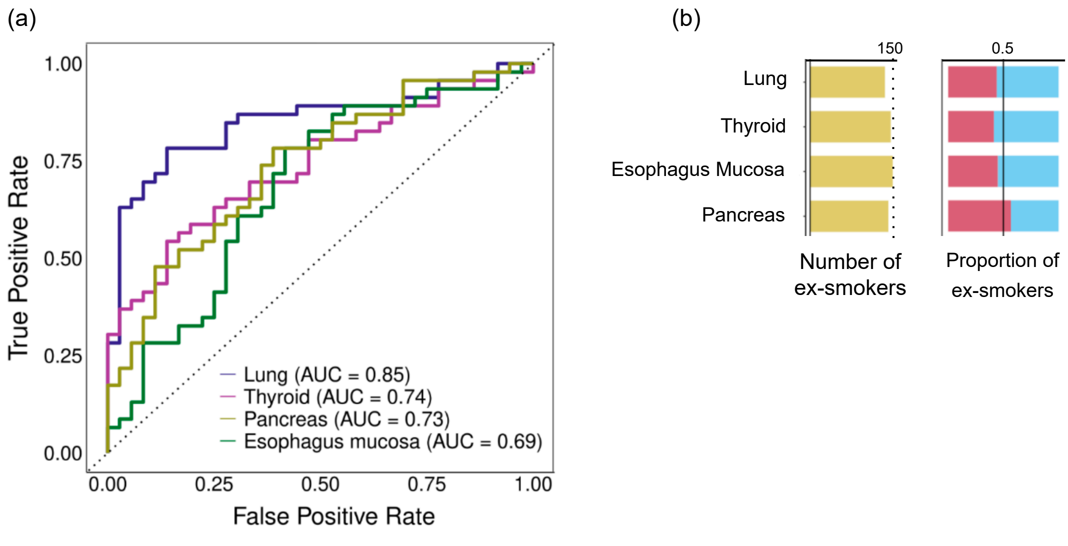


Figure 2.2: The ROC-AUC curves illustrate the performance of the best models by tissue.

2.3 Considerations

As previously stated, the discussion herein represents only a portion of these broader studies, which encompass more than just histological tissues. However, given our intensive involvement with histology and deep learning models, the focus of our discourse is well-founded.

Our work with histological images taught us how to manage their complexity, primarily due to their size, and how to derive comprehensive information from their hierarchical structure. Initially, we employed traditional CNNs such as VGG and Xception. Later, we transitioned to more advanced models like Vision Transformers, which incorporate the HIPT architecture.

Throughout these projects, our proficiency in deep learning has significantly advanced, particularly in adopting best practices for parameter optimization, such as k-fold cross-validation. This was facilitated by the utilization of HIPT features, which aggregate data at the WSI level, reducing the computational demands of our models. Additionally, confronting challenges such as imbalances in smoker classifications and age distributions highlighted the critical importance of a data-centric approach, which prioritizes the quality and preparation of input data over parameter optimization in order to improve the robustness of our results.

These insights and methodologies have been extensively applied in the current work and will be explored further in subsequent chapters. We will delve into working with WSI, the balancing strategies employed, and the implementation of the HIPT neural network across all modules of our solution.

Chapter 3

Literature Review

Digital Pathology (DP) is a subfield of Pathology focused on the acquisition, management, and interpretation of pathology information derived from glass slides [Ramamurthy et al., 2015]. It includes various technologies developed to digitize pathology slides and their associated metadata, addressing storage, analysis, and accessibility issues [Abels et al., 2019]. DP facilitates several types of imaging, each serving distinct diagnostic and research purposes, such as cytology, hematopathology, immunofluorescence, and high-resolution Whole Slide Images (WSI) in histopathology and dermatopathology.

Building on these foundations, Computational Pathology (CPATH) leverages advanced computational techniques to analyze and extract insights from pathology images. It incorporates machine learning algorithms to improve diagnostics, prognosis, and personalized treatment planning. This method enhances the depth of analysis beyond what manual review can achieve, aiming to increase the precision and effectiveness of medical assessments.

Although numerous studies exist for various CPATH tasks, the research presented in our dissertation explores a relatively novel task within this domain. To our knowledge, there are no existing studies directly related to the specific task of automated report generation in histopathology as detailed in our work. Our approach integrates and adapts various established tools and techniques from both inside and outside the field of computational pathology. Thus, this chapter aims to provide a comprehensive overview of these works and methodologies, outlining their objectives, applications, and their impact on our research.

3.1 Caption Generation

Caption generation (CG) systems are highly relevant to our study's focus on report generation, as they represent the closest analog to the tasks performed by report generation systems. These models fall under the category of Sequence-to-Sequence models [Neubig, 2017], processing an input sequence of image patches and outputting a sequence of textual tokens, and can also be named Vision Language Models (VLMs) since they connect both vision and language domains. Techniques in this area have evolved significantly over the years, transitioning from probabilistic models to deep learning-based models such as LSTMs and CNNs, and more recently to Transformer architectures. However, despite these advancements in model architectures, critical elements essential for the development of robust CG systems, such as token imbalance and inappropriate standard metrics, have been almost neglected in the scientific literature.

Next, we will explore how these models have been applied to histopathology and highlight some of their main limitations. For more details regarding VLMs, refer to the works of Hartsock and Rasool [2024], for a general overview of the VLM applications, and Zhang et al. [2024], for a mathematical perspective over the subject, including caption generation.

3.1.1 Applications in Histopathology

The application of CG systems in histopathology has just emerged in the past two years. Key studies in this area include Sengupta and Brown [2024], Lu et al. [2023], Guevara et al. [2023], and Zhou et al. [2024]. These systems not only generate captions but also share similar architectural frameworks, typically combining Vision Transformers with Large Language Models using attention mechanisms for effective multimodal fusion.

Particularly, the study by Sengupta and Brown [2024] has significantly influenced our research. This work adapts the VisionEncoderDecoderModel* from Hugging Face's Transformers library [Wolf et al., 2020] to process Whole Slide Images (WSIs) for generating pathologist's notes from the GTEx database as captions. The authors employed the HIPT architecture as the vision encoder and experimented with various decoder models based on BERT, including BioBERT, ClinicalBERT. Although BERT is traditionally an encoder-only model, it has been adapted here for use as a decoder in text generation within the Transformers ecosystem.

*Take a look on https://huggingface.co/docs/transformers/model_doc/vision-encoder-decoder for more details regarding this class.

3.1.2 Impact of Token Imbalance on CG Systems

Our experiments using the architecture and pre-trained weights from Sengupta and Brown [2024] uncovered significant limitations that hindered our progress towards project goals. In trials to generate captions for a lung image test set, we observed a consistent bias towards recurrent words and pheases, as we will discuss in the chapter 5. The captions predominantly reflected terms like 'congestion' and 'emphysema', which are frequently annotated by pathologists for this tissue type.

The issue of token imbalance in caption generation systems is a prevalent but under-explored problem in current research. Most studies on caption generation, both within the broader field and in medical-specific contexts, tend to overlook this issue, not recognizing it as a significant challenge that warrants attention. However, the work of Ding et al. [2023] stands out as they specifically address token imbalance within the domain of caption generation. They evaluate some common datasets that exhibit severe token imbalances and suggest alternative approaches to mitigate these problems. While their analysis does not extend to medical captions, they not only shed light on a typically ignored issue but also offer potential solutions that could be adapted to improve the reliability and fairness of medical caption generation systems.

3.1.3 Applicability of CG Metrics to Medical Contexts

Despite the remarkable advancements in CG techniques, the metrics used to evaluate Computational Pathology (CP) models have remained largely unchanged over the years. These metrics can be divided into two categories: n-gram-based and embedding-based *. Examples of the former include BLEU, METEOR, ROUGE, CIDEr, and SPICE, while examples of the latter are CLIPScore and BERTScore. In the work of González-Chávez et al. [2024], an extensive comparison between these metrics revealed that n-gram-based metrics are inappropriate for distinguishing between 'bad' and 'good' captions according to human judgment. Similarly, BERTScore and CLIPScore were found to struggle in making robust judgments, indicating that there is a risk in continuing to evaluate CG systems with

*N-gram based metrics like BLEU and METEOR assess image captioning by comparing the overlap of word sequences with reference captions, focusing on syntactic accuracy but often failing to capture semantic depth. Conversely, embedding-based metrics such as BERTScore and CLIPScore utilize embeddings from pre-trained models to evaluate semantic similarity between generated and reference captions, addressing limitations in capturing semantic richness inherent in n-gram based evaluations. For more details on how these metrics are calculated, check Appendix A.

these metrics. This risk arises not only because they do not accurately reflect the quality of a generated caption, but also because they might overlook issues such as overfitting.

The visualization presented in Figure 5.15 not only highlights the dominance of certain pathologies within each tissue type but also illustrates how less frequent conditions can be overshadowed, reinforcing biases twards specific terms. Such biases can, for instance, lead CG systems to the repetitive generation of phrases like 'pieces' in general or 'congestion' and 'moderate' in the specific case of lung, consequently skewing performance metrics such as Accuracy and BLEU, and resulting in overly optimistic evaluations.

In a study on report generation from chest x-rays, Harzig et al. [2019] question the effectiveness of traditional metrics like BLEU in medical contexts. They demonstrate that high BLEU scores could be achieved by merely repeating the same paragraph for all input images, suggesting these metrics might not accurately reflect the quality of pathologist's notes. Similarly, in our findings, which will be detailed in our results chapter, we observed comparable issues not only with BLEU but also with other metrics such as CIDER, ROUGE, and METEOR.

The study by Lu et al. [2023] points out significant limitations with commonly used evaluation metrics in the context of caption generation for histopathology. Despite advancements in techniques, the METEOR and ROUGE scores they reported fell below the thresholds necessary for practical deployment in clinical settings, confining their application to academic research. Furthermore, their work did not delve into the challenges of selecting appropriate metrics for medical contexts. Instead, they relied on traditional caption generation metrics without critically assessing their suitability or effectiveness for evaluating medical text, highlighting a gap in the adaptation of evaluation strategies to the unique demands of medical applications.

Conversely, Guevara et al. [2023] introduces an innovative approach by applying classification metrics, such as the F1-score, to evaluate captions based on the pathologies identified. This method is advantageous in medical settings because it more accurately reflects the biases and imbalances inherent in medical descriptions. In our study, we adopt a similar approach by utilizing classification metrics to evaluate the captions generated from WSIs, leveraging our access to both original and generated categories, as will be discussed in the results section.

3.2 Foundation Models

A foundation model (FM) is a system trained on vast datasets using unsupervised or self-supervised learning methods. This enables it to adapt to a wide range of downstream applications across various domains, often unrelated to the data on which it was originally trained. The primary applications of this technology are prominent in natural language processing, supporting tasks such as machine translation and automated text generation. In image processing, it has been utilized for zero-shot learning and image generation tasks.

Next, we will explore what embeddings are and how they are relevant to our work. We will then examine the applications of these FM models and embeddings in histopathology, and finally, we will discuss a limitation of these models related to data quality.

3.2.1 Embeddings in Deep Learning

In deep learning, the term ”embedding” is somewhat ambiguous, as it can refer to a transformation, a space, or an entity within that space. In the work of Tennenholtz et al. [2024], a subtle distinction is made in their definition of what they called ’domain’ embedding, represented by the function $E : \mathbb{V} \to \mathcal{W}$. Here, E denotes the process by which original data $v \in \mathbb{V}$ (such as words, users, or items) are transformed into embedding vectors, $E(v)$, within an embedding space, $\mathcal{W}$.

Practically, these transformations E are typically realized as specific layers within a larger architecture. These layers are designed to learn dense and meaningful representations of high-dimensional categorical data $v \in \mathbb{V}$, transforming it into vectors of continuous numbers $E(v) \in \mathcal{W}$ that preserves the semantic structure present in the original information. Ideally, two similar input instances will convert into vectors that are close to each other within this space. Such embeddings enable various downstream tasks including classification, regression, information retrieval, and more.

The relevance of Foundation Models in this context is significant. Trained on vast and diverse datasets, these models develop nuanced understandings of their source data, producing embedding spaces that capture rich context and relationships between different entities. These models are typically used as frozen backbones, from which embedding vectors are extracted for use in downstream tasks across various domains such as Natural Language Processing, Computer Vision, Graph Theory, and Time-Series Analysis ([Zhou et al., 2023], [Liang et al., 2024]). In this work, we applied this concept using the HIPT architecture, which will be explained next.

3.2.2 Application in Histopathology

In the field of histopathology, developing foundation models presents unique challenges due to the complexities associated with processing gigapixel Whole Slide Images (WSIs) and the scarcity of large, high-quality datasets. To address these issues, researchers have employed innovative approaches to data collection. Some of them, such as Ikezogwo et al. [2023] and Lu et al. [2023], have constructed extensive datasets by scraping digital resources including YouTube videos and PubMed articles. Others have utilized WSIs from well-known open datasets like The Genome Cancer Atlas (TCGA), utilized in developing the HIPT architecture [Chen et al., 2022]. A more recent development is the PLUTO architecture [Juyal et al., 2024], which merges scraped images with high-quality images from both public and private datasets to enhance training robustness.

Our research extensively utilizes the HIPT architecture, which serves as the foundational backbone across all modules of our solution. This architecture implements the DINO method for knowledge distillation using 10,678 WSIs from the TCGA database. It presents a three-tiered arrangement of Vision Transformers (ViTs): $ViT_{256} - 16$ at the cell level captures intricate cellular features from small 16x16 patches; $ViT_{4096} - 256$ at the patch level aggregates broader tissue contexts from 256x256 patches within larger 4096x4096 regions; and $ViT_{WSI} - 4096$ at the region level processes the largest patches, focusing on macroscopic tissue architecture and pathology distribution. This tiered approach ensures that once features from each of the 4096x4096 patches are extracted, a comprehensive representation at the WSI level can be achieved through final aggregation.

The significance of the HIPT architecture for our application lies in its ability to preserve both local and global information within the hierarchical structure of the WSIs, a crucial aspect that could be lost when working solely with individual tiles. This characteristic makes HIPT particularly effective for our specific use case in histopathology. For a more detailed explanation and intuition behind the functionality of this ViT, refer to Appendix A.1.

3.2.3 Concerns regarding data quality

The primary limitation of foundation models in histopathology relates to the reliability of the input data. A recent publication by Alfasly et al. [2024] presents a comparative study between foundation models used in this field, highlighting significant concerns regarding the quality of outputs from models trained with scraped data. The authors emphasize the necessity of high-quality, multimodal medical datasets for the effective development of vision-language foundation models in biomedicine. They state, ”To enable effective vision-language foundation models in biomedicine, high-quality, multimodal medical datasets are essential.” While Alfasly et al. [2024] have not conducted an extensive evaluation of all available models or their generalizability across a wider range of tasks, their critique is crucial and urges caution in the deployment of these models in real clinical settings. This underscores the importance of data integrity and quality in training robust and reliable foundation models for histopathology.

3.3 Cross-Modal Retrieval in Histopatology

Cross-modal retrieval (CMR) is a technique in artificial intelligence that enables the retrieval of information across different types of data, such as text, images, audio, and video [Li et al., 2023]. This process involves aligning and comparing data from these different modalities to extract relevant information and insights. CMR enhances user experience by allowing more comprehensive access to information and facilitating deeper understanding through the integration of various data types. It is commonly used in applications like search engines, digital libraries, and multimedia systems where users can, for example, input a text query and retrieve related images or videos. This capability supports a wide range of practical applications from educational tools to advanced research databases and can be extended to the biomedical domain as well.

In our study, which focuses on generating histopathology reports from WSIs, some investigations in the CMR field applied to pathology have proven particularly relevant. Notably, Ikezogwo et al. [2023] developed a dataset of 1 million images sourced from educational videos on YouTube, PubMed papers, and other educational resources. They employed contrastive learning to train a system that aligns text-image pairs for tasks such as zero-shot classification and text retrieval. Similarly, Lu et al. [2023] constructed a dataset from educational content and papers to facilitate the alignment of text and image modalities using contrastive learning techniques. Furthermore, Huang et al. [2023] compiled a dataset from medical Twitter content, applying the same contrastive loss to align text and image pairs.

All these studies utilize a framework based on the Contrastive Language–Image Pre-training (CLIP) model [Radford et al., 2021]. CLIP employs dual embedding layers, $f_I(\cdot)$ for images and $f_C(\cdot)$ for texts/captions, to extract latent representations of text-image pairs. Alignment is achieved by minimizing the cosine distance matrix's main diagonal square between these vectors, employing contrastive loss to optimize this process. This technique provides a simple approach for integrating multimodal information and has shown interesting results such as those mentioned in the previous paragraph, which are specific applications, and in its own paper where it demonstrated classification capabilities analogous to ImageNet models, but using Zero-Shot learning. This demonstrates its ability to generalize without the need for fine-tuning.

3.3.1 Limitations regarding WSI

While the aforementioned studies leverage innovative methods for aligning text and image data through scraping, they do not address the specific challenges associated with WSIs, which are crucial for our work in pathology report generation. This gap is significant, as WSIs contain a wealth of detailed histopathological information that could be lost if the images are segmented or not handled properly.

As we will discuss further in the methodology chapter, to effectively work with WSIs using CLIP-like models, we employed the HIPT as our Vision Transformer (ViT) to extract image embeddings that represent the entire aggregated WSI. We then developed a complete module that utilizes CMR to retrieve diagnoses based on these image slides. In this context, the CLIP model architecture proved essential for the continuation of our work, as the most critical module of our solution, the diagnostic retriever module, relies entirely on this architecture and forms the foundation upon which all subsequent stages are based.

3.4 Large Language Models in Clinical Practice

Clinical Natural Language Processing (NLP) is a subfield that intersects NLP and Biological Natural Language Processing (BioNLP), focusing specifically on the analysis of clinical textual data. This area has been an active field of research since the early 1990s and has evolved in line with broader trends in the NLP field. Initially, Clinical NLP employed rule-based algorithms; for example, the Canonical Phrase Identification System (CAPIS) from 1991 utilized matching index rules to extract findings from medical free-text [Lin et al., 1991], and the Natural Language Understanding System (NLUS) from 1996 combined graph theory with linguistic analysis to encode free-text medical records [Gundersen et al., 1996]. Although rule-based approaches remain prevalent, there has been a significant shift towards adopting AI methodologies, particularly since the 2000s [Wu et al., 2020].

The year 2016 marks a significant milestone with the adoption of deep learning-based approaches for Clinical NLP challenges. Since then, an increasing number of publications have demonstrated the applications of various deep learning architectures in NLP tasks such as text classification, named entity recognition, and relation extraction [Wu et al., 2020, 2022].

In the landscape of NLP, Large Language Models (LLMs) have proven their value in both BioNLP and Clinical NLP, with architectures such as BioLinkBert, PubMedGPT, PubMedBERT, BioGPT, and Med-PaLM gaining prominence. Specifically, BioGPT has been shown to outperform alternative biomedical language models in tasks such as Query Answering (Q&A), relational extraction, and named entity recognition [Raiaan et al., 2024]. Furthermore, researchers have shown that LLMs, such as ChatGPT, can successfully complete exams such as the United States Medical Licensing Exam (USMLE) [Kung et al., 2023], demonstrate utility in routine medical tasks including information extraction from electronic health records, support in literature searches, and assistance in medical writing regarding style and format [Biswas, 2023], and be employed in specialized areas within healthcare, like dental telemedicine [Eggmann et al., 2023] and radiology [Jeblick et al., 2024], to enhance patient-centered care and services.

Given that Pathology is a critical domain in Medical Diagnosis, it stands to benefit significantly from the advancements in Clinical NLP. Large Language Models (LLMs) can be utilized in pathology for various purposes, including as educational tools, assisting in diagnostics, generating reports, and extracting information [Cheng, 2024]. However, advancements in this area are still very recent.

3.4.1 Artificial Hallucinations

One of the major problems with LLMs, well known to their users and documented in the literature, is hallucination ([Liu et al., 2024], [Yan et al., 2024]). Hallucinations in LLMs generally manifest as generated content that is nonsensical or unrepresentative of the provided source, often due to errors in encoding and decoding between text and representations. This issue is particularly problematic in fields such as Clinical NLP, where an error in diagnosis can lead to a patient receiving incorrect treatment, further compromising their clinical condition.

Among the various alternatives available to address this problem, two are especially relevant to us, given their minimal computational resource requirements to be implemented:

Prompt Engineering (PE) and Retrieval Augmented Generation (RAG). These approaches will be discussed in the following sections.

3.4.2 Prompting

Prompting is a technique used in the realm of LLMs where specific inputs are provided to the model to guide its output. This method, known as Prompt Engineering (PE), is crucial for developing effective prompts that enable LLMs to better address specific tasks.

In Nori et al. [2023], the authors presented a specialized version of GPT-4 tailored for medical Question & Answering (Q&A) scenarios, which demonstrated superior performance over fine-tuned models such as Med-PaLM 2 [Singhal et al., 2023] across various benchmark datasets. Another study by Zheng et al. [2023] explored the influence of prompts that embody different social roles on model performance, concluding that prompts reflecting the relationship between speaker and listener often yield improved outcomes across diverse domains, including medicine. These studies underscore the effectiveness of a robust PE strategy, not only in reducing hallucinations in domain-specific applications but also in reducing the need for computationally expensive fine-tuning processes typically required for training LLMs with a large number of parameters.

3.4.3 Retrieval Augmented Generation

According to the review by Gao et al. [2024], Retrieval-Augmented Generation (RAG) represents a significant advancement in natural language processing by incorporating external knowledge bases to improve the responses of large language models (LLMs). RAG primarily addresses the limitations of LLMs, such as generating incorrect or outdated information by dynamically retrieving relevant document segments based on the semantic similarity of the query. This integration not only enhances the accuracy and relevancy of generated responses but also facilitates real-time knowledge updating and domain-specific information incorporation, making RAG particularly valuable in fields requiring precise and current knowledge as biomedicine.

In the landscape of RAG, we can find a variety of strategies aiming to optimize the retrieval and generation processes. Still in the mentioned work, we can find that these techniques range from Naive RAG, focusing on basic retrieval-read mechanisms, to Advanced and Modular RAG, which introduce sophisticated methods for improving retrieval quality and model flexibility. Advanced RAG incorporates pre-retrieval and post-retrieval optimization techniques to enhance the relevance and efficiency of the retrieved information. Modular RAG, on the other hand, adds flexibility through the introduction of specialized modules that can be reconfigured or substituted based on specific needs, thus supporting a more adaptable and effective framework.

RAG finds applications across several domains, primarily enhancing the performance of systems in tasks that demand extensive and up-to-date knowledge. It is particularly beneficial in areas such as customer support, where providing accurate and current information is crucial, and in domains like healthcare and legal services, where specialized knowledge is required. RAG also significantly contributes to educational technology, where it can provide detailed explanations or updated content in response to student inquiries. Additionally, its application in enhancing chatbot interactions makes it a valuable tool in improving user experience by providing more contextually appropriate and factually accurate responses.

Within the field of Clinical NLP, the applications of RAG techniques are relatively recent, dating back no more than two years. In a recent study, Miao et al. [2024] demonstrated a practical application in nephrology by utilizing RAG techniques to specialize ChatGPT for supporting the education of new professionals and aiding in clinical decision-making. In another work, Quidwai and Lagana [2024] introduced a RAG-based chatbot framework specifically designed to enhance personalized cancer treatment for Multiple Myeloma. Furthermore, in Unlu et al. [2024], the authors explored the use of RAG with GPT-4 to streamline subject screening in clinical trials, utilizing clinical notes to assess patient eligibility based on inclusion and exclusion criteria.

In the context of this work, to specialize a LLM in lung histopathology, we have developed a chatbot agent equipped with memory that utilizes interaction history to dynamically enhance and refine its responses. This implementation is an example of the application of the Modular RAG strategy. The details of this process will be explored further in the methodology section.

Chapter 4

Motivation

The motivation behind our research stems from an understanding of the various challenges encountered in histopathology, including logistical complexities and difficulties associated with implementing AI technologies. In what follows, we will first address the current logistics and associated problems, then delve into the details of our motivation, and finally, provide an overview of our proposed implementation for this work.

All the domain-specific information regarding histology, its logistics, workflow, challenges, and key points, were taken from Bancroft et al. [2018].

4.1 Current Workflow in Histopathology

The process begins with the collection of tissue specimens from the patient through surgical procedures or biopsies. These specimens are promptly immersed in a fixative solution to halt autolysis and preserve cellular integrity. Upon fixation, they are transported to the laboratory where each sample is tagged with a unique identifier to ensure traceability. At the laboratory, the specimens undergo tissue processing, which includes dehydration, clearing, and impregnation with paraffin wax. Once embedded in wax, thin sections are sliced and placed on slides. In the final treatment step, these slides are stained to highlight cellular details. From this stage, the slides can be digitized using whole-slide scanners and are also scrutinized under a microscope by a pathologist for diagnostic analysis. In Figure 4.1 (a., b., and c.) we can see an schema of this workflow.

4.2 Limitations

Given the described workflow, several limitations can be identified. Initially, there is the clear problem of the time it takes for the samples to reach the pathologist for analysis. Moreover, once the samples arrive at the laboratory and undergo all the necessary chemical treatments that enable their examination, pathologists can become overwhelmed with a high volume of cases [Cooper et al., 2023]. This leads to an increased workload and potential delays in diagnosis, a consequence of the rising number of cases and the scarcity of trained professionals to handle these tasks.

In addition to the direct challenges arising from the current logistics of this field and the shortage of skilled professionals, there is also the issue that manual handling of specimens and subjective interpretation of slides by pathologists can introduce errors and variability into the analysis and diagnosis [Kartasalo et al., 2021]. This lack of consistency can further affect the accuracy and reliability of diagnoses.

4.3 CPATH Promises

In recent years, advances in the fields of computer vision and artificial intelligence have shown successful applications in the realm of histopathology through CAPTH, engaging in tasks such as cancer detection, tissue segmentation, cell segmentation, survival analysis, among others [Cooper et al., 2023], with performance comparable to that of trained pathologists [van der Laak et al., 2021].

Given the various challenges faced by pathologists and other professionals involved in the traditional and current workflow of histopathology, and given the advances documented in the scientific literature in the field of CPATH, it is natural to consider the benefits and improvements that the implementation of such systems promises to offer to pathologists and patients, including workload reduction, increased productivity, greater precision in detecting abnormalities associated with images, and improvements in telepathology systems.

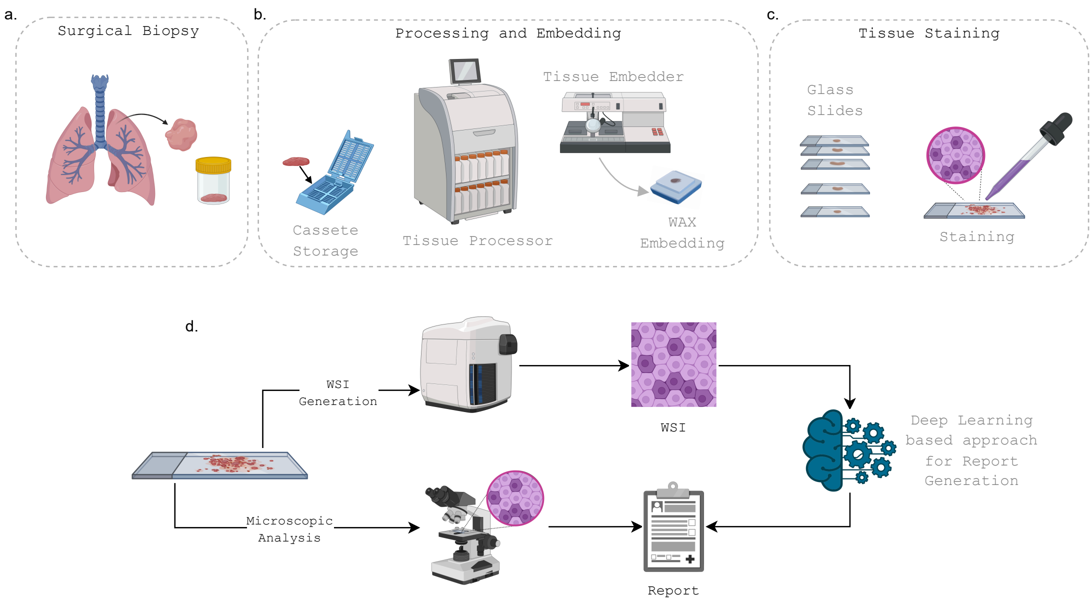


Figure 4.1: Overview of the current workflow in histopathology. The top figure shows the main stages of this workflow from sample collection (a) to tissue staining (c). In figure (d), the pipeline for generating pathological reports manually (d.1) and automatically (d.2) is depicted.

4.4 Implementation Challenges

Despite the promising results of CPATH and all the potential benefits associated with its clinical adoption, the field of histopathology remains largely reliant on manual processes and traditional microscopic analysis of glass slides, with few algorithms implemented and currently in use. As noted in the introduction of this work, various authors have addressed this issue with more academic rigor, and therefore it is not our intention here to delve into the details of this discussion. Interested readers can thoroughly explore these works ([Cooper et al., 2023], [Abels et al., 2019], [Jahn et al., 2020]). In our understanding, the problems identified can be classified according to their origin. Thus, we have problems (i) that originate from the limitations of CPATH technologies, and problems (ii) that arise from obstacles within the field of histopathology itself. Problems of type (i) include lack of high-quality data, poor annotations, training biases, image normalization issues, stain variability, among others. Problems of type (ii) are characterized by the lack of understanding of Deep Learning mechanisms, substantial IT infrastructure costs, ethical concerns regarding patient confidentiality, difficulty in establishing quality control for these solutions, concerns about the future of the profession and employability, among others.

4.5 Reasons Behind This Project

In this work, our primary motivation was to leverage recent advances in computer vision and deep learning within the field of CPATH with the aim of enhancing clinical histopathology practices. Considering the high-quality images and annotations provided, we selected the GTEx Portal as our data source to address challenges related to data quality. Furthermore, we aimed to create a solution that extends beyond the academic sphere where the focus typically remains on refining deep learning architectures and achieving benchmarks. By concentrating on real-world challenges that pathologists face daily, our goal was to develop a system that is straightforward, computationally affordable, and user-friendly, thus minimizing both theoretical and practical barriers to adoption.

With these objectives in mind and recognizing the current constraints of traditional histopathology workflows, we identified an opportunity to innovate with a system for the automatic generation of histopathology reports from whole slides. To the best of our knowledge, no similar systems currently exist in this domain. If successfully developed, such work could significantly enhance clinical accuracy and productivity during the analysis phase (Figure 4.1 (d)). Additionally, such a system holds potential utility in educational contexts, where it can be used to provide support and training for emerging professionals in the field.

4.6 System Overview

The architecture of our system is segmented into three principal stages: Inference (INF), Aggregation (AGG), and Generation (GEN), each designed to streamline the processing and analysis of data.

In the INF stage, we have developed three distinct modules, each targeting specific inference objectives: an Age Prediction (AP) module, a Smoker Classification (SC) module, and a Diagnosis Retrieval (DR) module. The DR module performs a similarity search among images to identify the most relevant pathologist notes for a given WSI. Although these modules operate on the same data source, they function independently, allowing for targeted improvements to one without affecting the others. This modular design also supports the addition of further inference components based on specific user requirements.

Following INF stage, the AGG stage is responsible for combining the outputs from the previous stage into a caption that includes all the demographic and pathological information associated with the patient and their image.

The processed caption is then advanced to the GEN stage. Utilizing three LLMs, the caption is enriched with additional insights drawn from academic sources such as PubMed, bioRxiv, and various university databases. This enrichment utilizes techniques such as PE and RAG to ensure that the LLMs produce content that is accurate and aligned with the intended research outcomes.

Specific details on the system's functionality and operational mechanics will be provided in the methodology section (chapter 6).

Chapter 5

Data Analysis

Throughout this project, numerous crucial decisions such as selecting training and testing instances, implementing strategies to mitigate data imbalance, and even choosing the model architecture, were guided by challenges revealed during the data analysis phase. Therefore, we believe it is essential to dedicate a chapter to this theme.

To mirror the sequence of our exploratory process, we begin this chapter by introducing the types of images and the database utilized. This is followed by a Data Analysis section, which outlines the technical challenges encountered and the strategies employed to address them. In particular, this section will focus on analyses related to lung tissue, which has been a central focus and utilized across all modules of our system.

This chapter is crucial to this dissertation as it lays the groundwork for understanding the strategies that will be detailed in the subsequent Methodology chapter.

5.1 Whole Slide Images: A Quick Overview

Whole Slide Images (WSIs) are gigapixel images that can measure up to 100,000 x 100,000 pixels. Due to their high resolution and fixed scale, these images inherently display a hierarchical structure when viewed at different levels of magnification, ranging from individual cells to complex tissue arrangements. To facilitate access and processing of these images across multiple levels, they are typically stored in a pyramid format. In this format, the same image is saved at various resolutions within a single file, stacked with the highest resolution image at the bottom and the lowest resolution image at the top, in an organization that resembles a pyramid. An illustration of this structure is shown in Figure 5.1.

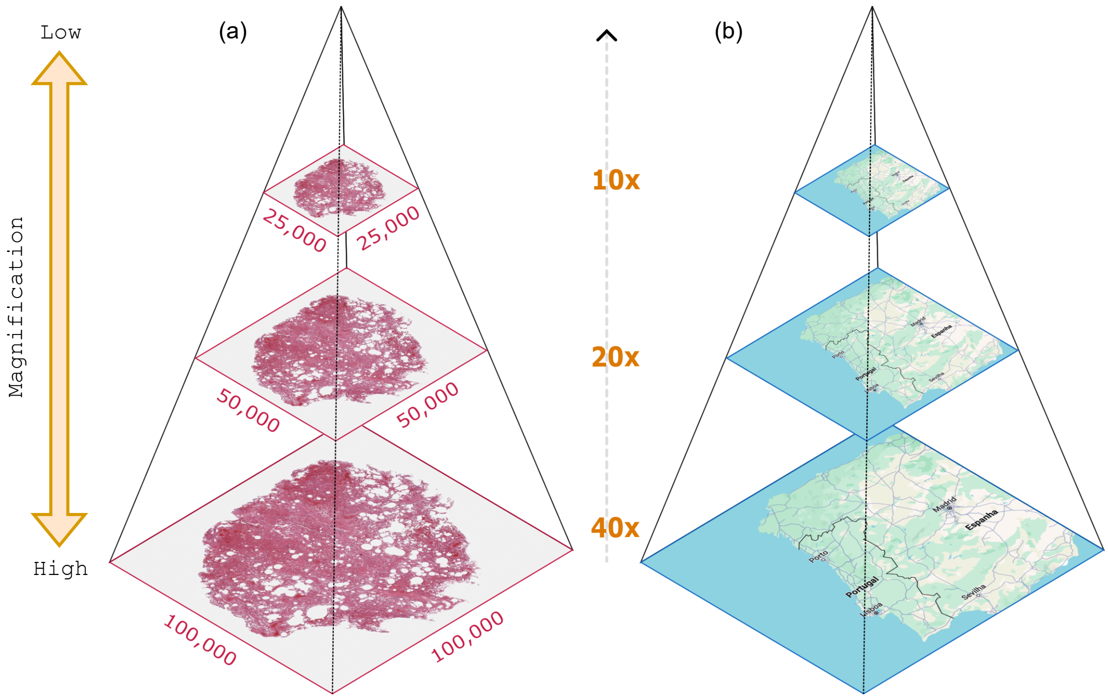


Figure 5.1: This illustration demonstrates the concept of magnification using a pyramid that displays the same WSI at different resolutions (a). In order to enhance the comprehension, an image pyramid featuring the map of Portugal is also presented (b), making an analogy with the downsampling process applied in (a).

The term resolution can sometimes be tricky because there are no standard definitions for it. In this work, when we refer to resolution, we mean the number of pixels within a given area, which denotes the level of detail, or granularity, that is visible in an image. Therefore, if we assume that a particular scanner can achieve 40x magnification, the most detailed and highest resolution level of the scanned WSI will be at this 40x, forming the basis of the image pyramid. If at this level the image dimensions are 150,000 x 150,000 pixels, downsampling it by a factor of 2x will result in an image size of 75,000 x 75,000 pixels at a 20x magnification. This process continues successively according to the rule $m = \frac{M}{d}$, where M is the original scanned magnification, d is the downsampling factor, and m is the new magnification level, forming the image WSI pyramid we see in Figure 5.1 (a). In Figure 5.1 (b) we present an analogy in the domain of cartography where, instead of a histological slide, we have the map of Portugal being downsampled twice.

Another important concept in working with WSIs is the Field of View (FOV). After selecting one of the available resolutions from the stacked pyramid, a pathologist zooms in and out of different parts of the WSI to identify regions of interest. The process of adjusting the FOV by zooming in and out on a WSI is illustrated in Figure 5.2 (a). This figure not only demonstrates the dynamic changes in FOV but also highlights the complexity inherent in a histological slide. Each time the FOV is narrowed, new topological features and interconnected elements emerge that were not visible at higher levels of magnification. This complexity is analogous to cartographic maps, where zooming in from the scale of a globe to the detailed granularity of ecosystems within a specific region reveals new details and relationships, as depicted in Figure 5.2 (b).

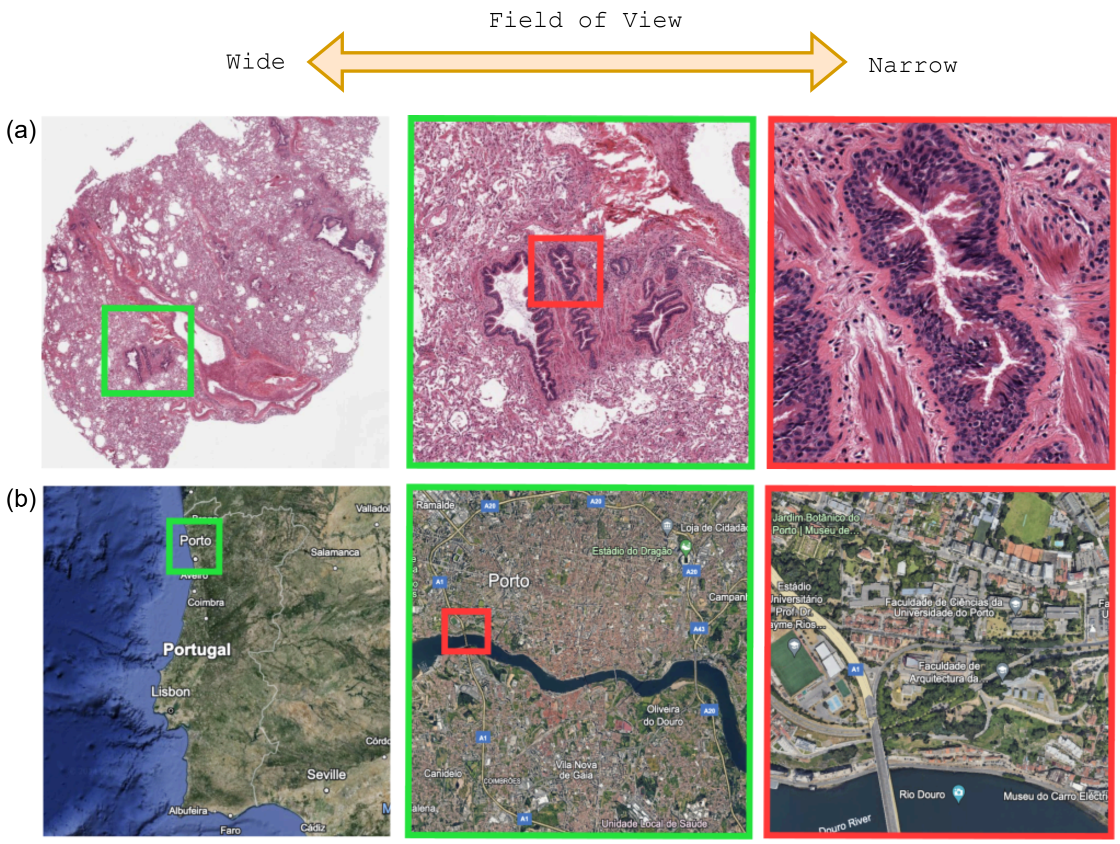


Figure 5.2: Here, the notion of Field of View (FOV) is illustrated with two layers of zoom on the same WSI (a). To clarify this concept, the same zooming process is applied to the map of Portugal (b), progressively narrowing the focus towards the Faculty of Sciences of the University of Porto.

As previously discussed, the images we work with are of considerable size, making it impractical to directly input them into image models. Consequently, it is necessary to segment these images into non-overlapping regions termed patches or tiles. As inferred from Figure 5.1 and Figure 5.2, selecting a magnification scale and a Field of View level invariably results in the loss of critical information within the hierarchical structure of the images. Specifically, as illustrated in Figure 5.2 (b), choosing the first image allows for an understanding of the connections between Portugal and Spain and among their major cities, yet it omits the finer details of the connections between the neighborhoods of these cities and the localities within each neighborhood, as depicted in subsequent images.

Recent studies have aimed to mitigate this issue. Notably, the HIPT architecture introduced by Chen et al. [2022], which we explored in subsection 3.2.2, addresses this challenge. This architecture is designed to generate an embedding space that accommodates various magnification levels, taking advantage of the pyramidal structure inherent to WSI. We will further discuss this architecture and its application in feature extraction from gigapixel images in the subsequent sections.

5.2 Data Collection and Preparation

Our dataset originates from The Genotype-Tissue Expression (GTEx) project [Lonsdale et al., 2013], a public resource designed for researchers studying tissue and cell-specific gene expression and regulation across individuals, development stages, and species. This portal includes data from three National Institutes of Health (NIH) projects: Adult GTEx, Developmental GTEx (dGTEx), and Non-human Primate Developmental (NHP-GTEx). In this work, we exclusively utilized the Adult GTEx dataset, which contains thousands of tissue samples from various organs of post-mortem subjects.

The data collection pipeline from the GTEx database, as well as its preparation for Deep Learning algorithms, involved the following steps, which were executed using Python and Bash scripts:

Image Download. In this stage, for a selected tissue type, we made a request per subject to the GTEx endpoint, which returned the WSIs in high-resolution .svs format.

Feature Extraction. Given that these images are hundreds of MBs in size, to avoid overloading the computer's hard drive, we performed feature extraction from each image using the HIPT architecture immediately after downloading, followed by its deletion from the system. The features, in the form of PyTorch tensors along with their thumbnails, were stored locally for each subject. The feature extraction process involved initially segmenting the images into patches of size $4096x4096$, the required size for using HIPT. Subsequently, from each of these patches, we extracted three types of features: $[CLS]_{4092}^{WSI}$, $[CLS]_{m256}^{WSI}$, and $[CLS]_{4092}^{WSI}$.

A summary of these two steps can be seen in Figure 5.3.

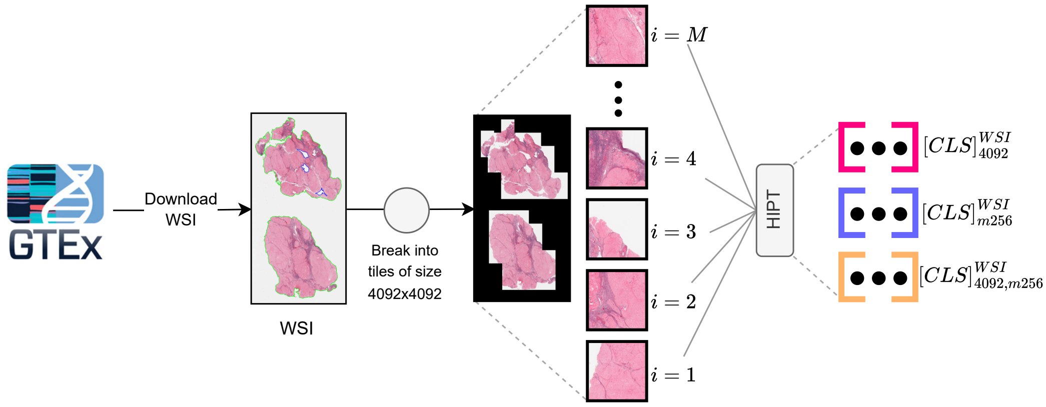


Figure 5.3: Initially, Whole Slide Images (WSIs) are downloaded from the GTEx portal. Subsequently, these slides are prepared for High-Performance Image Processing Technology (HIPT), which involves segmenting the slide into patches of 4096x4096 pixels. Finally, feature extraction is conducted using the mentioned architecture, yielding a total of three types of features.

5.3 Data Split

The initial stage of our system consists of three modules as previously mentioned: Age Prediction (AP), Smoker Classification (SC), and Diagnosis Retrieval (DR). Each module requires training from scratch using features collected from HIPT. It is important to note that not all data are applicable across all modules due to specific requirements and constraints of each. Figure 5.4 illustrates the filtering process used to prepare the training and testing sets for each module. Following this process, we obtained training sets comprising 727 images for AP, 635 for DR, and 254 for SC. Additionally, a common test dataset consisting of 72 subjects was established.

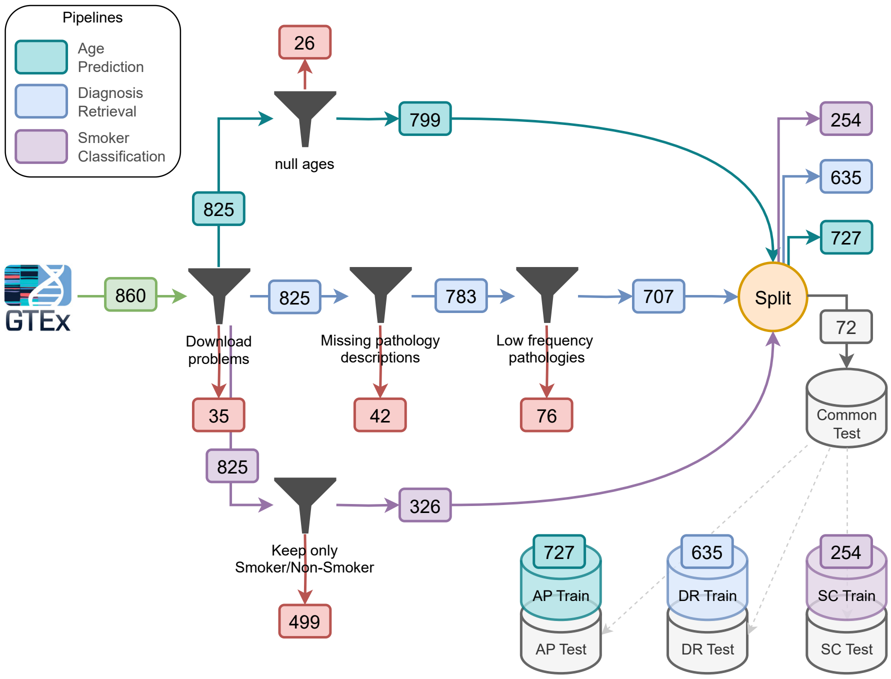


Figure 5.4: This figure illustrates the data selection pipeline used for selecting the training and testing sets for each module in the INF stage. We started with 860 lung WSIs from GTEx; however, due to download constraints, only 825 remained, resulting in the loss of 35 images. These 825 images then underwent various specific filters across three pipelines, leading to the final training sets for each modality and a common test set of 72 subjects. The numeric boxes are colored according to its module color (AP, DR, or SC), with red indicating data removal.

5.4 Dataset Overview

On the GTEx portal, it is possible to download a CSV file containing comprehensive information about the slides available in their system, as well as demographic and clinical data associated with each sample. This section begins with an overview of this dataset, subsequently focusing specifically on the lung tissue selected for this study, and the complexities involved in handling this type data.

Column	Content
Tissue Sample ID	GTEX-111CU-0326
Tissue	Lung
Subject ID	GTEX-111CU
Sex	Male
Age Bracket	50-59
Hardy Scale	Ventilator case
Pathology Categories	Macrophages, pneumonia
Pathology Notes	”... consistent with diffuse alveolar damage”

Table 5.1: Example of column contents in the GTEx database. All these information are openly available in their portal.

The GTEx portal provides a total of 25,713 slides distributed across 40 different human body tissues. Of these, 8,122 slides show some pathology, while 17,591 have no reported pathology associated. This leak of significant pathologies in the dataset is expected since GTEx is a dataset of healthy subjects.

5.5 Lung Tissue

As seen in Figure Figure 5.5, the GTEx portal has a considerable number of lung samples compared to other tissues, with a total of 860 slides in the current release. Figure Figure 5.6 shows that the lung slides have the highest number of associated pathologies, counting 19 different classes and 1 additional class that represents the abscense of pathologies.

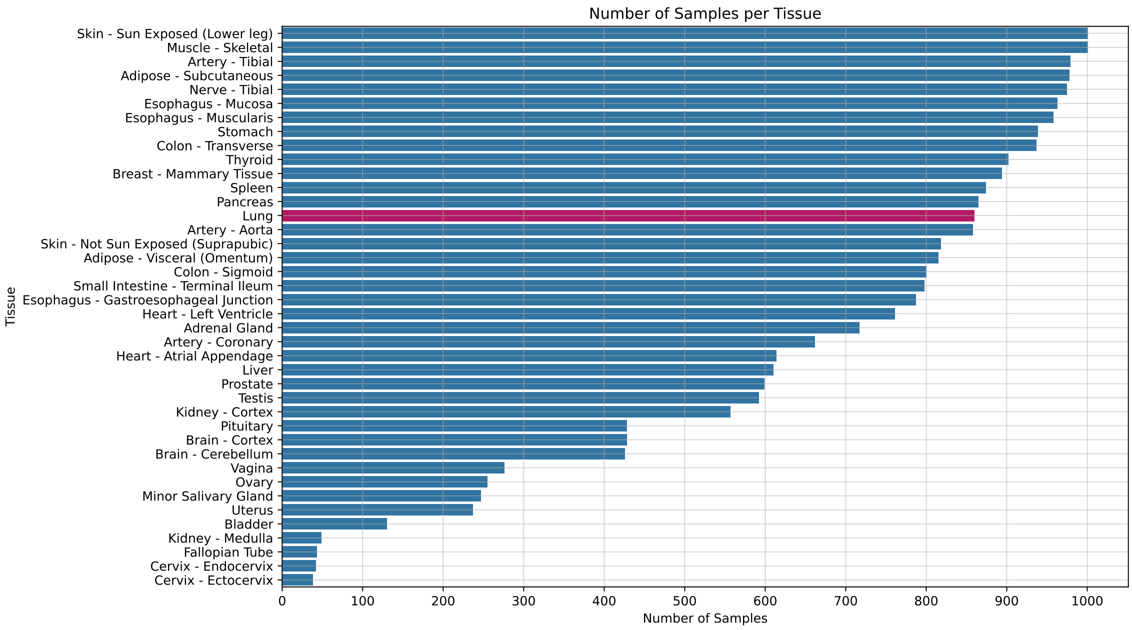


Figure 5.5: Number of samples per tissue in the GTEx database. Tissues with the most samples have up to 1000 images, while those with the fewest have fewer than 50 images.

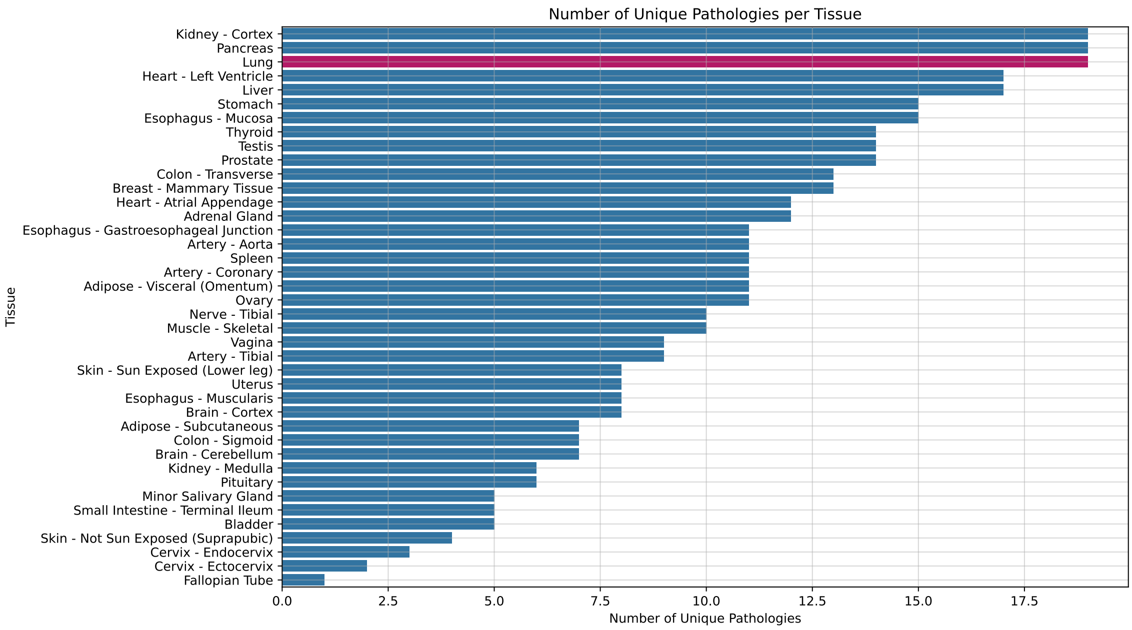


Figure 5.6: Number of pathologies per tissue in the GTEx database. The maximum number is 19 pathologies for Lung, Pancreas, and Kidney - Cortex, and the minimum is 1 pathology for Fallopian Tube.

In the following sections, we will address the impact of the distributions present in this dataset specifically regarding the variables of interest for our three modules: CG, AP, and SSP.

5.5.1 Data Imbalance

The dataset exhibits severe imbalances across different dimensions. Of these 860 slides, there is a significantly higher concentration of older subjects compared to younger ages, as we can see in Figure 5.7.

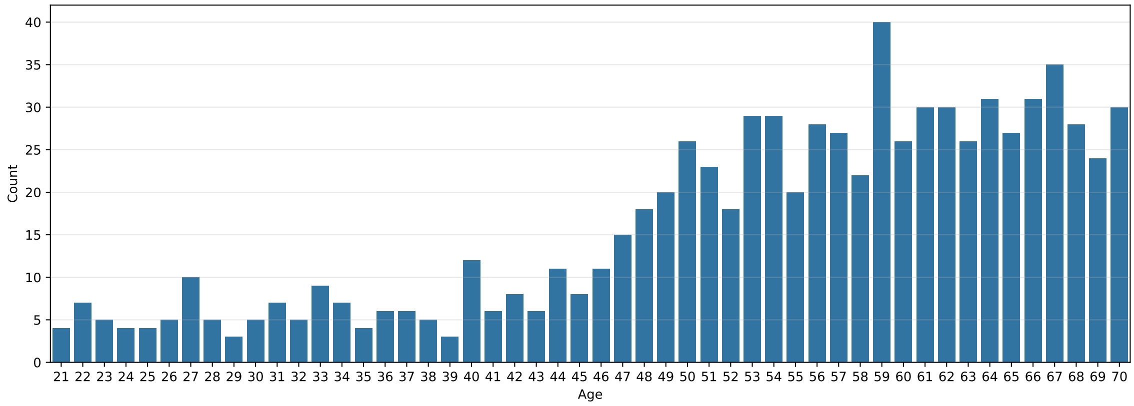


Figure 5.7: Number of occurrences of age across lung samples.

Regarding the pathologies, there is also a strong imbalance both within and between classes. In Figure 5.8, we can observe these two types of imbalances. The intra-class imbalance is straightforward, as we have a total of 860 samples; for congestion, for example, there are 379 positive cases and 481 negative cases, and so on for other classes. The inter-class imbalance is directly visible in the figure because we clearly see that some pathologies such as congestion, emphysema, and fibrosis appear in many more cases than others like pigment, inflammation, heart failure cells, and the following.

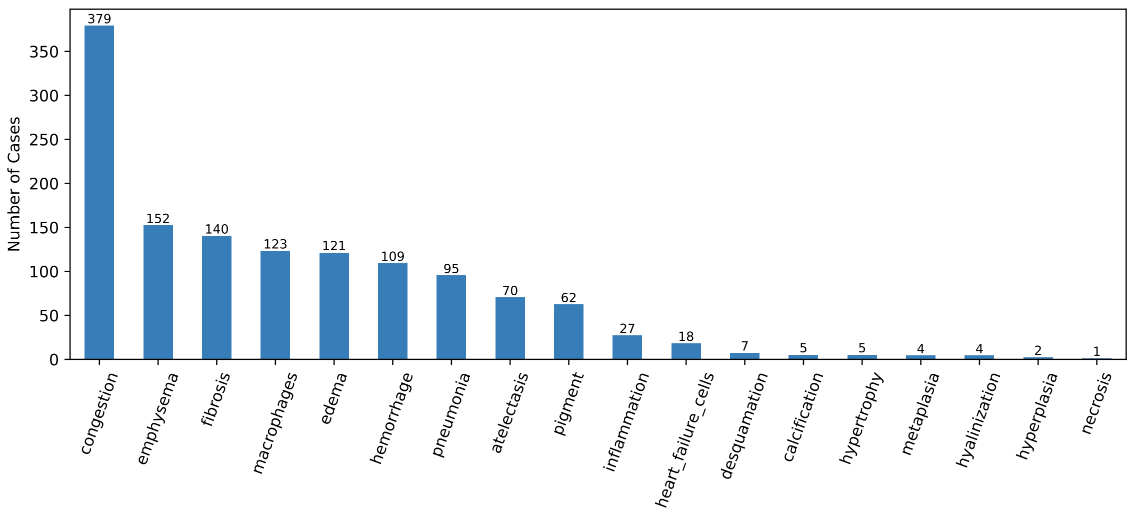


Figure 5.8: Number of occurrences of each pathology associated with lung tissues.

In addition to these intra-class and inter-class imbalances shown in Figure 5.8 for different pathologies, given the age imbalance coupled with predispositions associated with the aging process, most pathologies are found in samples from subjects aged over 45 years, as seen in Figure 5.9, where the area of greatest concentration is in the upper right quadrant of the matrix.

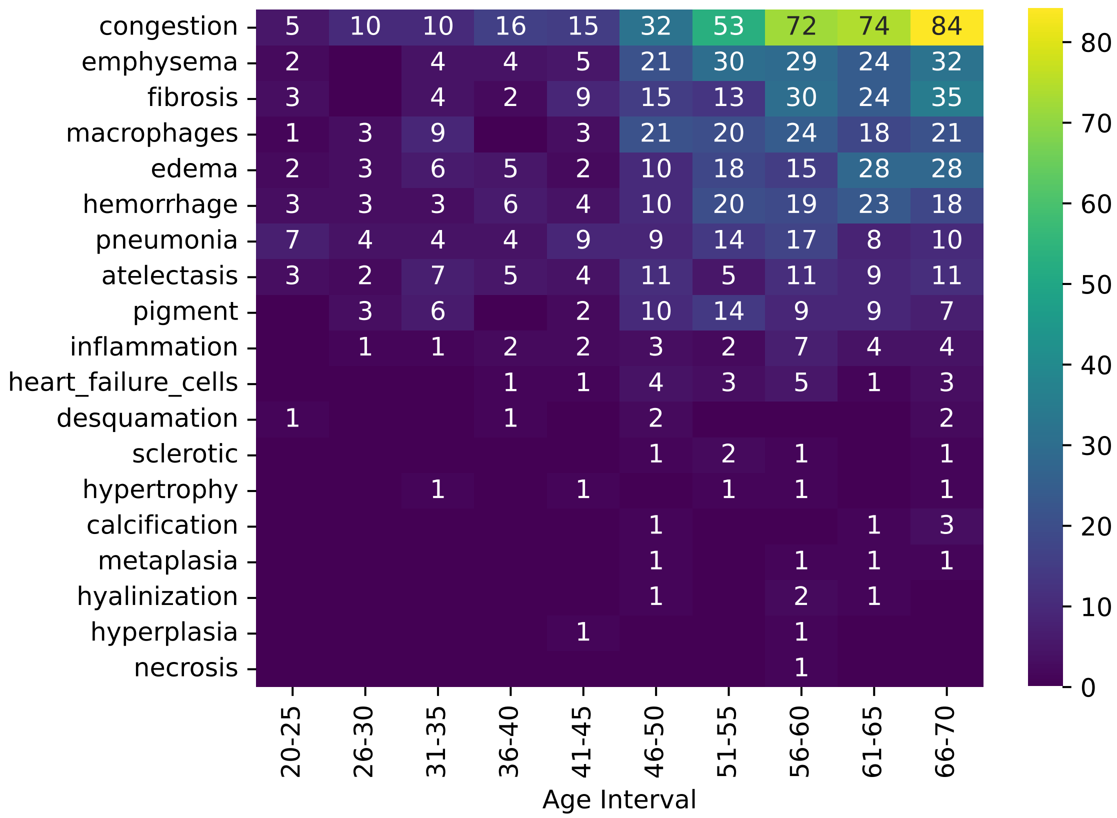


Figure 5.9: The heatmap displays the number of patients within each 5-year age range who have one of the 19 pathologies associated with lung tissue. There is a greater concentration of values in the upper right quadrant, where the older age groups are located. It is also evident here that many pathologies appear rarely in both younger and older subjects, as is the case for all those below "pigment". Note that empty cells denote no occurrence of the given pathology.

Regarding the pathologist's notes, we can also observe a significant imbalance in the use of terms. In Figure 5.10 (a), we can see that the frequency of tokens used generates a long-tailed distribution where the majority of terms are infrequently used while a few expressions are consistently repeated. Figure 5.10 (b) highlights the most frequent tokens that appear at the beginning of Figure 5.10 (a) with a higher value. We observe that phrases like '2 pieces', 'congestion', and 'moderate' dominate this distribution, followed by others such as 'macrophages', 'patchy', 'fibrosis', 'alveolar', among others. It is notable that many of the 19 pathologies occurring in lung tissue do not stand out among the pathologist's notes, either due to their low frequency of occurrence (Figure 5.8) or because they are eclipsed by other terms not equally relevant such as '2 pieces'.

a

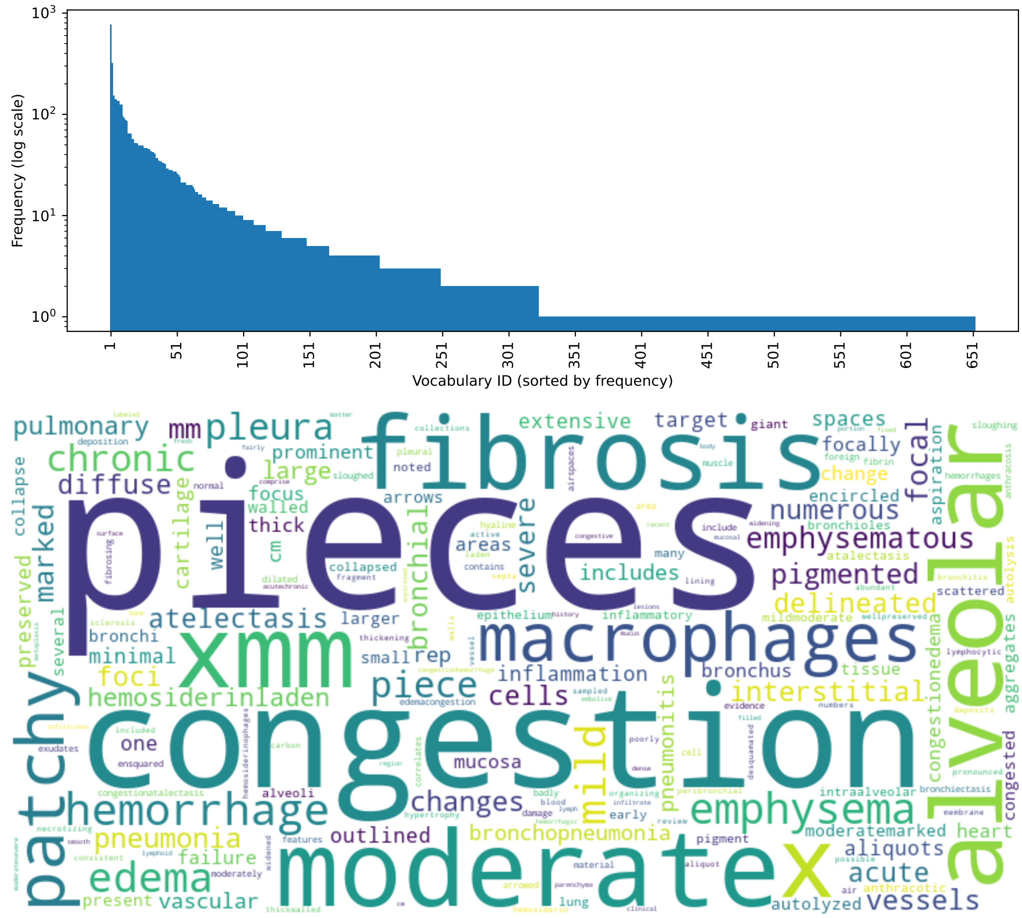


Figure 5.10: Frequencies of token (word) appearances in the pathologists' notes: (a) Distribution of tokens (words) from the pathologists' notes. Note that the y-axis is on a logarithmic scale. On the x-axis, the numbers represent the ID of each word after sorting from most to least frequent. (b) Word cloud showing the most recurrent words across the pathologists' notes.

Finally, there is also the imbalance associated with the smoking status of the patients. Figure 5.11 shows that the cases of non-smokers and ex-smokers are greater than those of smokers, and the distribution of these cases by age is also disproportionate.

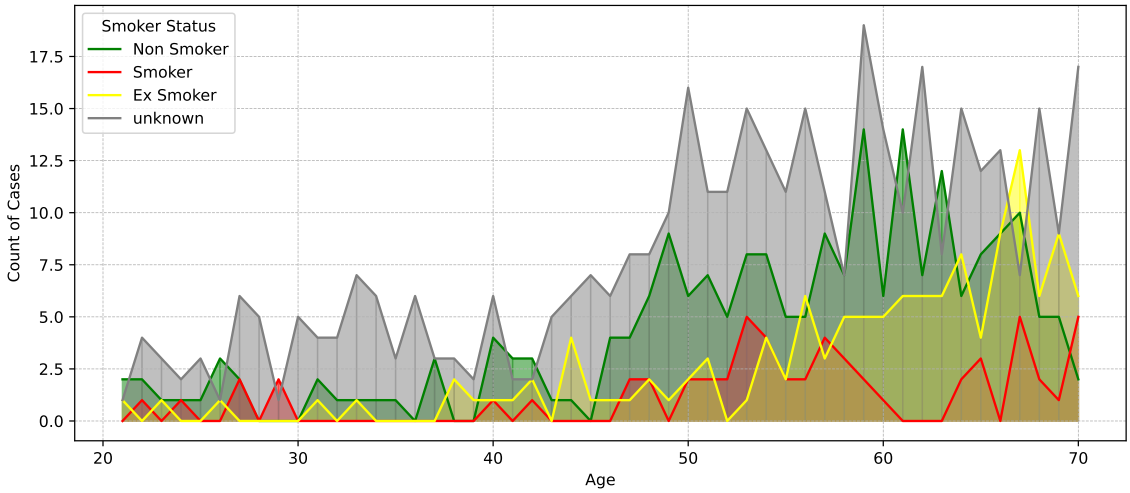


Figure 5.11: The figure displays the number of each smoker category across the ages of patients. We observe a higher concentration of unknown cases at all ages, and a greater concentration of all categories in older ages. It is also evident that smokers are generally the minority class.

All these imbalances have a direct impact on the modules of our system. The AP module is compromised by the age imbalance, the CG module by the imbalance of tokens in the pathologists' notes, and the SSP module by the imbalance among the three smoking classes. If not properly addressed, these imbalances can generate biases towards the direction of greater representativity of each of these variables, as we will see in the results chapter.

5.5.2 Co-occurrence of Pathologies

When developing a system aimed at identifying patient pathologies, it is important to consider the existence of correlations and cause-and-effect relationships between them. Although it is extremely difficult from a biological standpoint to infer such connections, their existence favors certain combinations of pathologies over others.

To understand the distribution of these combinations in our dataset, we can convert the combinations into classes and then count the number of patients present in each of these classes. The results of such processing can be seen in Figure 5.12.

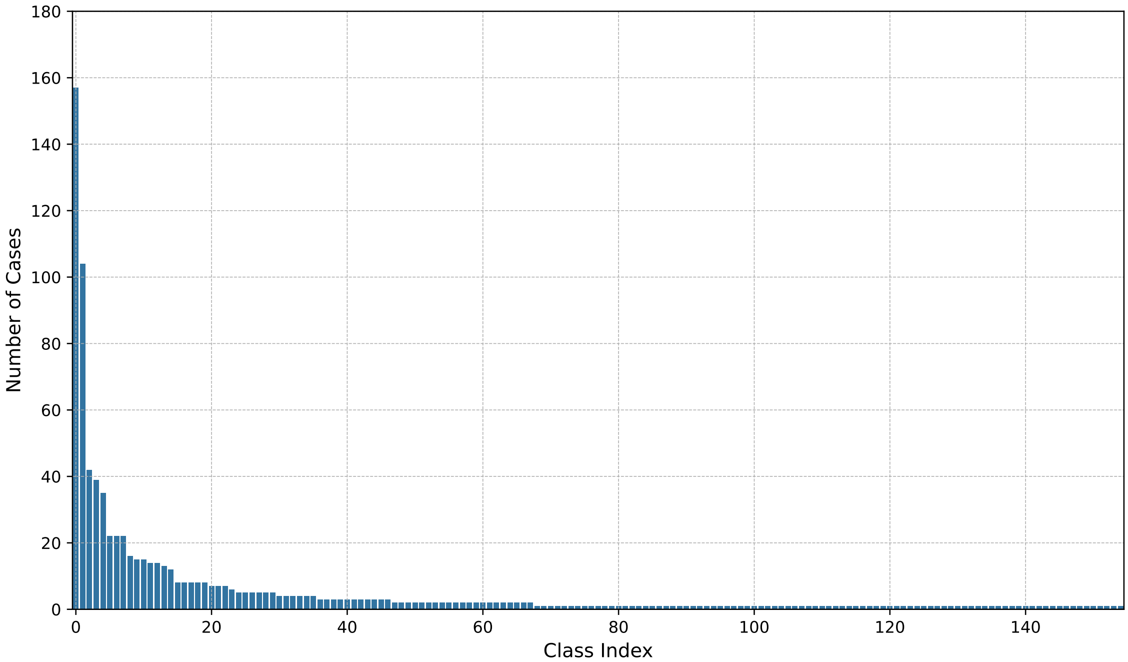


Figure 5.12: The figure displays the distribution of pathology combination classes. In total, there are 155 possible combinations represented here as class indexes in the x-axis, with the most frequent occurring in 157 patients and the least frequent occurring only once. We observe an extensive long tail with several rare combinations that appear in unique patients.

As observed in Figure 5.12, there is a large concentration of cases with similar pathology combinations and a long tail of rare combinations appearing in fewer than five patients.

One way to infer directionality between pathologies is to consider their counts in conditional probabilities. For example, given two pathologies A and B, the probability of finding A given the existence of B is $P(A \mid B) = \frac{\text{count}(A,B)}{\text{count}(B)}$. Similarly, the probability of finding pathology B given the pre-existence of A is $P(B \mid A) = \frac{\text{count}(A,B)}{\text{count}(A)}$. By defining a threshold, say 0.5, we can obtain a significant probability in both directions, in only one direction, or no significant connection between them. By conducting this pairwise procedure among all lung pathologies, we obtain a matrix called the co-occurrence matrix. From this matrix, by setting a significance threshold of at least 0.5, we obtain an adjacency matrix that translates into the directional graph shown in Figure 5.13.

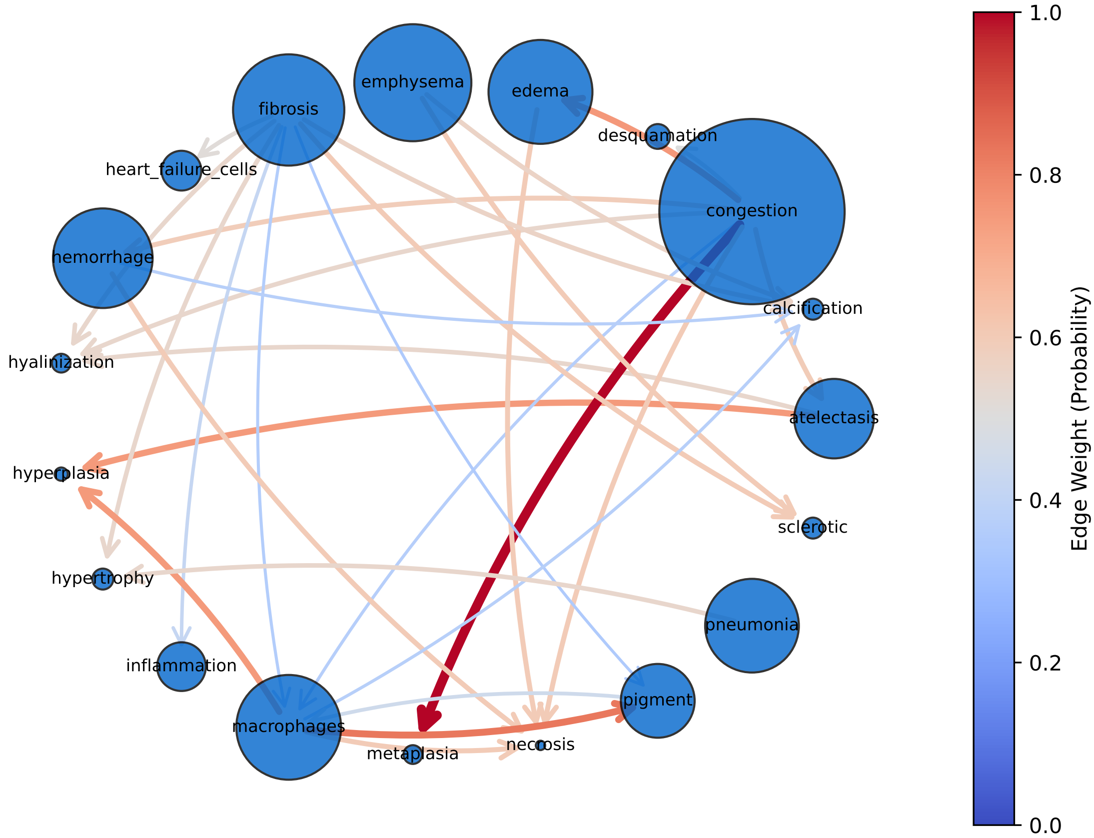


Figure 5.13: The graph represents the directional connections between different lung pathologies. The thickness of the edges represents the probability (the thicker the edge, the higher the probability, and thus the redder the hue), and the size of the nodes represents the frequency of occurrence of each pathology. From this graph, we see that there are preferred combinations, some unlikely, and others nonexistent. The directions were calculated using conditional probabilities.

Given the presence of preferred, more and less likely connections, any attempt to build a model for detecting such pathologies, whether a supervised classifier or a caption generation system, will face difficulties in predicting the correct combinations.

5.6 Attempt to Use a Caption Generation System

Our initial approach to the report generation system was to utilize a caption generation system as the foundation for producing reports. Such systems are particularly promising for generating reports from images, as they receive an image as input and return a description, which can then be further developed and enriched to produce a complete report. Initially, our idea was to generate a diagnosis in the form of a pathologist's note, detailing the pathologies present in the studied tissue and other structural details, as obtained through the 'Pathologist Notes' column (Table 5.1). For that purpose, we intended to use an Encoder-Decoder CG system as the examples outlined in section 3.1.

In our implementation, we primarily referenced the work of [Sengupta and Brown, 2024], which focuses on generating pathologist notes using the GTEx dataset. We adopted the structure and training methodology as detailed by the authors, with the sole difference being that while their CG system was trained using all tissues from the GTEx database, our experiment was limited to lung samples only.

Despite successfully generating captions, we encountered a significant issue: the captions were predominantly biased by the majority tokens found in the original diagnoses, almost invariably producing repetitive phrases. We can see this behaviour depicted in Figure 5.14.

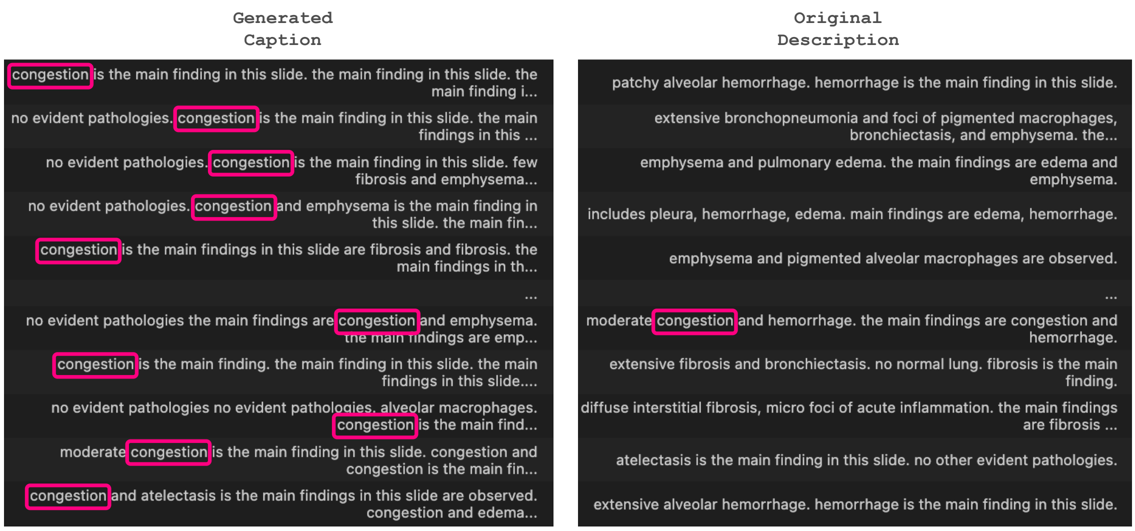


Figure 5.14: This figure highlights issues identified during tests with caption generation. On the left side, we present the predicted captions, while on the right side, we display the original captions, both of which have been processed. We note that the term 'Congestion' appears in all samples of the generated captions, often inappropriately, despite its presence in only one of the original captions. This recurrent inclusion of 'Congestion', a frequently occurring token in the lung dataset, was interpreted by us as an indicative of token imbalance affecting the caption generation process.

Confronted with this challenge, we decided to test the model weights trained by the authors, which were made available on their GitHub. This was to check if we could continue our project with only lung data using this pre-trained neural network. Although there was a slight improvement, and the network demonstrated a remarkable ability to identify the type of tissue in the images, it still failed to accurately detect most pathologies, encountering the same issue of generating repetitive phrases dominated by the most prevalent tokens in the training set.

As illustrated in Figure 5.15, all tissues in the GTEx database exhibit a certain level of token imbalance. For instance, the word 'pieces' consistently appears as the most frequently used term across all diagnoses. Furthermore, each tissue type shows certain pathologies and conditions occurring more frequently than others, reflecting an inter-imbalance among pathology classes within each tissue subset.

Figure 5.15: Word clouds generated from the pathologist's notes corresponding to the 12 tissues with the most samples in the GTEx dataset. In these word clouds, the size of each term is proportional to its frequency of occurrence across the notes, with larger words indicating terms that are mentioned more frequently. The number of samples from each tissue is indicated in parentheses.

The GTEx dataset comprises tissue samples from healthy subjects, and thus, substantial pathological elements in the annotations are not expected. However, as shown in Figure 5.7 for the lung tissue, the dataset predominantly consists of specimens from elderly individuals. This aspect is significant as the incidence of certain diseases escalates with age, potentially elucidating the imbalance in pathologies observed within the dataset.

However, while we cannot make any definitive conclusion, this experiment highlighted that the issue of imbalance in CG systems, though scarcely explored in the literature, is prevalent in various datasets used to train such models as underscored by [Ding et al., 2023].

Given our constraints of time and computational resources, we decided not to try other approaches to CG that are commonly used in histopathology, such as those detailed by [Lu et al., 2023], [Guevara et al., 2023], and [Zhou et al., 2024]. These either required architectures we could not train in our current setup or involved methods not directly applicable to Whole Slides, necessitating more complex adaptations.

Ultimately, we explored the use of CLIP model architectures, which employ contrastive learning to align text and image embeddings. This approach allowed us to work without necessitating tile-based inputs but, instead, utilizing embeddings that directly represents the WSIs by using the HIPT arquitechrua as our ViT [Chen et al., 2022]. This method not only yielded better results but also demanded fewer computational resources compared to the transformer-based caption generation systems. Detailed discussions on these outcomes will be provided later.

Although we did not proceed with the Encoder-Decoder type caption generation system as initially planned, the challenges we faced led us to discover various possibilities and guided us to the solution presented here. In future work, we may revisit these systems with a more mature perspective and perhaps focus on a more data-centric approach, aiming to mitigate the issue of imbalance before delving deeper into system architecture.

Chapter 6

Methodology

This chapter is designated to elucidate the methodology employed in developing the caption generation system. We begin with an overview of our system, from the input of images and their respective features to the output with the final report, aiming to highlight the contributions of each component.

Subsequently, each of these components is examined in a dedicated section, which describes the tasks performed and the methodologies applied, from architectural design through to training and performance evaluation. Finally, this chapter concludes with a subsection summarizing the challenges, limitations, and proposed enhancements for each module and the system as a whole, revisiting these points in the final conclusions of the thesis.

6.1 System Workflow

In this section, we provide an overview of our system's complete architecture, from the input of HIPT features to the generation of the final report. For didactic purposes, it is appropriate to divide this architecture into three main stages: inference (INF), aggregation (AGG), and generation (GEN).

In the first stage, INF (Figure 6.1 (a)), there are three main modules: Diagnosis Retriever (DR), Age Prediction (AP), and Smoker Classification (SC). These modules are tasked with generating crucial information for report composition. DR retrieves the pathologist's note from the most similar cases, AP estimates the chronological age of the patient, and SC predicts the smoking status of the patient, with categories including Non-Smoker, Smoker, and Ex-Smoker.

In the AGG stage (Figure 6.1 (b)), we synthesize all information generated in the inference module to produce a caption that encompasses both clinical and demographic characteristics of the patient and their lung tissue sample.

Finally, the GEN stage (Figure 6.1 (c)) relies entirely on LLMs to produce the final report. At the beginning of this stage, the generated caption from the previous step is provided to a Questioner Agent tasked with generating clinically relevant questions for understanding the patient's pathological condition. These questions are then individually addressed by another agent, the Q&E Agent, programmed to answer pathology-related questions based on a repository of references from literature that were scraped from various sources such as PubMed, BioArxiv, and selected histology books and manuals we manually collected. Ultimately, this content, along with the original caption, is delivered to a final agent, the Report Agent, charged with generating a histopathology report based on the provided caption and enriched with the content from the books.

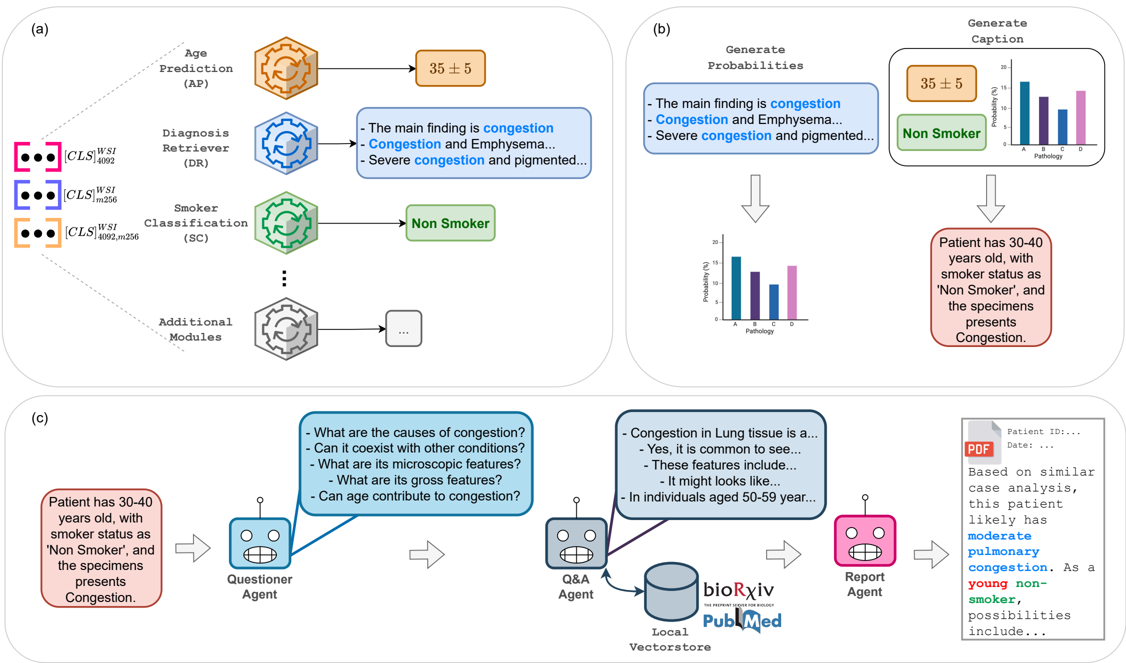


Figure 6.1: This figure scheme illustrates the three main stages of our solution: INF (a), AGG (b), and GEN (c). In the inference stage, we use our three main modules (AP, DR, and SC) to produce relevant information about the WSI based on the input HIPT features. Next, in AGG stage (b), all these information is gathered to compose the final caption (red box). Finally, in GEN stage (c), this caption is delivered to a pipeline with three agents responsible for information enrichment and final report production.

The architecture we developed was designed to make the INF stage as modular as possible, with an emphasis on flexibility and scalability. The modularity of its components (AP, DR, and SC) allows them to function as standalone and independent solutions. This independence permits significant flexibility within the system, enabling activation or deactivation of modules based on future interests and the possibility to enhance them individually as needed with advances in the field. Moreover, since the GEN stage receives only textual information, we can easily integrate new modules that produce other relevant pathological report details, such as tissue type identification, regions of interest on slides, and cell counting, and compose a new and more complete caption in the AGG stage. The GEN stage will then integrate it into the report as effectively as possible according to the custom prompt instructions.

In the following sections, we will delve into the details of each of these stages and their components.

6.2 Inference Stage

The inference stage is designed to predict all relevant information related to the subjects and their lung tissue samples. This stage includes two modules for inferring demographic features, Age and Smoker Status, and one module for clinical features, primarily through the pathologist's notes.

The selection of Age and Smoker Status as demographic features is informed by our previous research Moraes et al. [2023] and Ramirez et al. [2024]. These features are chosen for their known impact on lung health, which is crucial for accurate diagnoses of lung tissue conditions.

The use of the pathologist's notes for identifying pathologies is due to their established role as a clinical report. Incorporating these notes into our cross-modal retrieval system between text and image aids in enhancing the system's diagnostic capabilities by providing a reliable source of clinical information.

The following subsections will detail each module and the methodologies employed.

6.2.1 Age Prediction

The goal of the Age Prediction module is to predict the patient's age from their respective slides. For this purpose, we use the HIPT architecture as a feature extractor and employ these features in the downstream regression team targeting these ages.

In a previous study aiming to predict ages from WSIs, we hypothesized that the lack of robustness in our histology models was mainly due to two factors: (i) severe imbalance in the distribution of ages (ii) loss of information in the process of segmenting the image into tiles. As seen in Figure 5.7, the age distribution in this dataset has a severe imbalance favoring older ages, above 45 years. To address obstacle (i), in this work, we opted to integrate a technique called label distribution smoothing (LDS) [Yang et al., 2021], and to handle obstacle (ii), we decided to work with HIPT features in the hope of extracting more significant and contextualized representations, without a drastic loss of global information.

Following this, we detail the architecture of the model, data preparation, and the processes of training and performance evaluation.

6.2.1.1 System Architecture

The architecture of the age prediction model consists of four main modules: an input layer, a backbone, a neck, and finally, a head output layer. As the backbone, we used the previously mentioned HIPT, which is fully frozen so that its weights do not update. From this architecture, we extracted four different types of features, of which we used three: $[CLS]^{WSI}_{4096}$, $[CLS]^{WSI}_{m256}$, and $[CLS]^{WSI}_{4096,m256}$. Serving as the neck, we coupled a Multi Layer Perceptron (MLP) with variable numbers of layers, nodes per layer, and dropout rates that were dynamically selected during the hyperparameter optimization process. Finally, as the head, we have a linear layer with a single node. In Figure 6.2, we can contemplate a schematic of this architecture with the mentioned modules.

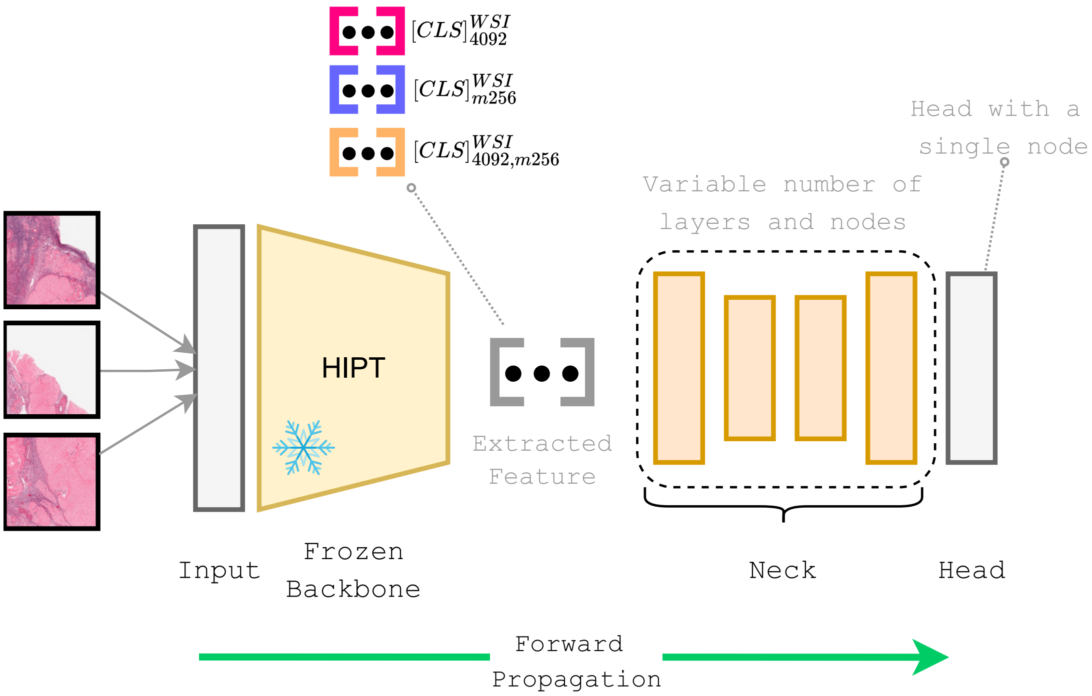


Figure 6.2: Architecture used in the age prediction module. Here we can see four main components: an input layer, a pre-trained backbone, a composition of layers forming the neck, and finally, a head layer with a single node. We also illustrate here the three different types of features extracted from the HIPT that we used in building the model.

6.2.1.2 Data Preparation

Beyond the standard preparation of histology images and subsequent feature extraction as outlined in Figure 5.3, we also removed patient instances associated with null age values and transformed the label space according to LDS theory to address data imbalance issues.

6.2.1.3 Training and Selection

To select the best model, we first conducted parameter and hyperparameter optimization using the Optuna framework, aiming to minimize the Mean Square Error (MSE). During this process, we employed a k-fold cross-validation strategy with $k = 5$. Throughout the trials, we endeavored to identify the optimal set of values for the variables listed in Table 6.1.

Variable Name	Available Options
reweight_opt	sqrt_inv, inverse
lds	true or false
lds_kernel	gaussian, laplacian, triangular
lds_ks	min: 3, max: 8
lds_sigma	min: 2, max: 5
custom_loss	mse, weighted_mse, weighted_focal_mse
hipt_features	CLS_{4096}, CLS_{m256}, $CLS_{4096,m256}$
learning_rate	min:10^{-4}, max:10^{-5}
n_layers	min: 1, max: 6
n_nodes	min:2^5, max:2^{10}
dropout_rate	min: 0.15, max: 0.35
n_epochs	min: 12, max: 13
batch	min:2^5, max:2^8

Table 6.1: Summary of the sample space for parameter and hyperparameter tuning in the age prediction module.

6.2.1.4 Evaluation

After several runs over the sample space of Table 6.1, we selected the parameters from the best trial. Then, based on these found parameters, we reduced the search space, considering the new values as central, and conducted another set of trials. In total, we ran 155 trials in this optimization stage.

Ultimately, we chose the parameters from the best trial of this second optimization round to be used in training the final model. From this point, we were able to assess the performance of this model on the common test set. The results are detailed further in the results chapter.

6.2.2 Smoker Classification

The smoker classification module was developed to identify the smoking status of individuals in our sample. We have three categories: ’Smoker’, ’Non Smoker’, and ’Ex Smoker’. The motivation for incorporating this module into the report generation solution is directly related to the type of tissue being analyzed. Since we are exclusively working with lung samples and are aware of the direct impacts of smoking on lung health from various studies, we believe it could be relevant to include this module as part of the solution, either influencing decisions in other modules or featuring directly in the final report.

In previous work, we used a pre-trained Xception model on the ImageNet dataset as a backbone for extracting features from histology images and subsequently classifying these images into 'Smoker' and 'Non Smoker' categories. The images used in that study were 512x512 tiles extracted from WSIs, and the final classification was determined by Majority Voting among these tiles for each subject. As previously mentioned in the discussion about the AP module and our prior work on that topic, working with isolated tiles inevitably leads to information loss since only local features are considered and global interactions within the complex hierarchical structure of a WSI are ignored. Therefore, in this study, we use features extracted from HIPT, which aims to address this issue by generating features at different hierarchical levels and retrieving embeddings that are both local and global. Moreover, we attempt here to extend the label space to also include the 'Ex Smokers' class, which was not directly addressed by the classifier in our previous work.

6.2.2.1 System Architecture

The architecture we use here is essentially the same as that used in the AP module Figure 6.2. We use the same HIPT features as input, and the network is also dynamic, meaning we are testing different combinations of the number of layers, nodes in each layer, and dropout rates. The difference here lies in the output layer, which has three nodes, one for each class, and in the loss function used, which instead of being a Binary Cross Entropy, is a weighted cross entropy loss.

6.2.2.2 Data Preparation

Just as with the specific chronological age used in the AP module, the smoking status of patients is also confidential and can only be obtained through a direct request to the GTEx Portal team. Once these data are acquired, we perform a minor adjustment to the names of the categories since they are spread across different columns, and finally, we perform a left join with the main table illustrated in Table 5.1 using the subject ID as the key.

In addition to the categories already mentioned, there is another category labeled 'unknown' used for subjects whose smoking status is not available. Since this category could potentially represent any of the other three categories and thus pose a risk for classification, we opt to remove it from the dataset, retaining only 'Smoker', 'Non Smoker', and 'Ex Smoker'.

6.2.2.3 Training and Selection

After excluding the 'Unknown' subjects, all other instances that are not part of the common test set were used as the training set. With this training data, we conducted a parameter and hyperparameter optimization analogous to that performed in the AP module, using a strategy of 5-fold cross-validation. The search space used in this module can be seen in Table 6.2.

Variable Name	Available Options
reweight_opt	sqrt_inv, inverse
alpha	min: 1, max: 4
hipt_features	CLS_{4096}, CLS_{m256}, $CLS_{4096,m256}$
learning_rate	min:10^{-4}, max:10^{-5}
n_layers	min: 1, max: 5
n_nodes	min:2^5, max:2^{10}
dropout_rate	min: 0.15, max: 0.35
n_epochs	min: 12, max: 13
batch	min:2^5, max:2^8

Table 6.2: Summary of the sample space for parameter and hyperparameter tuning in the smoker classification module.

In Table 6.2, the parameter alpha was utilized by us to assign additional weight to the minority class. If alpha is 1, then we maintain the weights as originally generated by strategies such as inverse frequency or inverse square of the frequencies of the classes. However, if alpha > 1, we multiply the weight of the minority class by alpha in order to increase it. This adjustment is designed to counteract the imbalance by enhancing the influence of the minority class in the model training process, potentially improving the model's ability to correctly classify instances of this less represented class.

6.2.2.4 Evaluation

After several runs, we selected the parameters from the best trial and conducted a second round of testing around these values similar to the process we followed in the Age Prediction module, ending up with a total of 69 trials. Ultimately, we chose the parameters from the best trial of this second optimization round to be used in training the final model. From this point, we were able to assess the performance of this model on the common test set. The results are detailed further in the results chapter.

6.2.3 Diagnosis Retriever

The diagnosis retriever (DR) module was designed to find the pathologist notes most adequate to a given image. This is accomplished using a model based on CLIP, which employs contrastive loss to align image and text pairs based on their pairwise cosine similarity matrix. Once a pair is well-aligned, it is possible to generate embeddings for a new image and conduct a search over the text embeddings (pathologist notes) present in the training corpus. The top K retrieved texts are then delivered to the Report Generation (RG) module to produce a consensus-based report.

6.2.3.1 System Architecture

The architecture of this module is essentially the same as that used by Radford et al. [2021] in the development of the CLIP model. Given a dataset $\{x_i^I, x_i^C\}_{i=1}^D$ of size D, consisting of images x_j^I and their respective captions x_j^C, the authors utilize two transformer architectures, a Vision Transformer (ViT) with an embedding layer f_I, and another language model transformer-based with an embedding layer f_C, to obtain the embeddings $z_i^I = f_I(x_i^I)$ for the images and $z_i^C = f_C(x_i^C)$ for their corresponding texts. After obtaining the embeddings, they perform the projection of both to the space $\mathbb{R}^{B\times 512}$, where B is the batch size, followed by normalization of these projected embeddings. Subsequently, the pairwise cosine distances between each of the embeddings z_i^I and z_i^C are calculated to produce a square matrix $S \in \mathbb{R}^{B\times B}$ with elements given by $s_{i,j} = z_i^I \cdot (z_j^C)^T$. Finally, these cosine distances are used as logits and fed into a contrastive loss function that optimizes the principal diagonal elements $s_{i,i}$, wich corresponds the actual text to its respective image.

To integrate this architecture with HIPT, we customized the CLIP model provided by the Transformers library from HuggingFace Wolf et al. [2020]. Our customization differs from the original version in the following aspects in two main aspects. First, we are using

HIPT as the ViT to convert Whole Slide Images (WSIs) into embeddings, and second, we are keeping out ViT with its weights frozen throughout the training process, only upding the weights of the language model as the projection layers for image and text.

In Figure 6.3, we can see in some details of the components of this structure. Fist, in (a), nós estamos a mostrar o processo

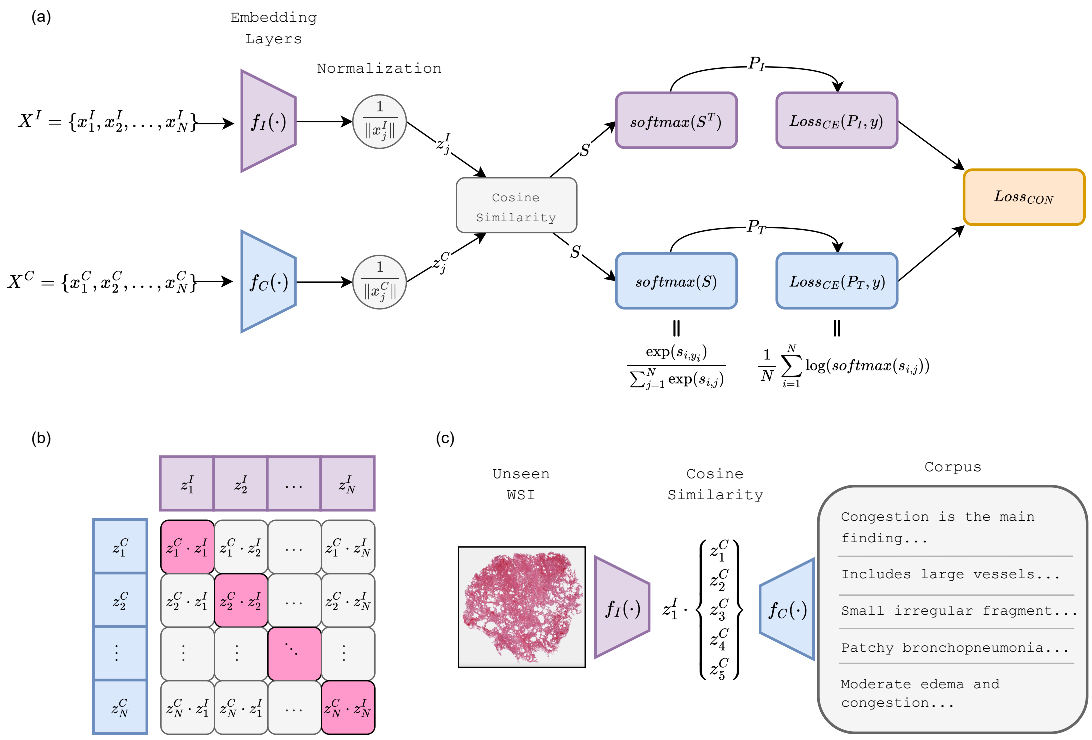


Figure 6.3: Illustration of the information flow occurring within CLIP-like models (a). We can see the images (I) and texts (T) passing through the embedding layers until reaching the final node where the contrastive loss is calculated. In (b) there is an illustration of the cosine similarity matrix between texts and images and, finally, in (c) we present a schema of the process of text retrieval given the embeddings of a new WSI.

6.2.3.2 Data Preparation

Data preparation for this module begins with feature extraction as illustrated in Figure 5.3. Subsequently, referencing the 'Pathology Category' column (Table 5.1), we perform a multilabel stratification of patients into training, testing, and validation sets.

For the diagnoses and pathologist's notes contained in the 'Pathology Notes' column, we employed a cleaning process using GPT-4o. A specialized prompt was designed to remove special characters and delete recurrent words of little relevance, such as 'pieces,' which frequently appears in the dataset. Additionally, we converted the 'Pathology Category' column into text and concatenated it with the original pathologist's note to enhance the diagnostic context of the pathologies described. Figure 6.4 displays an example of an original text from 'Pathology Note' alongside its processed version. Further details regarding the prompt used, as well as specifics about the model, temperature settings, and token count, are available in the appendix section B.1.

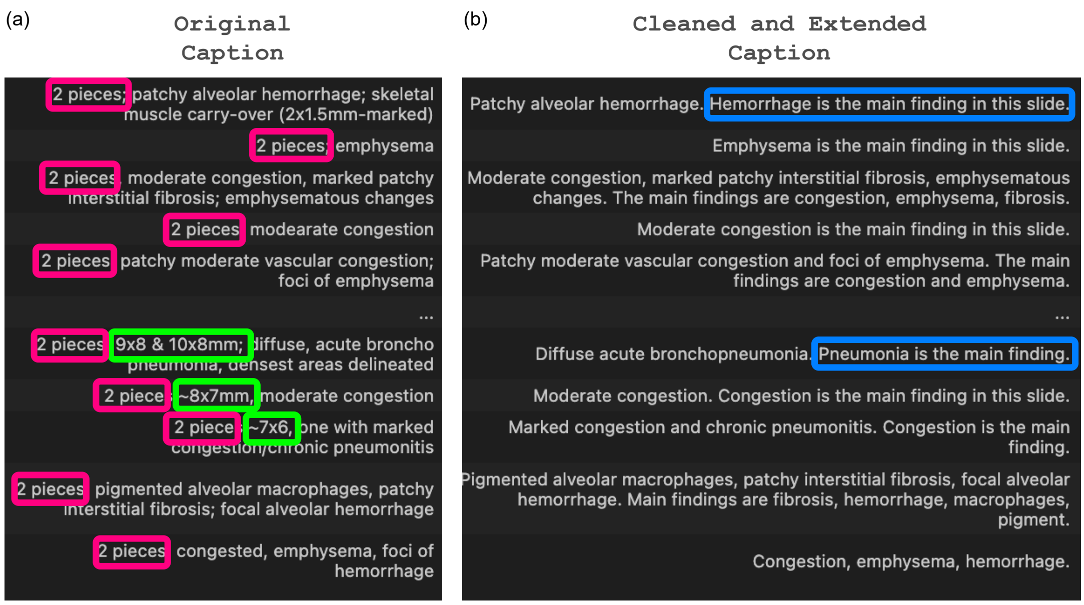


Figure 6.4: This figure illustrates the result of cleaning and extending descriptions using GPT-4o. Panel (a) shows the original descriptions with issues such as repetitive tokens highlighted in pink, and numbers, special characters, and units of measure marked in green. Panel (b) displays the cleaned and extended descriptions, with the pathologies present in the image highlighted in blue.

It is important to note that since we were constantly working with limited computational resources, we opted not to use a cross-validation strategy in this module, which utilizes two computationally expensive transformer architectures. Due to this, we used a predefined validation set, unlike in the AP and SC modules.

6.2.3.3 Training and Selection

We first conducted parameter optimization for batch size, learning rate, HIPT feature, and number of epochs using a Random Search strategy with the Optuna framework. The search space utilized can be seen in Table 6.3.

Variable Name	Available Options
hipt_features	CLS_{4096}, CLS_{m256}, $CLS_{4096,m256}$
learning_rate	min:10^{-3}, max:10^{-6}
n_epochs	min: 32, max: 32
batch	min:2^4, max:2^8

Table 6.3: Summary of the sample space for the diagnosis retriever module.

For each trial, we collected different categories of metrics. Since we are comparing descriptions that can be understood as image captions, we selected a set of metrics appropriate for caption comparison (Bleu, Cider, Meteor, Spice, Spider, Rouge). Given that we are also generating a probability vector where each entry in the vector represents the likelihood of a specific pathology occurring, and since we have the true labels for the correct pathologies associated with each image, we were also able to select a set of similarity metrics (Cosine, Euclidean, Manhattan, Hamming). Finally, as we aim to be accurate concerning the pathologies, we can also consider this problem as a multilabel classification task and use the aforementioned vectors to calculate metrics such as recall, precision, and F1-score, which provide a better view of our accuracy for each of the pathologies viewed as categories. A summary of these metrics can be found in Table 6.4.

Metric class	Metric type
Caption (CPT)	Bleu (1-gram a 4-grams), Cider, Meteor, Spice, Spider, Rouge.
Similarity (SIM)	Cosine, Euclidean, Manhattan, Hamming
Classification (CLS)	Recall, Precision, F1-score

Table 6.4: Summary of the metrics used in the training process of the diagnosis retriever module.

In Figure 6.5, a schema of the retrieval process and metric collection is presented. Figure 6.5 (a) illustrates how a new image x^I_{new}, not previously seen by the trained model, is processed. This image is input into the image embedding layer, yielding the embedding vector $z^I_{new} = f_I(x^I_{new})$. Simultaneously, every sentence from the corpus of pathologist's notes (diagnoses or captions) from the training set is processed through the text embedding layer, $z^C_j = f_C(x^C_j)$, resulting in a set of embedding vectors $\{I^I_j\}^D_{j=1}$, where D is the dimension of the training set. In the figure example, $D = 5$. We then compute the cosine similarities $s_{i,j}$ between z^I_{new} and each z^C_j, and select the top K most similar embeddings. Using their indices j, the corresponding texts are retrieved. In the example depicted, K=2, and thus 2 diagnoses are retrieved.

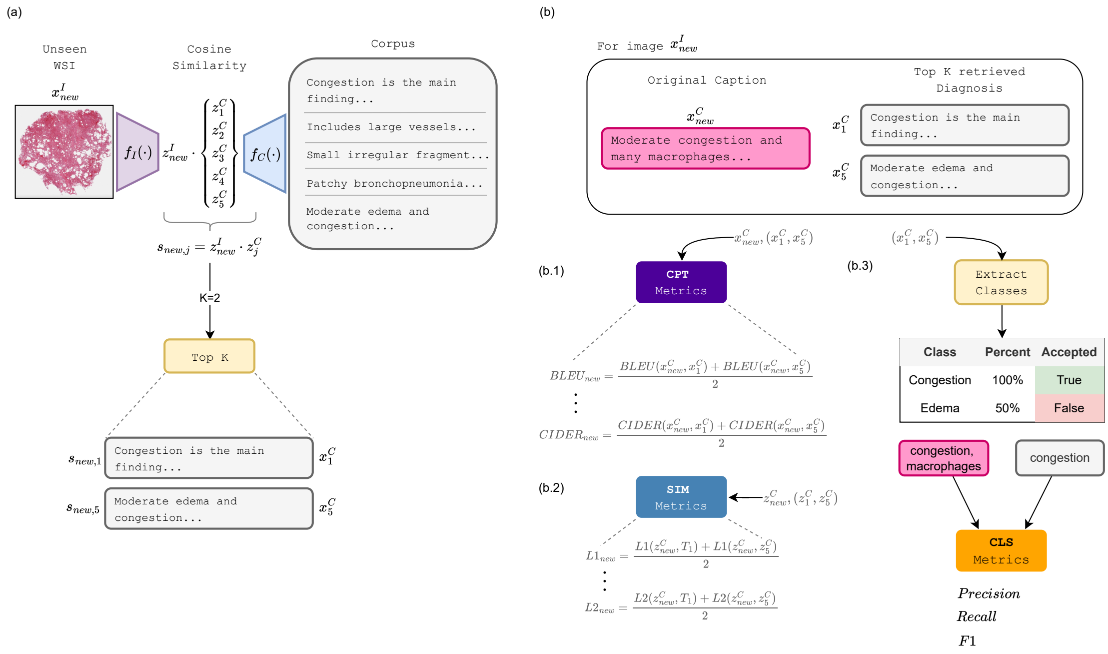


Figure 6.5: Illustrative schema of the cross-modal retrieval process (a) and subsequent calculation of metrics (b) for captioning (CPT) (b.1), similarity (SIM) (b.2), and classification (CLS) (b.3).

Next, the stage of metric collection is described. Given that x^I_{new} is associated with a true diagnosis x^C_{new}, comparisons are made between x^C_{new} and each retrieved text, as shown in Figure 6.5 (b), where the texts are x^C_1 and x^C_5. CPT metrics are then calculated (Figure 6.5 (b.1)) along with SIM metrics (Figure 6.5 (b.2)). For calculating the CLS metrics, we first assess the frequency of each pathology occurrence among the retrieved texts. Using a defined threshold, the predicted pathologies are identified. As depicted in Figure 6.5 (b.3), since congestion appears in both but edema only in one, and considering the greater similarity of x^C_1 to x^C_{new} compared to x^C_5, congestion is identified as the predicted pathology for this instance. With the real pathologies (congestion and macrophages) and the predicted one identified, the CLS metrics can be calculated as illustrated.

6.2.3.4 Evaluation

After 100 iterations, we analyzed the metrics, focusing on the F1-score, and the retrieved texts for each image. Based on this analysis, we defined an optimal set of parameters to run a final model using our entire training and validation dataset. From this final model, we were able to assess performance on the common test set across the modalities with which we are working.

6.2.4 Threshold Calibration

In the realm of classification problems, a conventional threshold of 0.5 is commonly employed to distinguish between positive and negative classes, based on the probability predicted by the classifier model for each instance. While this approach is generally effective, certain contexts necessitate the calibration of this threshold to a more suitable value, guided by specific criteria. In the clinical context, for instance, the primary objective is often to minimize the risk of misclassifying a patient as healthy when they are actually ill and require treatment. Consequently, the calibration strategy in such cases involves minimizing the false negative rate for each of the classes under consideration. This adjustment ensures that the model is more sensitive to detecting true positives, thereby reducing the likelihood of overlooking patients who need medical intervention.

Therefore, for the DR module, we calibrate the thresholds for each pathology, aiming to find the best probability cut-off for each. To acchieve that, for each predicted pathology, we first calculate the True Positive Rate (TPR) and False Positive Rate (FPR) across an array of different thresholds $\vec{t} = [0.1, 0.2, ..., 1.0]$, following the receiver operating characteristic (ROC) curve algorithm. This results in two vectors, $\vec{v}_{TPR}$ and $\vec{v}_{FPR}$, which contain the TPRs and FPRs for each threshold tested. Subsequently, we utilized $\vec{v}_{TPR}$ and $(1 - \vec{v}_{FPR})$ to obtain a vector of scores $\vec{s}^i$, where i indicates the index of the pathology in question, such that

$$\vec{s}^i = 2\vec{v}^i_{TPR} - \vec{v}^i_{FPR} \tag{6.1}$$

Finally, for each pathology i, we identify the index of $\vec{s}^i$ that holds the maximum value in the array and select the corresponding index in $\vec{t}$ to set as the threshold for pathology i. Equation 6.1 is designed to minimize the false-negative rate because in clinical applications, failing to identify a sick patient can prevent them from receiving necessary treatment, potentially leading to severe consequences.

In addition to implementing this process within the DR module, we extended its application to the SC module. The resulted calibrated confusion matrices for DR and SC are presented in chapter 7.

6.3 Aggregation Stage

As shown in Figure 6.1 (b), the AGG stage fulfills two primary functions. The first is to determine the probability of each pathology category to occur based on the K diagnoses retrieved from the DR module. The second function is to generate a caption that consolidates the results of all prior modules into a concise description containing both pathological and demographic data.

6.3.1 Obtaining Pathology Classes

The probabilities of pathology classes are derived from the texts retrieved in the DR stage. For the series of K diagnoses $\{x_j^C\}_{j=1}^K$, we initially count the pathologies present in each diagnosis. Each x_j^C is then assigned a weight within the range $[0, 1]$. Considering the order in which these texts are retrieved, which reflects their similarity to the reference image, x_{new}^I, a linear set of weights is created from 1 to 0, where x_1^C receives weight 1 and x_K^C receives weight 0. These weights are multiplied by the pathology counts in each description, and a final aggregation is performed by summing these weighted probabilities for each pathology. The results are then normalized to ensure that probabilities remain between 0 and 1.

Figure 6.5 (b.3) illustrates the process for calculating the count and frequency of pathologies.

6.3.2 Generating the Final Caption

Upon receiving outputs from the AP, SC, and DR modules, the goal of this phase is to integrate this information into a single caption, which will subsequently be used in the GEN stage as the foundation for constructing the final report.

Methods to achieve this objective includes (i) employing an LLM to deduce the optimal caption from the set of retrieved descriptions and the probabilities for each pathology, and (ii) manually concatenating outputs of the previous modules using a predefined template. Method (i) has the advantage of allowing the LLM to identify not only the most frequent pathologies but also intensity adjectives like 'moderate', 'severe', 'high', and 'low', which help scale the severity of the conditions. However, the downside is potential noise generation if the retrieved cases vary significantly or fail to accurately reflect the severity of the conditions, which could affect the accuracy of the final report. Method (ii) is simpler and minimizes the introduction of additional errors beyond those from the individual inference modules, but it lacks the detailed quantitative adjectives that can significantly aid in assessing the severity of the analyzed sample.

In our project, we chose the more conservative approach of method (ii) to limit error propagation within our system's workflow. The template used for manually composing captions from the outputs of the previous modules is as follows:

```
"Patient has "{age}" years old, with smoker status as "{smoker}", and its
  specimens, from Lung tissue, present "{pathologies}""
```

Listing 6.1: Template used for manual composition of captions using the outputs of the previous modules

6.4 Generation Stage

The GEN stage is tasked with two primary functions: enriching information and composing the final report. The workflow in this stage exclusively utilizes LLMs based on transformer architectures, which are adept at handling sequence-to-sequence tasks and operating in decoder-only modes.

Training LLMs on standard home laptops presents considerable challenges due to the immense computational resources these models require. For instance, the pre-training of LLaMa-2-70B, which has 70 billion parameters, alone consumes about 1.7 million GPU hours and 2.5 trillion Joules of energy. Typical home laptops, equipped with far less powerful CPUs and GPUs, are incapable of handling such extensive computational demands without significant compromises in speed, efficiency, and model accuracy. Moreover, the inference process, although generally less resource-intensive than training, still poses significant challenges in resource-limited environments, sometimes consuming comparable or even greater amounts of resources than training phases, depending on the model and tasks.

Given these constraints, a practical solution is to utilize cloud computing services, which offer access to powerful GPUs and enhanced computational capabilities, allowing users to pay for usage time. This approach is particularly valuable for home users and researchers without access to supercomputing facilities who need to train or fine-tune these models with domain-specific data.

Another alternative is to employ on-demand APIs from providers like OpenAI, Mistral, and others, which offer a wide array of models for inference tasks. These services charge a fee based on the number of tokens processed.

In our scenario, working from home with limited resources, we chose to use APIs instead of cloud-based GPUs primarily to reduce costs and effort. These APIs grant access to high-performance models, such as GPT-4, at a cost-effective rate for small-scale researchers who are financially constrained.

However, using APIs does not fully resolve the issue of specializing these models to our domain-specific data, as would be possible with direct fine-tuning. To address this, we employed the RAG technique alongside with PE methods. This approach involves maintaining our domain-specific data in a vector database accessible to the LLM. With a specialized prompt, we direct the model to prioritize our database data over external data.

In the subsequent sections, we will explore in detail each step of this generation process, employing techniques such as PE and RAG, in conjunction with the OpenAI APIs, to generate the histopathological report.

6.4.1 Information Enrichment

A histopathology report is more than a simple pathologist's note or a caption describing conditions present in specimens. While a caption may encapsulate the essential information necessary for generating a diagnosis from a WSI, it often lacks detailed insights such as the location of detected conditions, the intensity of observed pathologies, and nuanced gross and microscopic analyses. Given the clinical utility and educational implications, a more comprehensive report that includes relevant information about the observed conditions and extends to incorporate findings from scientific literature could significantly benefit both students and trained professionals.

In order to enhance the diagnostic captions generated in the AGG stage with useful and pertinent information, we have developed this stage of information enrichment. Our approach involves expanding the content of the captions by posing and addressing a series of pertinent questions about the target problems presented in the input diagnosis. These questions are designed to gather additional, enriching material, to further elucidate the conditions depicted in the captions.

However, the domain-specific nature of questions related to a histopathological diagnosis requires precision and control, which may not be fully achievable through a standard

LLM trained on vast data sets. Therefore, we opted for the control and precision that an LLM equipped with the RAG technique would give us, as we know exactly which references are consulted and can control the behavior of the LLM on how to consult these materials and how to process the information contained therein, thus enhancing the precision and relevancy of the responses.

To implement the information enrichment proposal described above, we instantiated two agents: a Questioner Agent specialized in asking questions relevant to the diagnosis, and a Q&A Agent specialized in answering questions about histopathology, specifically lung pathologies. An overview of this workflow scheme is illustrated in Figure 6.1 (c), where the nature of the questions and the general format of the answers can be observed.

The Questioner Agent, a GPT-4 model, takes the caption generated in the AGG stage as input and outputs a dictionary containing up to 10 questions related to the diagnosis from the caption. The Q&A Agent, instantiated with a GPT-3.5, processes these questions iteratively, providing answers based on the scientific literature it accesses using the RAG strategy. This agent is also equipped with a specialized prompt that directs the response process in a desired manner. Another crucial component of the Q&A Agent is its memory capability, which enhances the precision of its searches and responses. This improvement was confirmed during the testing and implementation phases of these models.

The documents utilized to build our vector database for the Q&A agent were gathered from PubMed and BioRxiv using a Python library named Paperscraper. Furthermore, we manually curated a selection of technical books focused on pulmonary histology to enrich the final vectorstore database. For more detailed information about the Paperscraper library, please refer to their GitHub page: `https://github.com/jannisborn/paperscraper`.

Finally, further details about the prompts used, along with other parameters such as the model specifications of the LLM, temperature settings, and maximum tokens, are provided in Appendix B at B.2 and B.3.

6.4.2 Report Generation

In this stage, we instantiate another instance of a LLM, specifically a GPT-4, tasked with generating the final report using the data produced during the enrichment phase. The agent receives both the caption from the AGG stage and the answers provided by the

Q&A Agent. Based on the guidance of a specialized prompt (refer to appendix B.5), it generates an enriched pulmonary histopathology report.

The report is structured into four main sections: Gross Description, Microscopic Description, Comments, and Summary. In the Gross Description, the agent is directed to detail the general and visually observable traits mentioned in the caption. If the caption lacks details on gross traits, the agent is instructed to infer these traits based on the pathology or combination of conditions depicted in the caption and to incorporate pertinent information from the responses of the Q&A agent. Similarly, the Microscopic Description section is handled by the agent using the responses from the Q&A agent to articulate content reflecting the pathologies observed. This section aims to specify the most probable microscopic characteristics given the pathological context provided in the caption.

In the Comments section, the prompt allows the LLM more freedom to include any relevant observations it deems necessary based on the caption's context and the information derived from the responses. This may include hypothesizing about the patient's age and smoking status. Finally, the Summary consolidates the most significant findings of the analysis before finalizing the report.

6.4.3 Evaluation

The performance evaluation of this module presents significant challenges. As this project was not developed in collaboration with its intended end users, namely pathologists and pathology students, it lacks a direct means for an appropriate evaluation of the reports. Ideally, each report would be scrutinized by a pathologist who would compare it against the corresponding Whole Slide Images to assess its validity and verify the accuracy of the information added during the enrichment stage with their expert knowledge. In the absence of such specialized human validation, we resorted to using another LLM as an evaluator agent. This agent is tasked with comparing the final report produced to the original pathologist's note and assigning a score on a custom scale.

Our evaluation methodology incorporates two distinct approaches. The first evaluates the classificatory power of the report, specifically, how accurately it identifies pathologies mentioned in the original diagnosis. This aspect largely depends on the performance of the Diagnosis Retriever (DR) module, which identifies the most similar cases and infers associated pathologies. To complement this, we introduced a second evaluation focusing on the informative quality of the report regarding the pathological and demographic conditions it discusses. Here, the goal is not merely to judge the accuracy of the predicted pathologies or demographic data, but to assess the quality of information provided in the enrichment stage, based on the assumption that the input caption data are accurate.

We adopted a categorical 4-level scale for these evaluations, ranging from 'Completely unrelated' to 'Accurately reflects' the information in the reference caption. After evaluating all reports, we count the scores in each category to quantify the overall performance of this stage.

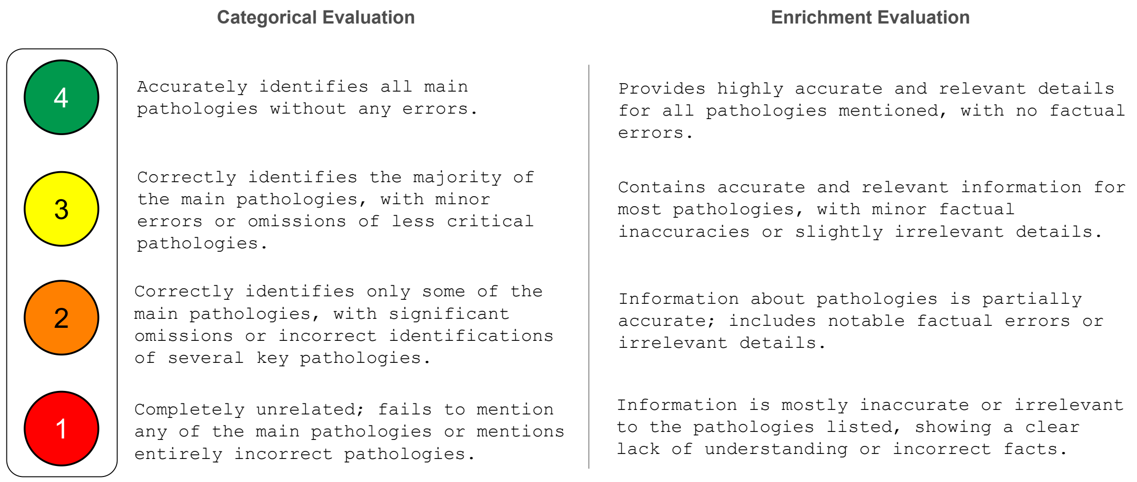


Figure 6.6: This figure illustrates the evaluation scale used to assess the system's performance in the generation stage based on the produced reports. Two evaluation systems utilizing the same scale are depicted: one (Categorical Evaluation) assesses the report's classificatory power, focusing solely on the pathologies compared to the original diagnosis. The other (Enrichment Evaluation) examines the informative quality of the report concerning the pathologies presented in the caption generated in the aggregation stage. Even if there is a low performance in correctly classifying pathologies in the categorical evaluation, a successful enrichment stage that accurately provides information about the input pathologies would result in a more favorable score in the second approach.

Chapter 7

Results

In this chapter, we present the results related to each module of our report generation system. Each section follows a similar structure, beginning with the results of the parameter optimization trials, followed by the performance outcomes of the best model for each module. This provides an opportunity not only to showcase results and model performance but also to generate insights and hypotheses from the gathered data.

7.1 Age Prediction

As outlined in subsection 6.2.1, we conducted 155 trials by adjusting parameters from the search space shown in Table 6.1, employing a k-fold cross-validation strategy. These trials informed the selection of an optimal set of parameters for training the final model. This section discusses the results from these trials and the final model configuration derived from them.

7.1.0.1 Parameter Optimization and Selection

Throughout the 155 trials in the parameter optimization stage, we evaluated the metrics MSE, RMSE, MAE, MAPE, and R2 for each of the 5 folds of each trial. The distributions of these metrics across the folds are depicted in Figure 7.1. It was observed that certain folds, such as fold 3, consistently outperformed others, showing better average values despite the presence of several outliers. Conversely, folds 5 and particularly 4, generally exhibited poorer performance. This variability among the folds is critical as it enhances the generalization capability of the final model, allowing it to perform consistently on unseen data once optimized across all folds.

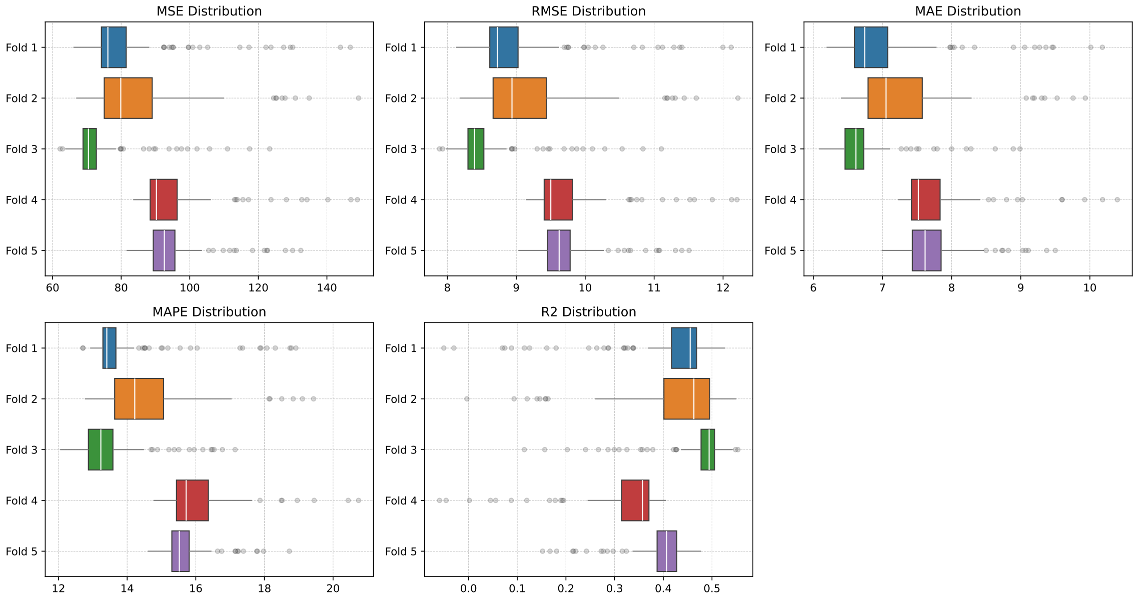


Figure 7.1: Boxplots illustrating the distribution of performance metrics across the 5 folds in the 155 trials.

Regarding the architecture of the MLP integrated with HIPT, we experimented with varying the number of layers from 1 to 6 to determine the most effective structure for this application. Figure 7.2 presents the average performance of these configurations by metric. Notably, MLPs with a simpler structure, specifically those with just 1 layer, generally provided the best optimization across all selected metrics. Detailed metric values for different numbers of layers are provided in Table 7.1. This table also indicates the number of trials conducted by the *Optuna* search algorithm. MLPs with a greater number of layers were tested fewer times, reflecting their less optimal results compared to MLPs with fewer layers.

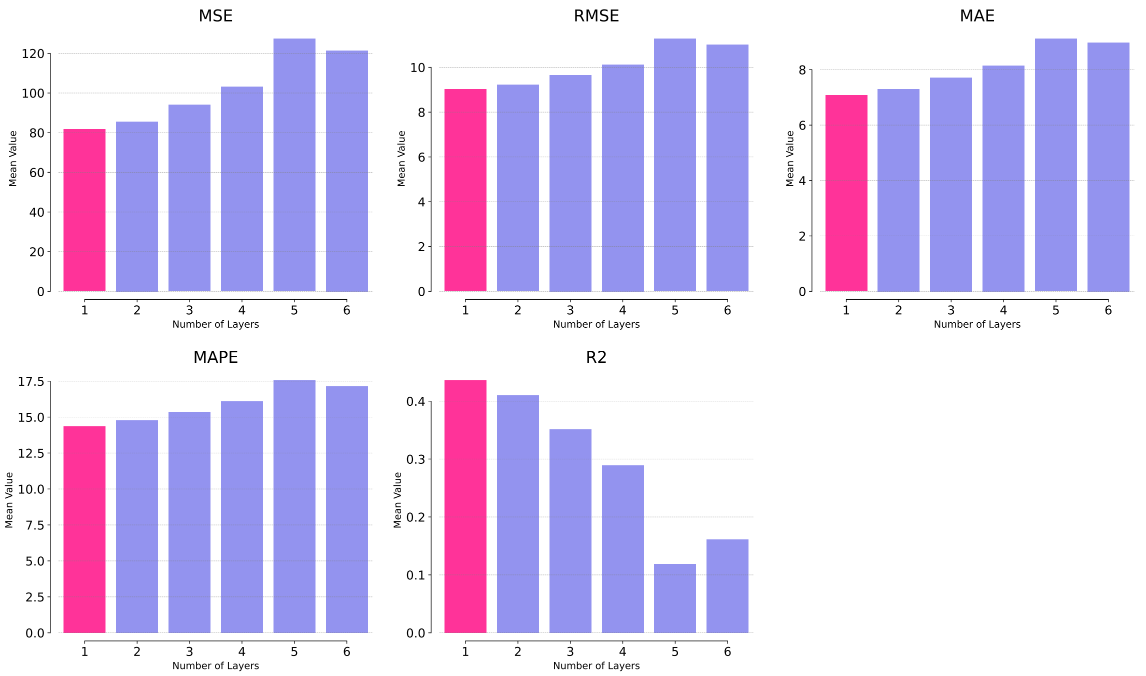


Figure 7.2: Bar graph showing the average metric values for different MLP configurations from 1 to 6 layers. The number of layers yielding the best performance per metric is highlighted in pink.

Layers	Trials	MSE ± Std	RMSE ± Std	MAE ± Std	MAPE ± Std	R2 ± Std
1	81	81.74 ± 4.85	9.02 ± 0.26	7.08 ± 0.24	14.34 ± 0.44	0.44 ± 0.03
2	53	85.57 ± 9.34	9.22 ± 0.47	7.30 ± 0.45	14.75 ± 0.79	0.41 ± 0.06
3	15	94.03 ± 13.64	9.65 ± 0.67	7.72 ± 0.64	15.36 ± 1.08	0.35 ± 0.09
4	3	103.19 ± 16.13	10.12 ± 0.78	8.15 ± 0.77	16.10 ± 1.76	0.29 ± 0.11
5	1	127.32	11.27	9.11	17.56	0.12
6	2	121.39 ± 1.26	11.00 ± 0.05	8.98 ± 0.03	17.12 ± 0.35	0.16 ± 0.01

Table 7.1: Average values and standard deviations of each metric collected in the AP module across the 155 trials for MLP neural networks with 1 to 6 layers. Note that the metric values for MLPs with 5 layers do not include standard deviations because the *Optuna* search algorithm conducted only a single trial for this configuration and, due to its low performance, directed subsequent searches towards MLPs with fewer layers.

With respect to the number of nodes per layer, analysis focused exclusively on MLPs with one intermediate layer which demonstrated superior performance. The optimal number of nodes in this configuration appears to be 256, as evidenced in Table 7.2. This configuration, with 256 nodes in a single intermediate layer, was tested only 7 times. Another possible selection criterion could involve choosing an architecture that underwent extensive testing and exhibited low variance. For instance, MLPs with 1024 nodes in the intermediate layer presented the smallest deviation and underwent the highest number of trials. However, we ultimately chose to proceed with the 256-node model, prioritizing the simplicity of the architecture over a more complex one.

Nodes	Trials	Mean R2 ± Std
32	15	0.41 ± 0.028
64	17	0.43 ± 0.043
128	15	0.44 ± 0.043
256	7	$\mathbf{0.47 \pm 0.013}$
512	11	0.45 ± 0.012
1024	16	0.45 ± 0.011

Table 7.2: Summary of R2 values by number of nodes for MLPs with 1 intermediate layer.

Regarding the HIPT features, Table 7.3 demonstrates that the metrics MAE, MAPE, MSE, R2, and RMSE are effectively optimized by utilizing features labeled $[CLS]^{WSI}_{4096}$ and $[CLS]^{WSI}_{4096,m256}$. As discussed in subsection 3.2.2, the HIPT features are extracted from image fields of view at three distinct scales: 16x16, 256x256, and 4096x4096. The foundational work by Chen et al. [2022] correlates these scales with cellular features, cellular organization, and tissue phenotypes, respectively. Our results indicate that the features $[CLS]^{WSI}_{4096}$ and $[CLS]^{WSI}_{4096,m256}$, which represent embeddings from fields of view of 4096x4096 and a mean of all 256x256 fields, provide the most effective signal for age prediction. This suggests that the broader representations of cellular organization and tissue phenotypes may yield a stronger predictive signal than the more detailed cellular features typically captured in smaller fields of view of view. For a detailed explanation and intuition behind the concept of field of view, refer to section 5.1.

HIPT Feature	MSE	RMSE	MAE	MAPE	R2
$[CLS]^{WSI}_{4096}$	86.12 ± 5.15	9.26 ± 0.27	7.3 ± 0.27	14.81 ± 0.32	0.41 ± 0.04
$[CLS]^{WSI}_{m256}$	111.04 ± 12.8	10.51 ± 0.61	8.45 ± 0.61	16.91 ± 0.9	0.23 ± 0.09
$[CLS]^{WSI}_{4096,m256}$	**83.35 ± 8.58**	**9.11 ± 0.43**	**7.18 ± 0.43**	**14.47 ± 0.68**	**0.42 ± 0.06**

Table 7.3: Average regression metric values for each HIPT feature and for MLPs with one intermediate layer only.

Finally, after discussing insights from the parameter optimization stage, we now present the results of the best model generated during this phase. For this purpose, Figure 7.3 provides scatter plots for each of the five folds, with a fitted line illustrating the relationship between the predicted age values (y-axis) and the actual age values (x-axis) for each subject in the validation sets. The plots allow for a direct visual assessment of the model's performance in terms of accuracy and prediction across the different folds. Average values for the collected metrics across these folds and the parameters used to train this model can be seen in Table 7.4 and Table 7.5.

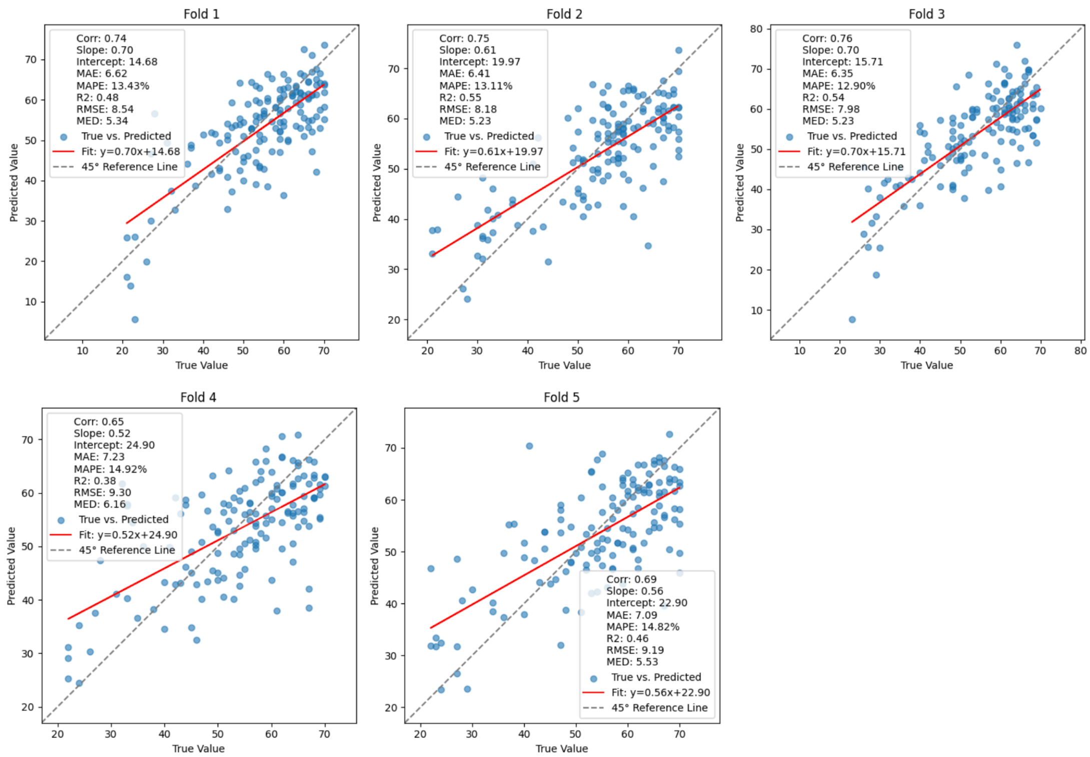


Figure 7.3: The figure displays scatter plots of the best model's performance across the five folds obtained during the parameter optimization stage. The x-axis represents the true ages, and the y-axis the predicted ages for each subject in the validation sets. Legends accompanying each plot provide metrics that assess the performance of each fold individually.

MSE	RMSE	MAE	MAPE	R2
74.90	8.64	6.74	13.84	0.48

Table 7.4: Metric Values of the best model found in the parameter optimization stage. Displayed are the average values of the metrics across the 5 folds.

batch	epochs	learning_rate	dropout_rate	n_layers	n_nodes	hipt_features
32	3005	2×10^{-5}	0.10	1	128	$[CLS]^{WSI}_{4096,m256}$

Table 7.5: Parameter Values of the best model. The parameters presented are consistent across all folds, except for the number of epochs, which represents the average of the optimal number of epochs per fold determined by an early stopping criterion with a patience of 100 epochs.

7.1.0.2 Final Model Analysis

Using the parameters of the best model identified during the optimization phase (??), we executed a final model run for age prediction and evaluated its performance on the test set. Figure 7.4 depicts the scatter plot of predicted versus actual age values. The model achieved performance consistent with that observed across the validation folds (Figure 7.3), with an MAE of 5.88, R2 of 0.52, MAPE of 12.8, and RMSE of 7.78.

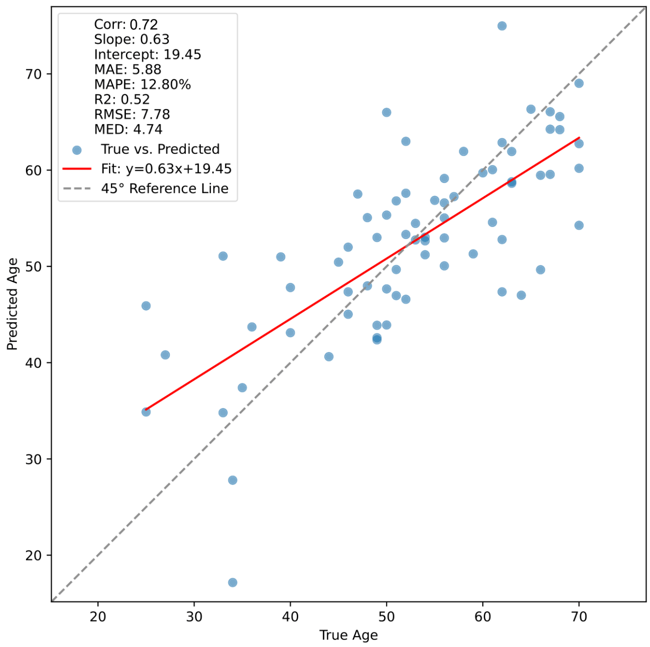


Figure 7.4: The scatter plot of predicted versus actual age values, illustrating the final performance of the Age Prediction module's model.

To further analyze where the model encounters difficulties, we grouped the ages into bins of 5, 10, and 20 years and computed the number of patients and the MAE for each bin. The findings are displayed in Figure 7.5, which includes plots with a dual axis: one for the bars indicating the count of subjects per bin, and another for the line representing the MAE per bin. These visualizations indicate an inversely proportional relationship between the number of subjects in each bin and the associated error; bins with fewer subjects (below 45 years and above 60) exhibit higher MAEs, particularly compared to the central age group spanning from 45 to 60 years.

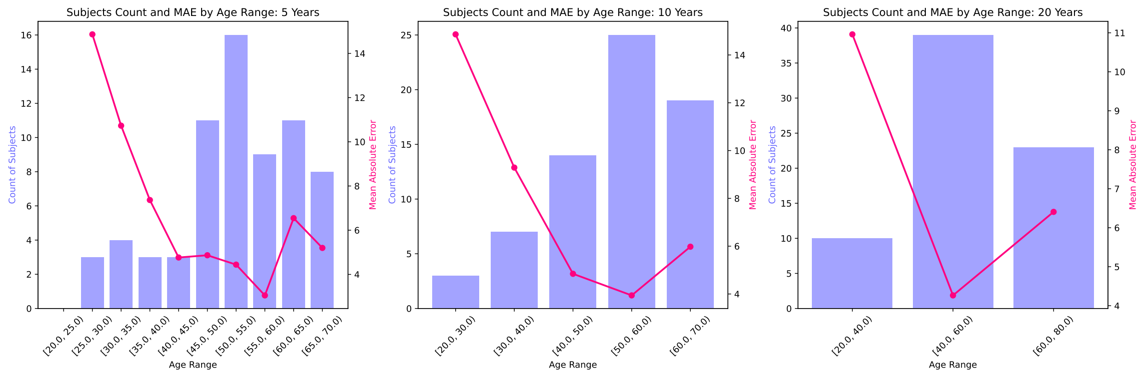


Figure 7.5: The figure displays the distribution of subjects by age bins alongside the corresponding MAE for each bin considering 3 different ranges of age, 5, 10, and 20.

In our previous research [Moraes et al., 2023], we also noted the impact of age imbalance on the model's performance. In the current study, while the LDS technique helped mitigate some of the imbalance issues, outperforming our previous results, challenges remain in predictions for underrepresented age groups. This suggests that future efforts should focus on further methods to address these discrepancies in age distribution.

7.2 Smoker Classification

7.2.1 Parameter Optimization and Selection

In the Smoker Classification (SC) module, we conducted 69 trials, collecting metrics such as F1 Micro, F1 Macro, Specificity, Sensitivity, and Balanced Accuracy using a k-fold cross-validation strategy with $k = 5$. The distribution of these metrics across the folds is presented in Figure 7.6. In general, Folds 1, 3, and 4 exhibited the most stable and effective performance across all metrics, characterized by the highest mean and lowest variance, suggesting that the model configuration or the data in these folds might be particularly well-suited for the model. Conversely, Folds 2 and 5 displayed the weakest performance, evidenced by lower mean values and higher variances, indicating potential issues with data quality or distribution in these folds. This variability among folds is expected and necessary for ensuring that the model optimizing metrics across all folds will be most effective at generalizing to unseen data.

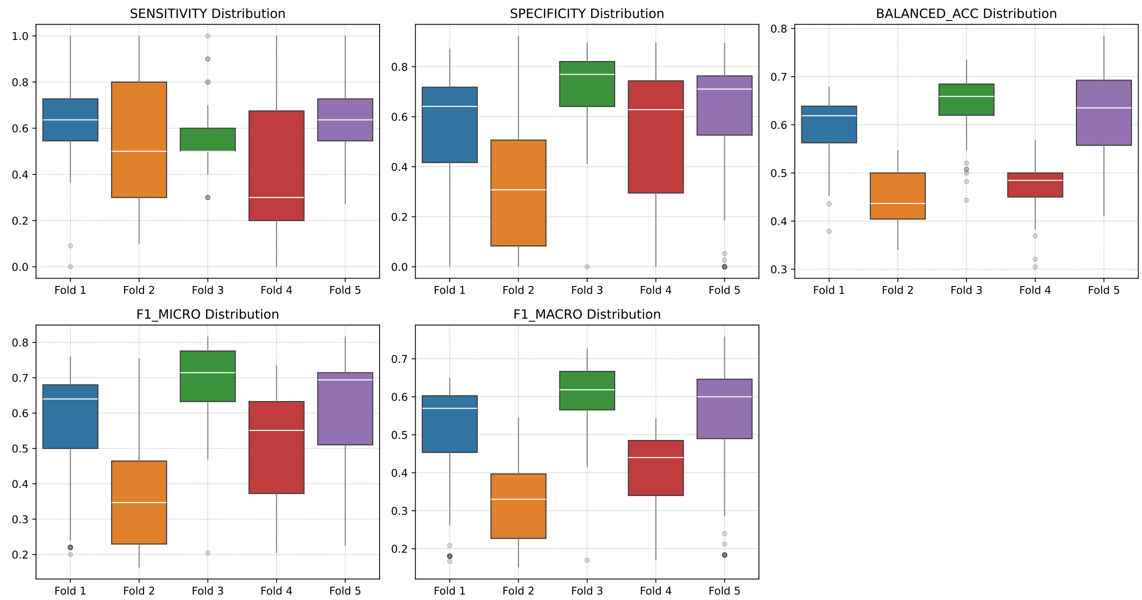


Figure 7.6: Boxplots illustrating the distribution of performance metrics across the 5 folds in the 69 trials of the parameter optimization phase.

Regarding the network's layer configuration, Figure 7.7 highlights in pink the bars corresponding to the number of layers that performed best for each metric. It appears that architectures with 3 layers are most suitable for achieving higher averages in F1 Macro and F1 Micro metrics across the folds. However, for other metrics such as Sensitivity, the best number of layers was 4 and, for Balanced Accuracy, simpler architectures with only one layer proved to be more favorable. Detailed statistics for these metrics, along with their respective standard deviations, are available in Table 7.6.

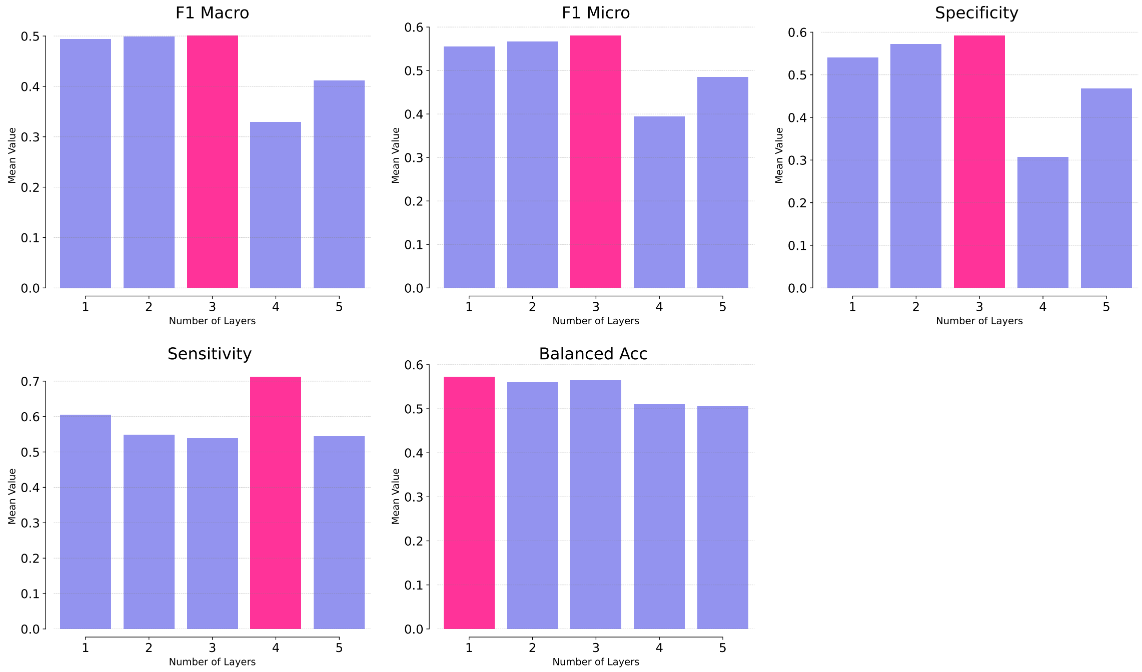


Figure 7.7: Boxplots illustrating the model's performance using networks with different complexities given by the number of layers utilized. The values presented here are the average for each metric.

Layers	Trials	F1 Micro	F1 Macro	Specificity	Sensitivity	Balanced Acc.
1	14	0.49 ± 0.10	0.55 ± 0.13	0.54 ± 0.20	0.60 ± 0.16	0.57 ± 0.04
2	22	0.50 ± 0.06	0.57 ± 0.10	0.57 ± 0.15	0.55 ± 0.16	0.56 ± 0.04
3	20	0.50 ± 0.08	0.58 ± 0.11	0.59 ± 0.19	0.54 ± 0.19	0.56 ± 0.03
4	9	0.33 ± 0.12	0.39 ± 0.15	0.31 ± 0.26	0.71 ± 0.29	0.51 ± 0.05
5	4	0.41 ± 0.13	0.48 ± 0.15	0.47 ± 0.25	0.54 ± 0.24	0.51 ± 0.05

Table 7.6: This figure displays the performance metrics for the smoker classification module, showing mean values and standard deviations for each metric, organized by the number of layers in the network. This presentation helps illustrate how the complexity of the network impacts its predictive accuracy.

As shown in Table 7.7, we present the average values of the metrics collected, grouped by feature type. The $[CLS]^{WSI}_{4096,m256}$ feature generally performed best in terms of F1 Macro, F1 Micro, and Specificity metrics. For Balanced Accuracy, this feature performed comparably to $[CLS]^{WSI}_{4096}$. Regarding Sensitivity, the best performing feature is $[CLS]^{WSI}_{4096}$.

Sensitivity is critical as it measures the proportion of actual positive cases (smokers) correctly identified, while Specificity indicates the proportion of true negative cases (non-smokers) accurately detected. These results suggest that models trained with the

$[CLS]^{WSI}_{4096}$ features are particularly effective at identifying smokers. Conversely, models using $[CLS]^{WSI}_{m256}$ features excel in recognizing non-smokers. However, models trained with $[CLS]^{WSI}_{4096,m256}$ features demonstrate a balanced performance between Sensitivity and Specificity, indicating a robust capability to distinguish between these classes. Moreover, considering the values for F1 Macro and F1 Micro, the $[CLS]^{WSI}_{4096,m256}$ features are deemed most suitable for this classifier, providing a strong signal for distinguishing between smoker and non-smoker classes.

HIPT Features	F1 Macro	F1 Micro	Specificity	Sensitivity	Balanced Acc
$[CLS]^{WSI}_{4096}$	0.46 ± 0.08	0.51 ± 0.10	0.46 ± 0.16	0.68 ± 0.15	0.57 ± 0.03
$[CLS]^{WSI}_{m256}$	0.34 ± 0.10	0.41 ± 0.13	0.37 ± 0.24	0.56 ± 0.29	0.47 ± 0.03
$[CLS]^{WSI}_{4096,m256}$	0.50 ± 0.09	0.58 ± 0.12	0.58 ± 0.20	0.55 ± 0.18	0.57 ± 0.03

Table 7.7: This figure illustrates the average values of each performance metric for the smoker classification module, grouped by HIPT feature. It highlights how different feature configurations influence the overall performance of the model.

Based on the analyses conducted during the optimization phase, we selected the model that demonstrated the optimal balance between Sensitivity and Specificity, while also achieving the highest F1 scores, as the best model. The confusion matrices for the 5 folds of this model are depicted in Figure 7.8. Analysis of these matrices shows that only Folds 1 and 3 achieved optimal performance along the main diagonal, indicating a balanced performance between the true positives and true negatives. In contrast, Folds 4 and 5 prioritized Sensitivity but suffered a considerable loss in Specificity. Conversely, Fold 2 exhibited notably low performance in terms of Sensitivity.

The aggregate performance metrics across all 5 folds for this best model are summarized in Table 7.8. The parameters used for this model, which include layer configurations, learning rates, and other relevant settings, are detailed in Table 7.9.

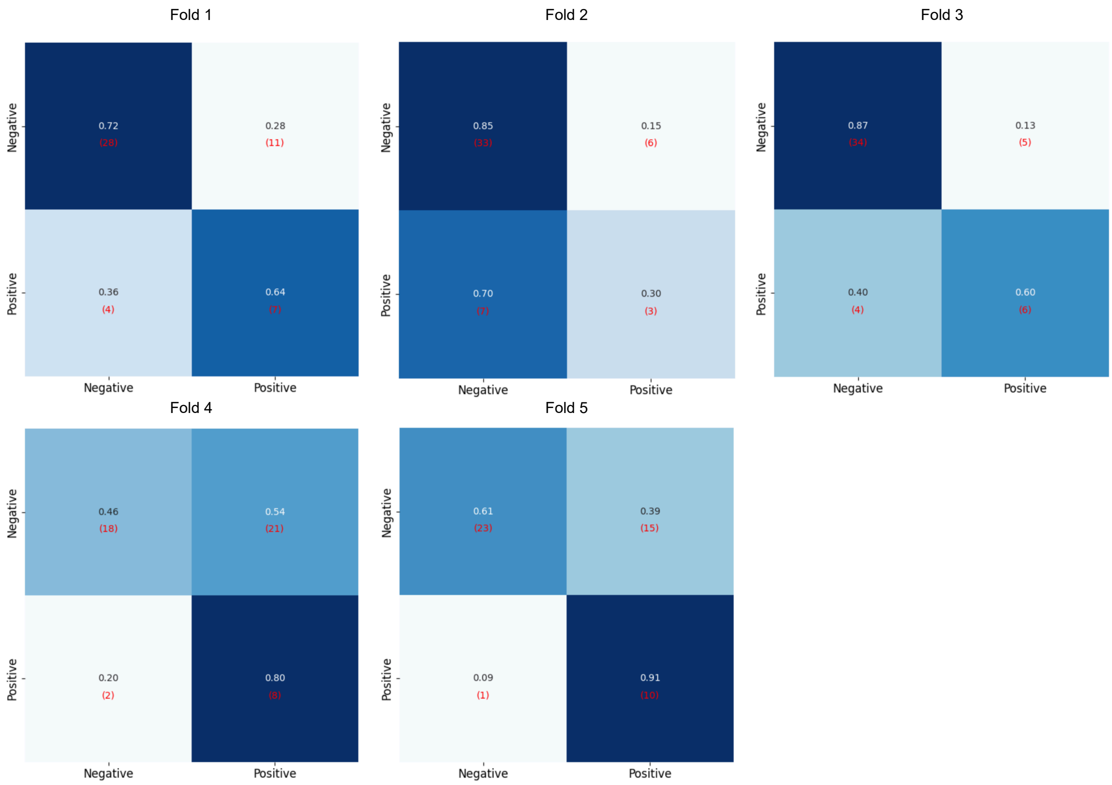


Figure 7.8: Confusion matrices for the 5 folds of the best model from the optimization stage in the smoker classification module, illustrating the model's performance across different metrics. This visual representation highlights the balance achieved between Sensitivity and Specificity, as well as the challenges in maintaining consistent performance across all folds.

F1 Macro	F1 Micro	Sensitivity	Specificity	Balanced Acc.
0.59	0.60	0.45	0.75	0.60

Table 7.8: Metric Values of the best model found in the parameter optimization stage for the SC module. Displayed are the average values of the metrics across the 5 folds.

batch	epochs	learning_rate	dropout_rate	n_layers	n_nodes	hipt_features
32	166	2×10^{-5}	0.21	1	344	$[CLS]^{WSI}_{4096,m256}$

Table 7.9: Parameter Values of the best model in the optimization phase for the SC module. The parameters presented are the same across all folds, except for the number of epochs, which represents the average of the optimal number of epochs per fold determined by an early stopping criterion with a patience of 100 epochs.

7.2.2 Final Model Analysis

Using the parameters from the best model identified during the optimization phase (see Table 7.9), we trained the final model for the SC module. The performance of this model is depicted in a confusion matrix in Figure 7.9. The matrix illustrates that the final model effectively optimized the main diagonal, achieving a satisfactory balance between sensitivity and specificity. This balance was achieved through a combination of using the best parameters found during the optimization phase and implementing a post-processing step for threshold optimization. Specifically, we calculated the optimal threshold for each fold of the best model in the optimization stage based on AUC values, then averaged these thresholds to determine the most effective overall threshold for use in the final model. This optimized threshold, rather than the conventional cutoff of 0.5, was then applied to the final model.

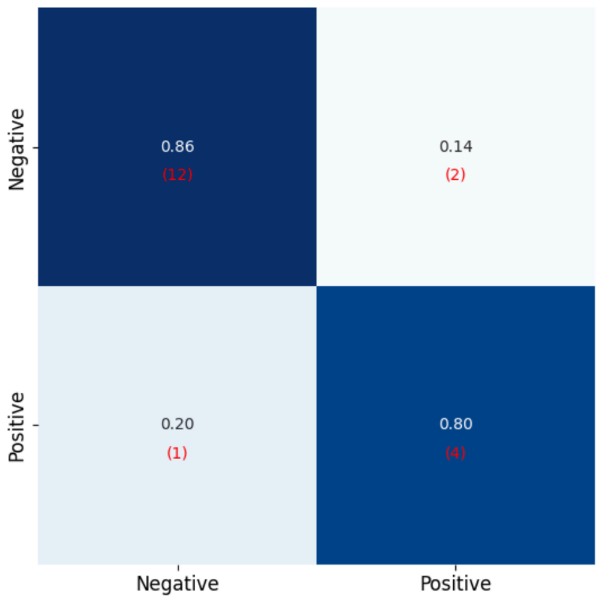


Figure 7.9: Confusion matrix for the final model of the SC module, demonstrating the optimized balance between sensitivity and specificity achieved in testing.

It is important to note that from the common test set of 72 subjects, we were only able to assess the performance of this model on 19 subjects, as the remaining subjects belong to the 'Ex Smoker' and 'Unknown' categories. Subsequently, in the GEN stage, we used this model to classify these subjects into 'Smoker' and 'Non Smoker'. This strategy was previously employed in Ramirez et al. [2024], and we decided to maintain the same approach here. We will revisit this point in the discussion chapter to further justify and explore the implications of this decision.

The average results of this final model for each one of the metrics utilized are summarized in Table 7.10. As we can see, the model obtained a F1 Macro score of 0.80 and an F1 Micro score of 0.73, indicating a robust performance in precision and recall across both categories. Sensitivity and specificity scores are 0.80 and 0.86, respectively, which validates the model's accuracy in correctly identifying positive (smokers) and negative cases (non-smokers).

F1 Macro	F1 Micro	Sensitivity	Specificity	Balanced Accuracy
0.80	0.73	0.80	0.86	0.83

Table 7.10: Performance Metrics of the Final Model for the SC Module

7.3 Diagnosis Retrieval

Recapping briefly what was discussed in the methodology section, we conducted 100 trials within a defined search space as outlined in Table 6.3. For each trial, we collected various metrics, detailed in Table 6.4. After evaluating these metrics, we identified an optimal set of parameters that led to the development of the best model for the Diagnosis Retrieval (DR) module. We will first present the results from the trials that guided our selection of these parameters, followed by the outcomes from the best model trained with these optimal parameters. For further details, please refer to subsection 6.2.3.

7.3.1 Parameter Optimization and Selection

Selecting the optimal set of metrics for our study was a complex process. As demonstrated in Figure 6.5 (a), the Diagnosis Retrieval (DR) models retrieve the top $K = 10$ most similar pathologist notes for each of the j images in the validation set. For each j tuple $(x_j^I, \{x_i^C\}_{i=1}^K)_j$, we applied caption metrics (CPT) and similarity metrics (SIM) to each pair (x_j^I, x_j^C), subsequently calculating the average for each set of j as illustrated in Figure 6.5 (b.3). However, in our specific context of evaluating pathology captions, the accuracy of capturing pathology categories in the captions is more crucial than the structural integrity of the caption itself. Therefore, it was insufficient to match numerous words while missing the associated pathologies. To address this, we also incorporated classification metrics

(CLS), such as precision, recall, and f1-score, which consider the categories of the most frequent pathologies among the K retrieved captions.

In Figure 7.10, the box plot displays the distribution of each selected metric across our sample of 100 trials. Some metrics, such as BLEU-1, BLEU-2, ROUGE, Hamming, Manhattan, and Euclidean, consistently exhibit elevated values. Regarding CPT metrics, BLEU scores, particularly those considering smaller n-grams like 1 and 2, perform well because they easily match single words and pairs of words, which frequently recur across all pathologist notes. The SIM metrics also have high values because they compare embedding vectors that already possess a high degree of similarity, as they were selected based precisely on the criterion of similarity.

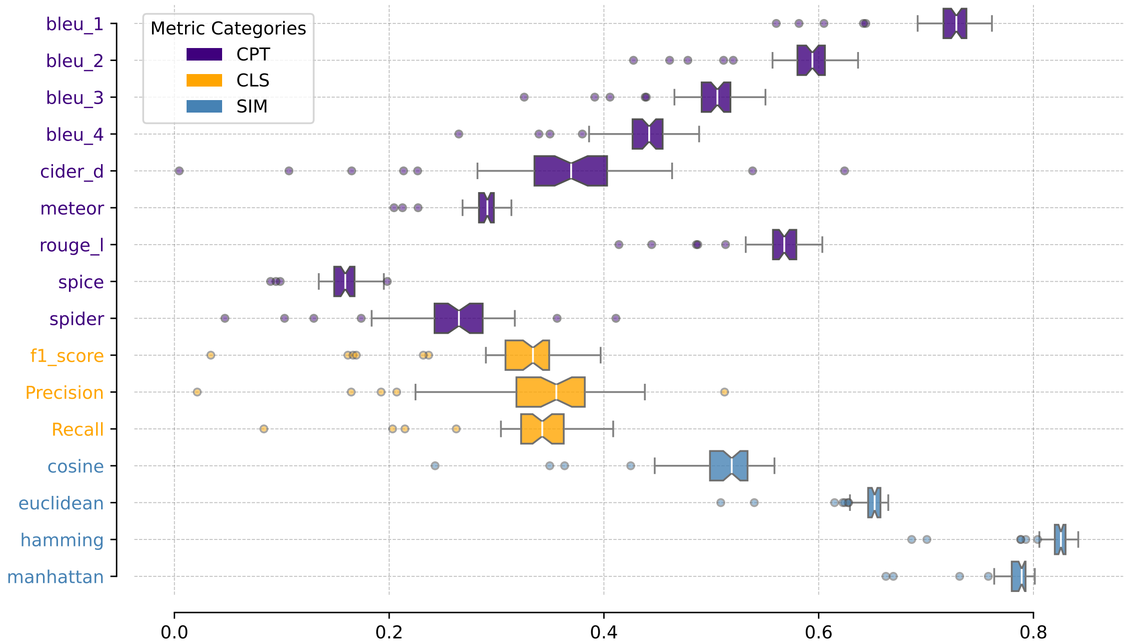


Figure 7.10: Box plot of the metrics selected for evaluation of the 100 parameter optimization trials on the validation set. The metrics are categorized into three groups: caption (CPT), similarity (SIM), and classification (CLS).

Moreover, Figure 7.10 reveals significant variability among the metrics, with differences of up to 60 percentage points between the lowest (SPICE) and the highest (Euclidean) average metric values. Despite their different natures and purposes, which led to their categorization into three groups, this wide variance indicates that relying solely on individual metrics can be a dangerous approach. In our medical context, we place greater trust in CLS metrics, which more accurately reflect the quantity of pathology categories accurately identified during each validation instance.

Regarding the parameters batch size and learning rate, Figure 7.11 illustrates the average values of each metric for each batch size and learning rate selected in our search space. As shown in Figure 7.11 (a), there is not a significant variation in the performance of these metrics with changes in batch size. However, for some metrics, particularly those in the classification (CLS) category, a subtle improvement is noticed with batch sizes ≥ 32. In Figure 7.11 (b), the improvement achieved with learning rates $\leq 2 \times 10^{-4}$ is more evident across all metrics.

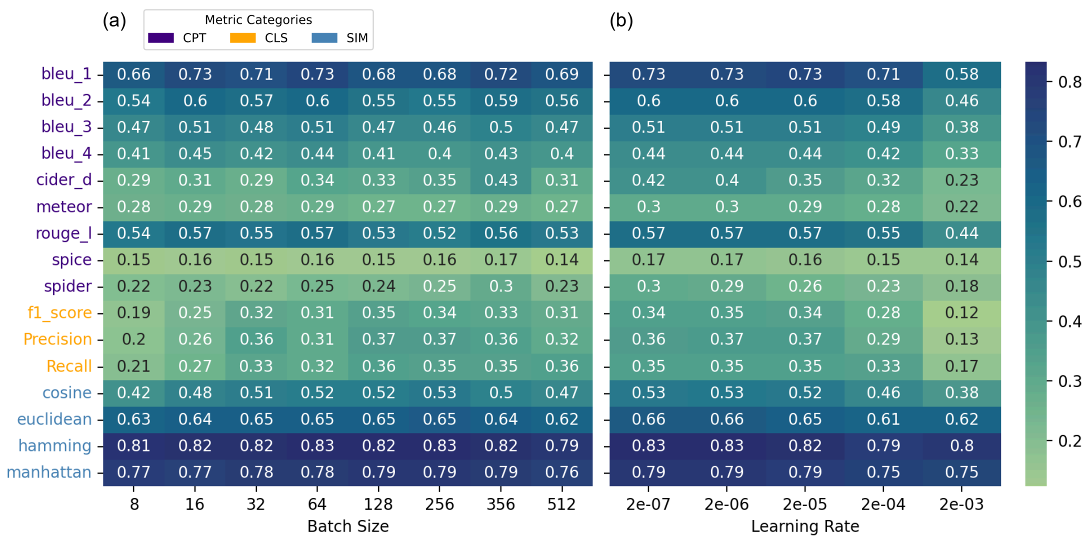


Figure 7.11: The heatmaps display the average values of each metric across 100 validation trials by batch size (a) and by learning rate (b), spanning three categories of metrics: caption (CPT), similarity (SIM), and classification (CLS).

Figure 7.12 presents the average f1-score values across these 100 trials for each of the HIPT features tested. The bars indicate that consistently, features $[CLS]^{WSI}_{4096}$ outperformed the others, with $[CLS]^{WSI}_{m256}$ coming in second.

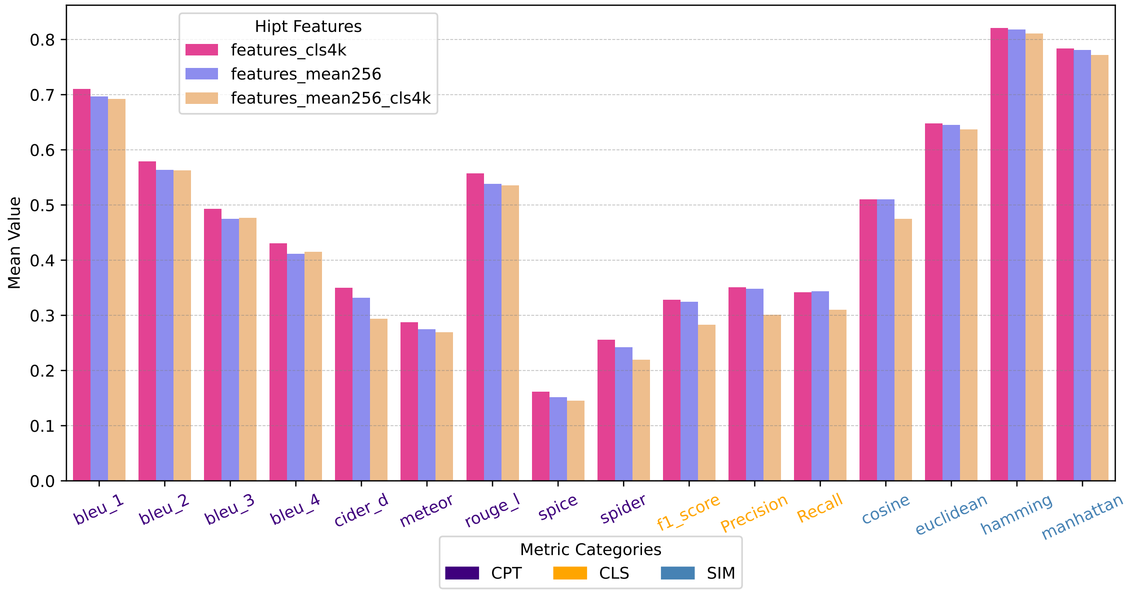


Figure 7.12: This bar graph shows the average f1-score values for each HIPT feature tested, categorized into three metrics groups: caption (CPT), similarity (SIM), and classification (CLS).

These results, particularly those in Figure 7.12, were surprising as we initially expected better performance from the features $[CLS]^{WSI}_{4096,m256}$. These features incorporate more information as they not only include embeddings of the 4096x4096 tiles that represent global structures but also average embeddings of the 256x256 tiles that highlight local patterns of cellular arrangements. Despite this intuition, we opted to train the final model with the $[CLS]^{WSI}_{4096}$ feature.

Finally, the number of training epochs was optimized using the technique of early stopping, which terminates the training process as soon as the loss begins to worsen. We employed a patience of 3 epochs, which means if there is no improvement in the loss after 3 consecutive epochs, training is halted. In Figure 7.13 (a), the top 10 best trials out of 100 on the validation set are shown, selected based on the f1-score. It is evident that each trial varies in duration; some last just over 10 epochs, others extend beyond 30, and one even persisted for over 50 epochs before being terminated by the early stopping mechanism. This variation in the optimal number of epochs is directly influenced by the variables of batch size and learning rate.

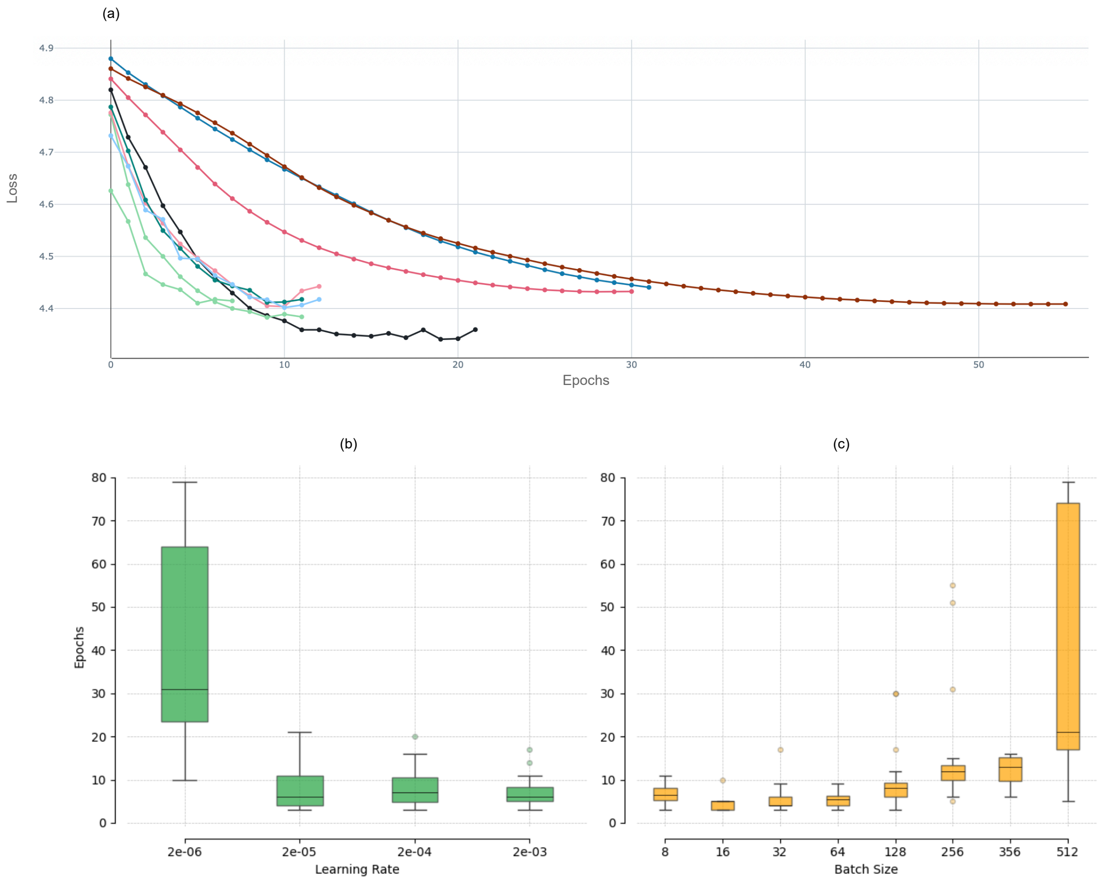


Figure 7.13: Analysis of the number of epochs per trial. Panel (a) displays the top 10 best trials of the Diagnosis Retrieval (DR) module on the validation set, evaluated based on f1-score. The x-axis indicates the number of training epochs and the y-axis shows the values of the Contrastive Loss. Given the application of the early stop technique, each trial was conducted over a different number of epochs, tailored to the optimal duration for each setup. Below, the boxplots in panels (b) and (c) illustrate the distributions of the number of epochs across the 100 trials, categorized by variations in learning rate and batch size, respectively.

In Figure 7.13 (b) and (c), we observe the distribution of the number of epochs with respect to different settings for learning rate and batch size. From these boxplots, it is apparent that a larger batch size and learning rate tend to increase the number of training epochs required.

7.3.2 Final Model Analysis

Based on the previously presented results and specifically the performance on the f1-score, we selected the best parameters for the final model: a batch size of 256, a learning rate of 2×10^{-5}, and the HIPT feature $[CLS]_{4096}^{WSI}$. The number of training epochs was set to 10, as the top-performing trials (Figure 7.13 (a)) all lasted at least this duration. Moreover, this epoch count aligns with the average values seen in the boxplots of Figure 7.13 (b) and (c) for the selected learning rate and batch size.

With these parameters, we combined the training and validation sets to train the final model. Figure 7.14 displays the mean value for each metrics obtained with this model across all test subjects. For the CLS metrics, the mean is taken over pathology categories. Notably, the model achieved superior results across several metrics, particularly the f1-score, which reached a value of 0.56.

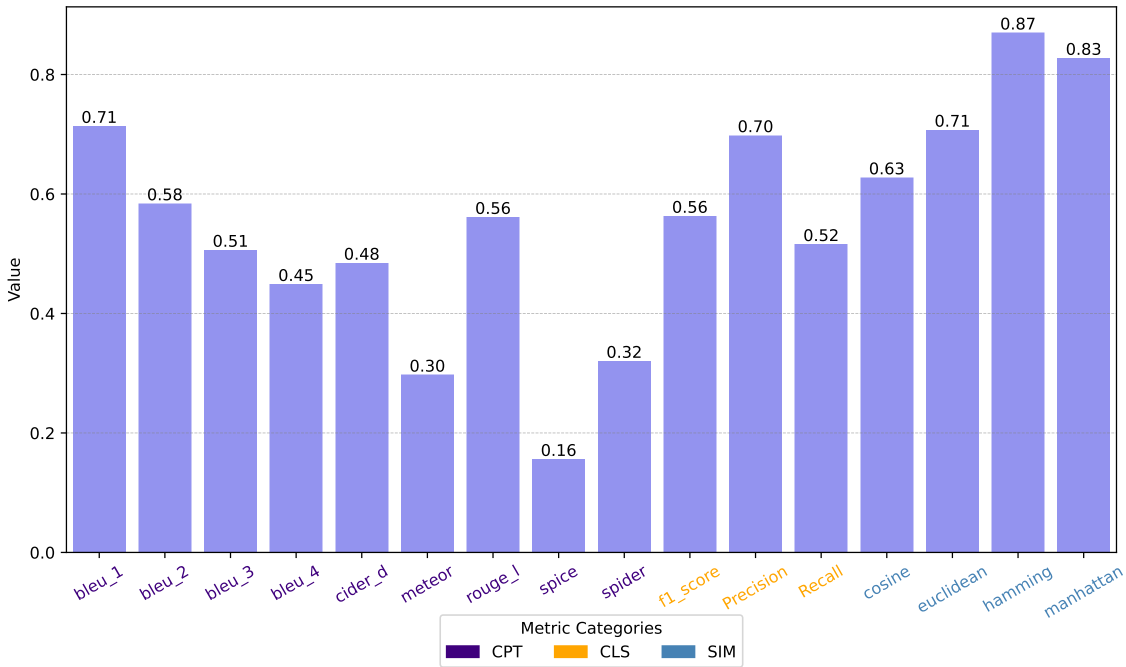


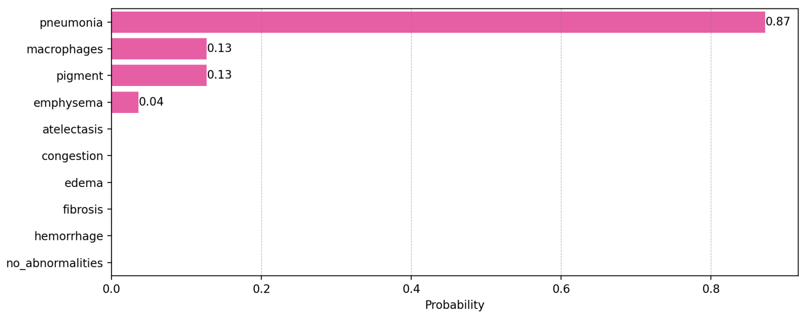
Figure 7.14: Final DR model's performance across various metrics. The bars display the mean values of each collected metric (CPT, CLS, and SIM) across all test subjects.

Figure 7.15 presents an example of outcome of the DR model. Panel (a) shows the original diagnosis and the 10 retrieved diagnoses. Pathologies that match are highlighted in green, while discrepancies are marked in red. Panel (b) depicts the probabilities for each pathology, generated from the frequency of occurrence in the retrieved diagnoses as outlined in Figure 6.5 (b.3). Despite some retrieved diagnoses including pathologies not present in the original, the most pronounced pathology, Pneumonia, matches the original diagnosis. In panel (c), we observe the CPT metrics values for this case, noting that metrics such as BLEU, METEOR, and ROUGE, while consistently high in the validation sets, exceed the upper limits of their respective distributions (Figure 7.10).

(a)

Sample ID	Original Diagnosis	Retrieved Diagnoses
GTEX-WFG8-0926	Patchy acute pneumonia. Pneumonia is the main finding in this slide.	[Patchy bronchopneumonia. Pneumonia is the main finding in this slide., Patchy bronchopneumonia. Pneumonia is the main finding in this slide., Patchy bronchopneumonia. Pneumonia is the main finding in this slide., Includes larger vessels, many pigmented macrophages. Main findings are macrophages, pigment., Bronchopneumonia. Pneumonia is the main finding in this slide., Bronchopneumonia. Pneumonia is the main finding in this slide., Bronchopneumonia. Pneumonia is the main finding in this slide., Patchy bronchopneumonia, heart failure cells, prominent arteries. The main findings are heart failure cells and pneumonia., Bronchopneumonia and patchy emphysema. The main findings are emphysema and pneumonia., Heart failure cells, mild patchy pneumonia. The main findings in this slide are heart failure cells, pneumonia.]

(b)



(c)

Metric	Value
BLEU-1	0.91
BLEU-2	0.80
BLEU-3	0.75
BLEU-4	0.72
METEOR	0.49
ROUGE-L	0.86
CIDEr-D	0.92
SPICE	0.32
SPIDER	0.62

Figure 7.15: Example of the outcome from the diagnostic retriever model. In (a), we observe the original diagnosis alongside the 10 retrieved diagnoses with the final DR model, highlighted in green for matches and in red for discrepancies. In (b), the probabilities for each pathology are shown, derived from the frequency of descriptions in (a). In (c), the values of CPT metrics for this specific case are displayed.

Regarding the probability distributions of each pathology as generated by the model, Figure 7.16 reveals significant variances. The density plot in (a) clearly shows that certain pathologies such as Congestion, Edema, Hemorrhage, and Macrophages are more frequently predicted with high probabilities, notably within the 80% to 100% range. These pathologies dominate the space in the figure. The corresponding boxplot in (b) provides a clearer visualization of the distribution across the four main quartiles and indicates the presence of outliers in the data; notably, Edema displays seven outlier points extending the tail to the right, while Hemorrhage shows one outlier at a probability of 100%, significantly elongating the right tail of this distribution. Conversely, other pathologies like Emphysema, Fibrosis, and Pigment, along with the absence of abnormalities, are rarely predicted with probabilities above 80%.

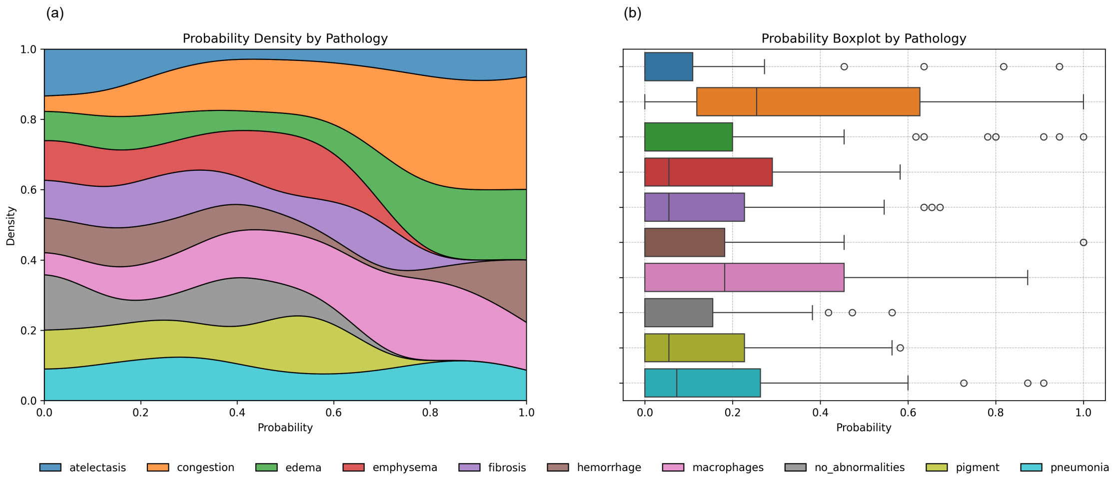


Figure 7.16: Probability distribution of the occurrence of pathologies in the test set. Panel (a) is a density plot of these probabilities by pathology, while panel (b) presents the box plots of these distributions. The figures complement each other by providing a visual representation of the areas of higher and lower density for each pathology and a precise depiction of the quartiles, respectively.

These probabilities are derived from the frequency of appearance of pathologies among the 10 retrieved diagnoses for each subject. The noted imbalance reflects a certain inaccuracy in the model's ability to identify text embeddings that correspond most closely with the image embeddings of patients exhibiting conditions such as pigment, emphysema, fibrosis, and others which manifest with low probability. As a result, the retrieved texts for these conditions are more varied, and the probabilities generated for them are consequently lower.

Given the observed asymmetry in the probabilities generated for each pathology, one potential remedy could be to define customized probability thresholds for each pathology, rather than relying on a universal threshold. The evidence presented in Figure 7.16 suggests that a universal threshold is impractical for our purposes. We could establish these individual thresholds by using the validation set as a reference and calculating the optimal cutoff points based on the Area Under the Curve (AUC) for each pathology. However, this approach may still not be entirely suitable for healthcare applications like ours, where the stakes of diagnostic accuracy are particularly high. In such contexts, it is crucial to minimize the risk of false positives, that is, incorrectly stating that a person does not have a pathology when, in fact, they do. Therefore, in our specific scenario, prioritizing the reduction of the false-negative rate is most appropriate.

Using the optimized thresholds for each pathology, arrive at the consufion matrices in Figure 7.17. For six out of the ten conditions shown (Actelectasis, Congestion, Hemorrhage, Macrophages, No Abnormalities, and pigment) the main diagonal, which represents correct classifications, was maximized. This indicates a high true positive rate and a successful application of the customized thresholds for these conditions. However, for Pneumonia, our technique was unable to minimize the false-negative rate effectively. Finally, for Fibrosis, Edema and Emphysema, while we successfully minimized the false-negative rate, this was followed by a significant increase in the false-positive rate.

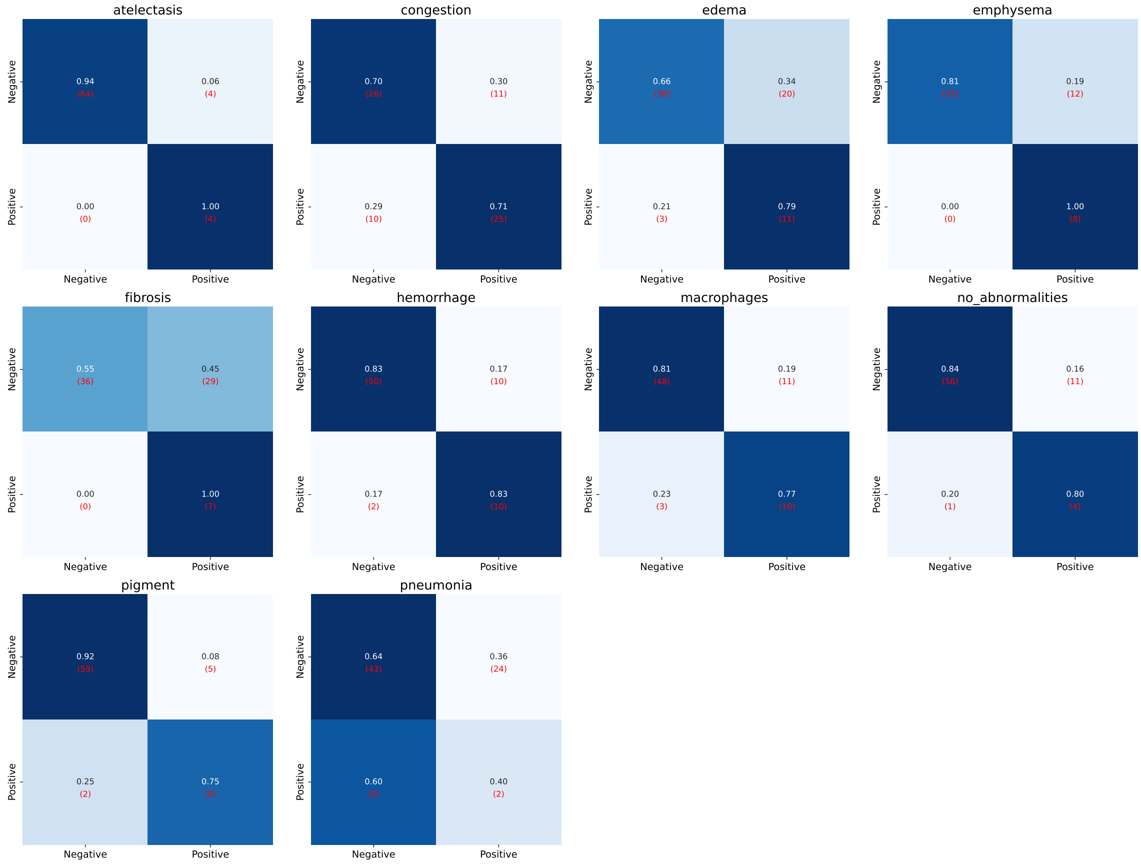


Figure 7.17: Confusion matrices for the classification of pathologies after threshold optimization.

7.4 Generation Stage

This stage differs significantly from the INF stage, which employs individually trained modules for specific inference tasks. Here, we evaluate the functionalities of agents involved in the information enrichment process and assess the generated reports using the predefined scoring scale detailed in Figure 6.6.

7.4.1 Information Enrichment and Report Generation

The information enrichment process involves two specialized agents: a Questioner Agent and a Q&A Agent. The Questioner Agent is tasked with generating relevant questions based on the input caption to deepen the analysis of the pathology presented. This agent operates independently without direct access to the vector-store containing scientific publications, relying instead on its own knowledge base acquired during its training to formulate questions that can effectively probe the underlying medical issues.

The Q&A Agent, on the other hand, answers the questions posed by the Questioner Agent. Unlike its counterpart, this agent utilizes the RAG strategy and has exclusive access to a local vector-store containing scientific publications from databases like PubMed, BioRxiv, and other academic resources. This setup enables the Q&A Agent to provide informed and contextually relevant answers, which are crucial for the subsequent report generation.

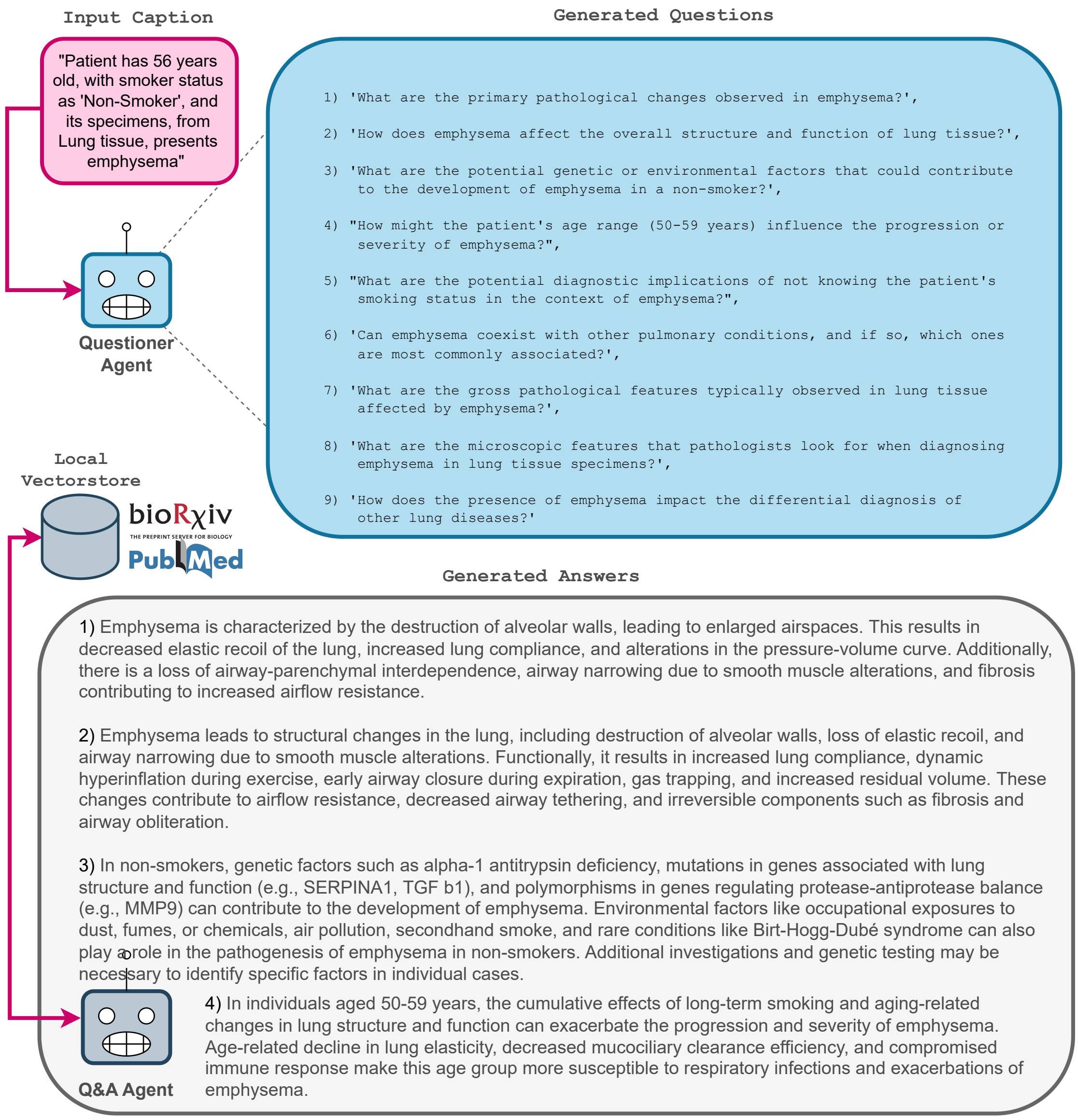


Figure 7.18: This figure illustrates the outputs from the Questioner and Q&A agents during the enrichment stage. The top left pink box shows the input caption from the AGG stage, which is processed by the Questioner Agent to generate pertinent pathological questions. The Q&A Agent then responds to each question using the RAG strategy, drawing on its exclusive access the vector-store of scientific literature.

Following the question-and-answer session, the information is passed to the final agent responsible for generating the final report. This agent compiles the insights from the enriched dialogue into a detailed and coherent medical narrative, as depicted in Figure 7.19.

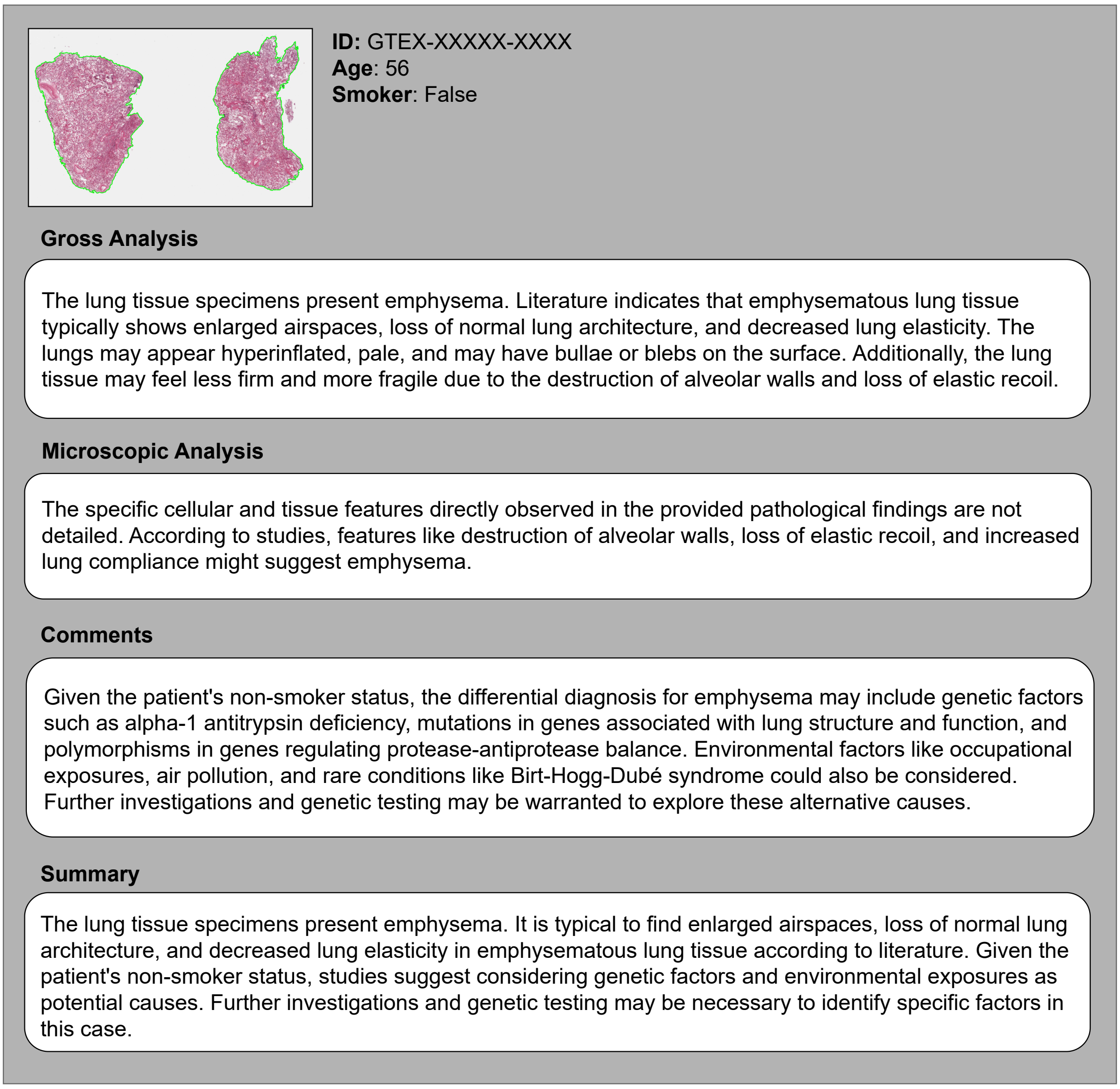


Figure 7.19: Example of a final report generated by the Report Generator Agent, utilizing the dialogue of questions and answers provided by the preceding agents and an input caption used as reference. This report synthesizes the enriched information into a final document that encapsulates the diagnostic insights. Given the confidentiality of the smoker status and specific age of the subjects in the GTEx database, we are omitting the ID of this patient, even though the age and smoker status presented here are solely predictions.

7.4.2 Report Evaluation

As previously discussed, we defined a scoring scale (Figure 6.6) to evaluate our generated reports. We employed this scale in two distinct approaches to assess the reports concerning classification and informativeness. Regarding the classification aspect, Figure 7.20 illustrates the percentage and absolute values assigned by the scorer agent responsible for evaluating the reports. It is evident that the majority of scores fall into category 2, indicating that while the reports correctly mention pathologies, they tend to include two or more conditions than actually exist. The second highest bar, at 24%, is in category 3, which reflects minor errors generally due to missing intensity adjectives for pathologies or the addition/omission of up to one pathology.

Categories 1 and 4, constituting 21% and 12% of the scores respectively, represent opposite outcomes. Category 1 indicates a complete mismatch, where reports add pathologies unrelated to the specimens, and category 4 represents totally accurate diagnoses without any errors or omissions.

These results are focused solely on pathological accuracy and do not account for potential errors related to the demographic data of the subjects. Consequently, they primarily reflect the performance limitations of the DR module, which exhibits high specificity for each pathology considered. Although not immediately apparent in Figure 7.17, the complexities of pathology prediction become apparent when these pathologies are combined to generate a multilabel for each patient. The individual false negative rates of each pathology are propagated and combined during this process, thereby amplifying their impact factor when we compose the vector of multilabel pathologies for each WSI. This factor can explain the high scores in category 2, indicating an over-prediction of pathology possibilities.

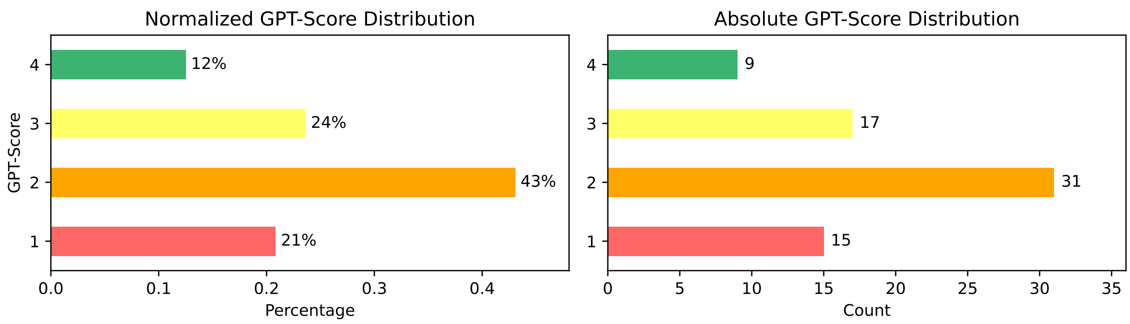


Figure 7.20: This figure shows the distribution of classification scores for the reports, as evaluated by the scorer agent. Each bar represents the percentage of reports falling into each classification category.

Figure 7.21 elucidates the rationale behind the Scorer Agent's decisions when assigning specific scores. We analyzed a random sample of 7 scores associated with 7 of the 72 generated reports. The most penalized cases (score 1) occurred when the agent failed to identify any common elements between the generated report and the original diagnosis's pathologies. Cases receiving score 2 involved correct identification of some pathologies, but also erroneous inferences about additional pathologies not present in the original diagnosis. For score 3, the example shows that while Congestion and Edema were correctly identified, the erroneous assumption of Hemorrhage was considered a minor error. Cases awarded score 4 matched all original pathologies and conditions flawlessly, with no errors or omissions.

ID	Score	Justification
48	2	The generated report identifies congestion but misses sloughed bronchial epithelium, adding unrelated pathologies like edema, fibrosis, hemorrhage, and pneumonia.
17	3	Correctly identifies congestion and edema but adds hemorrhage, which is not mentioned in the original note. Minor error.
43	1	The generated report fails to mention chronic inflammatory cells and hemosiderin-laden macrophages, focusing instead on emphysema, which is unrelated.
32	4	Accurately identifies all main pathologies: congestion, macrophages, and pigment. No errors or omissions noted.
8	1	The generated report fails to mention congestion, the main finding, and instead focuses entirely on emphysema, an unrelated pathology.
5	2	The report identifies macrophages but misses congestion, focusing instead on atelectasis and fibrosis, which are not mentioned in the correct note.
57	4	The generated report accurately identifies all main pathologies: congestion, emphysema, and hemorrhage, without any errors or omissions.

Figure 7.21: Justifications generated by the Scorer Agent for each score choice, demonstrating how each report was evaluated based on its accuracy in classifying pathologies.

In terms of informational accuracy, the reports' performance is evaluated on how effectively they provide quality information about the pathologies mentioned in the input caption. As shown in Figure 7.22, the category 4 of the GPT-Score reaches nearly 100%, suggesting that the enrichment phase of the GEN stage successfully integrates relevant information from the academic literature available in the local database.

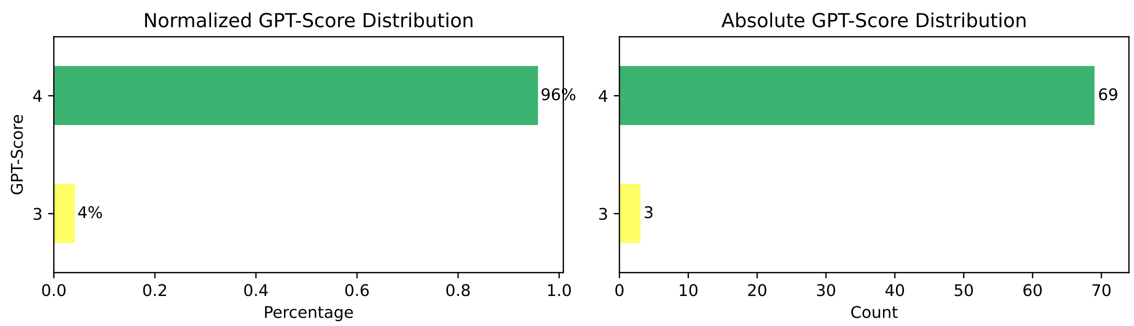


Figure 7.22: Illustration of data used and produced during the information enrichment stage. The figure shows a caption from the AGG stage, processed by the Questioner Agent to generate pertinent questions, followed by the Q&A agent's answers utilizing the RAG strategy.

Figure 7.23 provides examples with justifications from the Scorer Agent for their score selections, focusing on the informational aspect of the reports.

ID	Score	Justification
48	4	The report accurately describes congestion, edema, and hemorrhage, linking them to smoking history and relevant conditions like ARDS.
37	3	The report accurately describes the absence of abnormalities and potential microscopic changes in ex-smokers but lacks specific microscopic findings.
43	4	The report accurately describes fibrosis and pneumonia, including their gross and microscopic features, and discusses clinical implications.

Figure 7.23: Justifications provided by the Scorer Agent for each score, highlighting the performance of the reports in delivering comprehensive and accurate information about the pathologies.

Finally, we include in Appendix C additional examples of generated reports, showcasing both successful cases and instances where the original diagnosis was not accurately identified. In the discussion section, we will revisit these performances, examine the associated limitations, and explore potential improvements for future versions of the system.

Chapter 8

App Development

In this chapter, we aim to present the application developed to embed our system. The application was developed as a proof of concept to demonstrate how we might lower the access barriers to the results of our research, facilitating the use and testing of our work by end users who may not possess advanced knowledge in computer science.

8.1 Overview

The application we developed encapsulates the entire system described throughout this thesis, from feature extraction from images (Figure 5.3) to the generation of final reports. It is structured into four main parts, each designed to enhance the user experience. These parts are organized into different pages within the application and will be explored sequentially.

8.1.1 Image Upload and Preprocessing

The analysis of WSIs begins on the main page of the app, where users can upload an image up to 1GB in size. Typically, WSIs of lung tissue range from 200MB to 600MB, depending on the sample (Figure 8.1 (1)).

Following the upload, the system automatically detects and displays associated metadata such as the scanner used, pixel size at different levels, and more (Figure 8.1 (2)). Concurrently, the WSI is segmented into patches measuring 4096×4096 pixels, a necessary step for subsequent processing with the HIPT system. Post-segmentation, details such as the number of patches generated, the size of each patch, and the original image dimensions are displayed on the screen for user verification (Figure 8.1 (3)).

Figure 8.1: Initial application screen. Step 1 involves the WSI upload, triggering the release of metadata (2). The WSI then undergoes segmentation, with relevant metadata from this phase also displayed (3).

Additionally, on this initial page, users have the ability to navigate through uploaded WSIs for visual analysis. Figure 8.2 demonstrates this functionality, starting with lower resolution levels and broader fields of view at the top, and progressively increasing resolution while narrowing the field of view to focus on specific structures at the bottom.

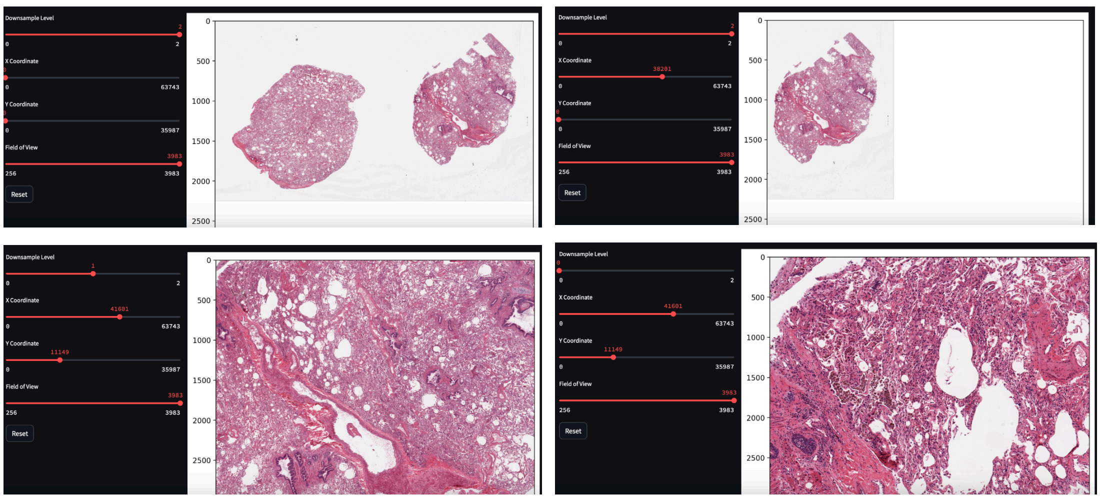


Figure 8.2: Demonstration of WSI navigation functionality on the main page of the app, showcasing various levels of detail.

8.1.2 Statistical Information

A dedicated page within the app allows users to access descriptive statistical information about the tissues from the GTEx database (Figure C.2).

Figure 8.3: This page displays statistics associated with various tissues from GTEx. Users can select a specific tissue type (1), upon which statistical data such as the most frequent words in diagnoses and additional metrics are displayed on the screen (2).

8.1.3 Report Generation

Following the extraction of patches from WSIs, users can navigate to the report generation section of the app. Initially, users are prompted to initiate the feature extraction process, activating the HIPT architecture to derive embeddings critical for feeding the AP, SC, and DR modules of our system.

Upon completion of the HIPT feature extraction, a button to generate reports becomes active. Figure 8.4 illustrates two examples of reports generated by the system for different WSIs. The responsible agent constructs these reports by integrating information from academic literature stored in a vectorstore, completing various required sections of the report.

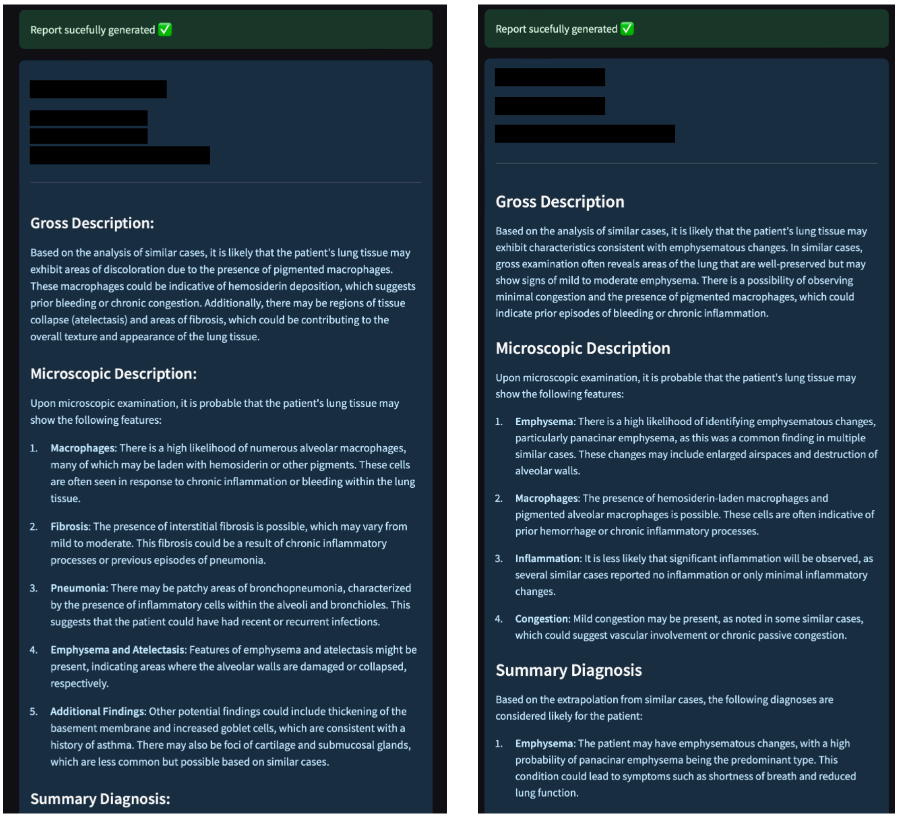


Figure 8.4: Examples of reports generated within the APP for two distinct WSIs. To maintain confidentiality, patient demographic data are obscured.

8.1.4 Specialist ChatBot

The app features a dedicated page for interacting with the Q&A Agent developed to enrich the information in the reports. This agent uses a modular RAG strategy to draw from a local vector database that contains articles and educational materials on lung pathologies, which have been compiled from various online sources.

Figure 8.5 presents three different chat interactions. The first poses a general question about the most and least common pathologies in lung tissue. Subsequent chats involve more specialized inquiries, such as details about Granulomatous Inflammation and features of Honeycombing in idiopathic pulmonary fibrosis. Additionally, the references utilized by the agent are displayed within the chat, enabling users to verify the sources of the information, thus enhancing transparency and control over the content provided by the bot.

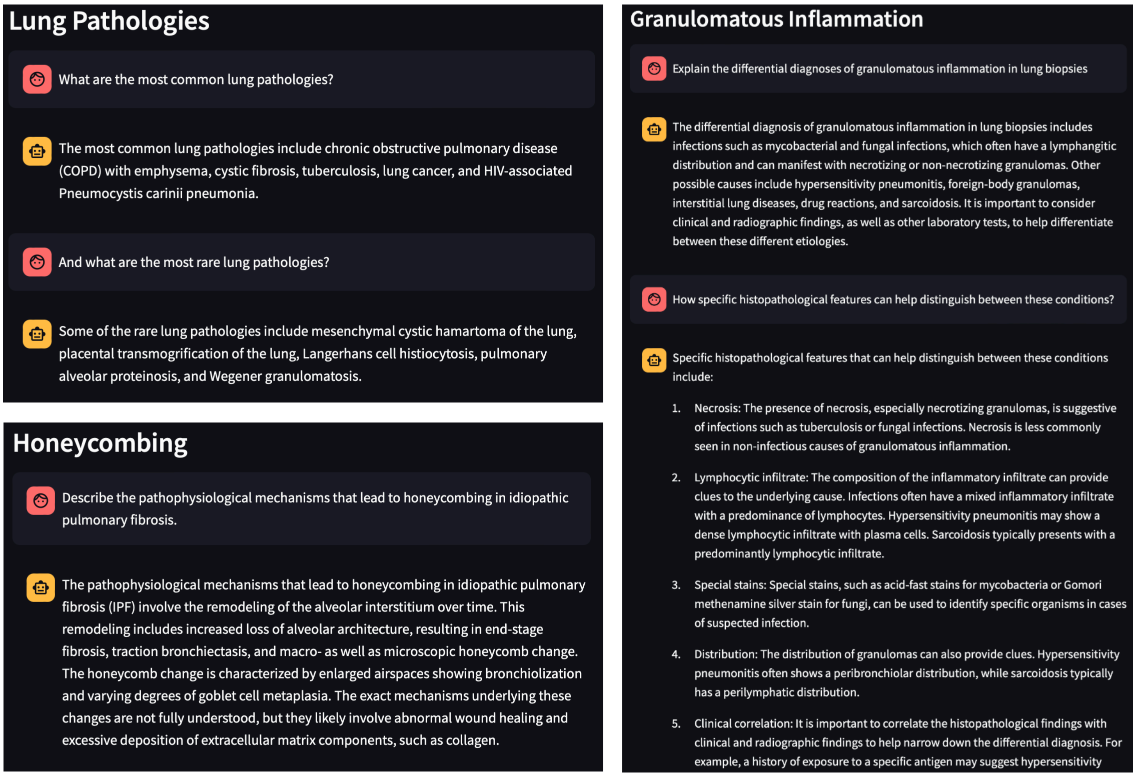


Figure 8.5: Examples of interactions with the Q&A Agent showcasing both simple and complex questions related to pulmonary histopathology.

8.2 Conclusion

The development of this application was intended to demonstrate the potential for making our solution more accessible to individuals without specialist computing knowledge. Traditionally, end users would need familiarity with technologies such as *GitHub*, *Python*, the *LangChain* framework used to build the agents, and *PyTorch* to implement our neural network models in the AP, SC, and DR modules. This knowledge base would be necessary just to access the final report or to engage with a specialized chatbot for addressing specific inquiries. By integrating our solution into an app, we have significantly lowered these technical barriers, providing a practical example of how our system can be made user-friendly for the intended audience. We believe that this effort to make our research accessible through the app has been a worthwhile endeavor.

Chapter 9

Discussion

This chapter is dedicated to exploring the implications of the results previously presented. We have organized this discussion into four main sections: three of them corresponding to the three stages of the solution we implemented (INF, AGG, and GEN), each delineating the strengths and limitations, and a final section that contextualizes these discussions in relation to our initial objectives.

9.1 Inference Stage

9.1.1 Age Prediction

The prediction of age is crucial in assessing the quality of life and the duration of healthy living in humans. Identifying the age of individuals from biological features facilitates an understanding of the mechanisms of aging and aids in planning personalized therapies to enhance quality of life during the aging process. In our previous work on predicting age from histology images [Moraes et al., 2023], we pioneered the integration of these images with other omics data to achieve more accurate results. In the current study, we have expanded upon these results by incorporating a larger sample size, addressing the issue of age imbalance, and obtaining more precise values, as evidenced in Figure 7.4 with a MAE of 5.88 across the 72 subjects in the test set.

Despite these improvements, the problem of age imbalance, although reduced by the application of the LDS technique, persists and continues to hamper optimal performance. This issue is illustrated in Figure 7.5, where the MAE increases with fewer samples in certain age bins. Consequently, there remains scope for exploring techniques that could further alleviate this issue to develop more robust models in the future.

Additionally, the enhanced performance of the current model is attributed not only to our strategic handling of the imbalance but also to the application of HIPT to extract features from the images utilized. At present, we cannot definitively quantify the contribution of each factor to the model's performance, thus presenting an opportunity for detailed analysis in future research.

In line with the objectives of our work, this module has successfully inferred demographic data from the subjects from whom our WSIs were derived, thereby enriching the content of the report and the analyses performed in the GEN stage.

9.1.2 Smoker Classification

Given the association of smoking with a myriad of comorbidities, increased mortality risk, and accelerated aging, it is imperative to study the effects of cigarette smoke on human tissues from both molecular and histological perspectives. This understanding is crucial for deciphering the underlying mechanisms responsible for these adverse outcomes. In collaboration with others, our prior research [Ramirez et al., 2024] has primarily focused on these impacts at the histological level, utilizing deep learning techniques to ascertain the smoking status of patients from their WSIs.

In our current study, we specifically focused on lung tissue and continue to employ the predictive strategy developed in the previous research, aiming to refine a binary classifier capable of discerning between smokers and non-smokers. We have obtained encouraging results, as illustrated by the confusion matrix in Figure 7.9, where we optimized the main diagonal, achieving a good balance between Sensitivity (80%) and Specificity (86%) (Table 7.10). However, these results must be interpreted with caution, as the optimal model from the optimization phase predominantly enhanced performance for folds 1 and 3 (Figure 7.8). This observation underscores the necessity for further investigation to ensure robust generalization across more diverse test sets.

An important limitation of this module is its binary classification system, predicting solely for Smoker and Non-Smoker categories while the original dataset includes significant numbers of Ex-Smokers and individuals with Unknown smoking status. The decision to focus on these two categories was driven by the relative simplicity of binary classifiers over multi-class systems. Although our primary aim was not to develop an exhaustive smoker status classifier but to demonstrate the utility of such a module in enhancing histological reports for lung tissue, we opted for a binary approach. Consequently, Ex-Smokers and those with Unknown smoker status in the test set were classified either as smokers or non-smokers based on their image feature similarities to the learned characteristics of smokers and non-smokers. Future work could focus on expanding this binary classifier into a multi-class task. This expansion would help analyze the distinctions and similarities among different smoker categories and their effects on the final report.

This module, similar to the Age Prediction module, has effectively shown its capability in predicting relevant demographic data for lung histopathology reports, aligning with our initial research objectives.

9.1.3 Diagnosis Retriever

As discussed in the review chapter (section 3.3), CRM systems can greatly benefit clinical settings due to their ability to identify similar data across diverse datasets, which facilitates case classification and accelerates diagnostic processes. In our project, the challenge was significant as we not only trained our system from scratch in a computationally restricted environment using modern Transformer architectures for both text and image data, but also dedicated efforts to integrating WSIs directly into our system. To our knowledge, no other study has developed a CMR model that directly does that. The closest study in this area, Lu et al. [2023], focused on retrieving information from WSIs by processing tiles, rather than integrating the entirety of whole slide data into their CMR system.

We focused on adapting the HIPT architecture within a CLIP model, replacing CLIP's original ViT with HIPT and keeping its weights frozen during training. Given the complete replacement of the ViT, utilizing the original language model weights from CLIP seemed unsuitable, as these were not tailored for histology images nor aligned with HIPT's embeddings. Consequently, we chose to train the language model from scratch, aligning it with the image data. This approach increased the complexity of our task and potentially affected the final model's performance, considering our training dataset (635 pairs image-text) was significantly smaller than the dataset used in the original CLIP study (400 million pairs of image-text).

It is important to note that, according to its authors [Radford et al., 2021], the contrastive loss in the CLIP model generally benefits from larger batch sizes, which enhance learning and predictive robustness. Originally, CLIP was trained using a batch size of

32,000 pairs $\{x_j^I, x_j^C\}$. In contrast, the sample size used to train our DR model was significantly smaller, involving only 727 subjects for both training and validation. This limitation may adversely affect the performance of our model.

In addition to the technical challenges mentioned earlier regarding the implementation of this model, accurately measuring its performance also presented significant hurdles. CMR models like CLIP are multifaceted, designed for various tasks including zero-shot classification and information retrieval, and can be evaluated from multiple perspectives. In our project, the primary goal was to identify cases most similar to a given WSI to retrieve corresponding diagnoses or pathologist notes. These notes were then processed in the AGG stage, where they were categorized into pathology classes based on word frequency. Therefore, we assessed the quality of this CMR model both through its retrieval capabilities using standard caption metrics and its classification accuracy using Precision, Recall, and F1-Score.

However, critiques by Harzig et al. [2019] and González-Chávez et al. [2024] suggest that traditional caption metrics based on n-gram comparisons may distort the true performance of a model, especially in medical contexts. Our findings, depicted in Figure 7.10 and Figure 7.14, demonstrate that while these metrics can yield high values, they may still coexist with lower classification performance. This raises concerns in a pathology setting, where the objective is to precisely predict patient pathologies, not merely to match words.

Given these insights, we shifted our focus towards enhancing the classification capabilities of our CRM model for both selection and evaluation. In the results presented in Figure 7.17, we have seen the robust classification for conditions such as Atelectasis, Congestion, Hemorrhage, Macrophages, and No Abnormalities, moderate performance for Edema, Emphysema, and Pigment, and relatively weaker performance for Pneumonia and Fibrosis.

This module is crucial within our system, as the quality of reports directly depends on the DR's ability to accurately retrieve cases similar to a given input WSI. While the AP and SC modules refine the report with more specific demographic data for each subject, the DR module steers the GEN stage based on the identified pathologies and conditions, which are later integrated in the AGG stage. Therefore, addressing any performance shortcomings in classifying certain pathologies is essential to enhance the overall ability of the system to generate precise reports. Future work could focus on the impact of token imbalance on this module, the number of images in our training set, and explore whether changes to internal components of our neural network could yield broader improvements.

In conclusion, while this module requires further enhancement, we have made significant advances in a relatively unexplored area concerning CMR models in histopathology using to WSIs. Our work demonstrates the feasibility of developing such systems with limited computational resources and a small data set, thereby contributing to future works on domain-specific CRM models.

9.2 Aggregation Stage

This stage, although less complex within our system, can still introduce errors that may propagate into the final report. The aggregation stage involves two critical tasks: firstly, generating probabilities for each of the pathologies based on texts retrieved from the Diagnosis Retriever (DR) module; secondly, concatenating these probabilities with the patient's demographic data to form a comprehensive report.

The process of generating probabilities can inadvertently lead to underestimating or overestimating certain pathologies. Such inaccuracies can propagate errors, magnifying the impact as they transition into the composition of the multilabel vector. This becomes particularly problematic given the existence of false negatives, which, while low for pathologies such as Hemorrhage (17%) and Macrophages (19%) (refer to Figure 7.17), may accumulate when several false positives are also considered. This could result in reports suggesting more pathologies than are actually present, necessitating careful attention to improve accuracy in future work, especially concerning the handling of multi-label pathology representations.

The second function, involving the concatenation of pathological and demographic information into a single report using a predefined template, is less likely to introduce additional errors not already present from previous stages. We have adopted a conservative approach by manually performing this concatenation without relying on LLMs. This method minimizes the risk of introducing new errors but also limits our ability to incorporate descriptive adjectives like 'severe', 'moderate', 'much', or 'little', which could enhance our understanding of the clinical context. Therefore, exploring new methods that could integrate these nuances without significant error introduction remains an area for future development.

9.3 Generation Stage

The GEN stage faces two primary limitations. Firstly, its performance is highly dependent on the outputs from preceding stages. Errors inherent in earlier modules, such as AP, SC, DR, and the captions generated in the AGG stage, can cumulatively affect the GEN stage, potentially leading to compounded inaccuracies in the final reports. This interdependence highlights the critical need for accuracy in each preceding step to ensure the quality of the overall system.

The second major limitation concerns the challenge of quantifying the quality of the final reports in a manner that is meaningful to pathologists. Given the complex and nuanced content of these reports, which go beyond simple captions to include detailed medical information, assessing their quality without direct evaluation by specialists is challenging. Ideally, pathologists and pathology students, the intended end-users, would conduct thorough reviews of each report in conjunction with the corresponding WSIs, evaluating them for both deficiencies and merits.

In the absence of direct collaboration with pathology experts, we employed a new instance of a LLM as a proxy for performance evaluation. Although not ideal, this approach allowed us to gather descriptive statistics to identify areas for improvement within our system. For instance, a significant percentage (43%) of the reports received a score of 2, as shown in Figure 7.21, indicating that while most reports correctly identify some pathologies, they also inaccurately add conditions not present in the original diagnoses. This highlights a high false-negative rate, prompting further analysis to reduce these errors. Additionally, the second-highest score (24%), score 3, reflects frequent omissions within the reports. Although the false positive rate is lower, its impact is still significant and necessitates adjustments in both rates to optimize the confusion matrices shown in Figure 7.17.

Moreover, in the absence of specialists to directly assess the informational value of our reports, we used the same scoring system to evaluate the enrichment process. The data in Figure 7.23 suggests that the majority of reports are adequately enriched with relevant content, indicating that the integration of Questioner and Q&A agents in the GEN stage is effectively meeting its objectives. Its success suggests that there is room to work on its specialization and independent adaptation as a consultation tool for pathologists. This opens up possibilities for developing the system further, aiming to meet the specific needs of clinical practice and educational environments.

In conclusion, while the current results do not yet reach the standards necessary for clinical or educational implementation, the system demonstrates functionality and potential for improvement at multiple levels. It shows promise as a valuable tool that could eventually enhance the daily practice of pathologists.

9.4 Perspective on Our Objectives

Our primary goal was to develop an application that could bridge the domains of CPATH and Clinical NLP. While we did not achieve results sufficient for deployment in a production environment or integration into routine histopathological analysis and educational settings, we believe our efforts successfully demonstrated a pathway for multimodal integration between these fields.

We established three methodological guidelines for constructing our system: first, to develop and reproduce our work in environments with limited computational resources; second, to ensure that our solution could be easily maintained and improved; and third, to make our research accessible to end users without extensive computational expertise.

We believe we have met all these guidelines. Firstly, we developed the entire system on a personal computer with limited resources. Secondly, the architecture of our system, consisting of modular components, adheres to our second guideline. This modular design allows for individual components to be upgraded, new ones to be added, or existing ones to be removed without disrupting the entire system. Lastly, we achieved our third guideline by embedding our system in an application with a user-friendly interface, thereby facilitating easy access, usage, and replication by end users.

Furthermore, it is worth noting that, to the best of our knowledge, no prior study has attempted to evaluate the performance of their models in the manner we did, with a particular focus on the value for the end user. Generally, research in the field of CPATH, especially recent studies involving FMs, tends to concentrate more on the development and refinement of neural network architectures and on surpassing established benchmarks. Our approach differed significantly. Throughout this work, we consistently considered the end user, for whom this effort is intended, as a critical reference point. The construction and evaluation of our solution thus reflected our dedication to the interpretability and utility of our work, aiming to ensure that the system is not only effective but also accessible and beneficial to its intended users.

In conclusion, we believe that we have successfully met our established objectives, making significant theoretical contributions in terms of system development and deep learning. Additionally, we have provided practical tools that have the potential to enhance the field of histopathology, demonstrating the benefits of our research.

Chapter 10

Conclusion and Future Work

In this project, we applied recent DL techniques to the fields of CPATH and Clinical NLP. Our goal was to develop a system that supports pathologists by integrating image and text modalities in a novel way. We achieved this by creating a system that on one side utilizes CPATH algorithms to process WSIs, and on the other side employs Clinical NLP to generate histopathology reports from these images.

During INF stage, our components exhibited varying performance levels. The AP module, notably robust, achieved a MAE of 5.88 on the test set, demonstrating better accuracy for ages above 45 years. Our DR module, when evaluated as a classifier, attained an F1-score of 0.56, and as a caption generator, it recorded BLEU-1 of 0.71, CIDEr of 0.48, and METEOR of 0.30. The SC module demonstrated a good balance with a sensitivity of 0.80 and specificity of 0.86. These individual modules, especially the DR, are crucial for generating accurate reports, as they primarily inform the content of these reports. Although perfecting each module was not our central focus since demonstrating the system's feasibility and its multimodal integration was the priority, future enhancements to the DR module will be necessary to improve report quality.

In GEN stage, evaluating the quality of our reports without a aid of skilled pathologists was challenging. To address this, we employed an additional instance of a LLM, this time as an evaluator, providing a quantitative measure of report performance. Our evaluations indicated varied accuracy levels, with 12% of the reports being completely accurate, 23% committing minor errors, 43% committing major errors, and the remaining ones showing complete divergence. Furthermore, the informational aspect of our reports demonstrated a high score in category 4 of 96%, indicating that the enrichment phase of our system was successfully conducted.

Working in a computationally constrained environment necessitated strategic choices to optimize resources. The increasing complexity of neural networks, like the Transformers used in the DR module, requires substantial computational resources. Our project demonstrates that careful planning and strategic decisions can yield valuable and domain-specific results without significant investments.

Additionally, embedding the report generation system within an application featuring a high-level user interface has made our system more accessible to end users. Unlike many CPATH systems that demand computational expertise from users, our approach reduces barriers by making it easier for non-specialists to integrate CPATH systems into the daily practice of pathologists and educational settings

Looking ahead, the insights gained from each component of this system provide a rich foundation for further development. The DR module, in particular, holds significant potential for both scientific and clinical advancements due to its modern architecture and the practical value of CMR systems in improving diagnostic accuracy. Future work could also focus on enhancing the GEN stage components, particularly the modular RAG used in the Q&A agent. This could lead to the development of decision-support tools and educational aids for pathology students.

In conclusion, we believe that this work has successfully met its primary objectives and holds considerable promise to enhance the field of histopathology. Through our experimentation in the CPATH and Clinical NLP domains, we have explored many theoretical aspects of DL and gained valuable experience that can be further used to enhance our solution. By engaging in a multimodal approach to bridge these areas, we have proposed plausible avenues for future initiatives aiming to integrate these domains into systems that can be used in clinical and educational settings. Although not yet fully ready for deployment in such environments, we remain optimistic about the potential that our initial findings have demonstrated and look forward to improve our solution.

Appendix A

Technical Background

A.1 Hierarchical Image Pyramid Transformer

The Hierarchical Image Pyramid Transformer (HIPT) developed by Chen et al. [2022] is a ViT architecture designed to tackle the complexities of gigapixel WSIs in CPATH. WSIs are exceptionally high-resolution images, often exceeding $100,000 \times 100,000$ pixels, which capture detailed tissue microenvironments at multiple scales. Traditional methods, such as multiple instance learning (MIL), face significant challenges in handling the vast size and hierarchical nature of WSIs, often failing to capture the full spectrum of morphological features. HIPT addresses these challenges by exploiting the inherent hierarchical structure of WSIs and employing a sophisticated bottom-up aggregation strategy to derive comprehensive slide-level representations.

HIPT is designed to process visual tokens at various scales, from small patches that capture individual cells to larger patches that encompass broader tissue microenvironments. The architecture comprises three primary stages: cell-level aggregation using $ViT_{256} - 16$, patch-level aggregation using $ViT_{4096} - 256$, and region-level aggregation using $ViT_{WSI} - 4096$. Each stage utilizes Transformer self-attention mechanisms to model the critical dependencies between visual concepts at different resolutions. This hierarchical approach enables HIPT to capture both fine-grained details, such as cell identity and shape, and coarse-grained features, such as tumor-immune localization, which are essential for accurate cancer subtyping and survival prediction. The model is pretrained using a two-stage self-supervised learning process, leveraging the DINO framework to pretrain each aggregation layer. This pretraining strategy allows HIPT to generalize effectively across different stages and learn robust representations for high-resolution images.

By generating more comprehensive and hierarchical representations of WSIs, HIPT demonstrates improved performance over existing state-of-the-art methods in various diagnostic and prognostic tasks. Its ability to model long-range dependencies between phenotypes proves beneficial in context-aware tasks, such as survival prediction. Furthermore, HIPT can be employed as a feature extractor backbone for a vast range of downstream tasks, including semantic segmentation of histopathology tissue, identifying both fine-grained and coarse-grained visual concepts similar to self-supervised ViTs on natural images.

A.2 Label Distribution Smoothing

Label Distribution Smoothing (LDS) is a technique developed by Yang et al. [2021] to address the challenges posed by imbalanced data in regression tasks, particularly when dealing with continuous target variables. Traditional methods for handling data imbalance, such as re-sampling and re-weighting, are primarily tailored for classification problems with categorical targets. However, these approaches often fail when applied to regression tasks, where the target values are continuous and may exhibit complex, skewed distributions. LDS aims to bridge this gap by leveraging the inherent similarities between nearby target values, thereby providing a more nuanced understanding of data imbalance in the continuous domain.

The primary problem that LDS seeks to solve is the inadequacy of traditional imbalance handling techniques in capturing the true distribution of continuous target variables. In regression tasks, the empirical label density does not accurately reflect the imbalance as perceived by the model. This discrepancy arises from the dependence between data samples at nearby labels, which is not accounted for in conventional methods. For instance, in age estimation tasks, the similarity between individuals of close ages is not utilized effectively, leading to suboptimal performance. LDS addresses this issue by employing kernel density estimation to smooth the empirical label density distribution, thereby capturing the effective imbalance that affects regression problems.

In practical applications, such as age estimation, LDS can be instrumental in improving model performance. By convolving a symmetric kernel with the empirical density distribution (gaussian, laplacian, etc.), LDS generates a smoothed version that accounts for the overlap in information of data samples with nearby ages. This smoothed distribution correlates more strongly with the error distribution, indicating that LDS captures the real imbalance that impacts the model's predictions. Once the effective label density is obtained, techniques for addressing class imbalance, such as cost-sensitive re-weighting, can be directly adapted to the regression context. For example, the loss function can be reweighted by multiplying it by the inverse of the LDS-estimated label density for each target age. This approach not only mitigates the effects of data imbalance but also enhances the model's ability to generalize to the entire range of continuous target values, providing a more comprehensive and unbiased evaluation for deep imbalanced regression tasks.

A.3 Caption Generation Metrics

A.3.1 Caption Metrics

- **BLEU** [Papineni et al., 2002]: Measures the quality of machine-translated text by comparing it to reference translations, focusing on the precision of n-grams. It is calculated using the formula BLEU $= \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$, where p_n is the precision for n-grams and BP is a brevity penalty.

- **CIDEr** [Vedantam et al., 2015]: Evaluates the similarity of generated captions in image description tasks using consensus among references, formulated as CIDEr $= \sum_{n=1}^{N} w_n \left(\frac{g^n \cdot s^n}{\|g^n\| \|s^n\|}\right)$, with g^n and s^n being the TF-IDF vectors for n-grams.

- **METEOR** [Denkowski and Lavie, 2014]: Aligns generated text to reference translations and computes a score based on a harmonic mean of precision and recall, using METEOR $= F_{\text{mean}} = \frac{10 \cdot P \cdot R}{R + 9 \cdot P}$, where P and R are precision and recall, respectively.

- **ROUGE** [Lin, 2004]: Measures the overlap of n-grams between the generated and reference summaries, with the formula ROUGE-N $= \frac{\sum_{s \in \text{Ref Summaries}} \sum_{\text{gram}_n \in s} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{s \in \text{Ref Summaries}} \sum_{\text{gram}_n \in s} \text{Count}(\text{gram}_n)}$.

- **SPICE** [Anderson et al., 2016]: Assesses the semantic quality of captions by comparing sets of semantic propositions, calculated typically through set-based F-scores over semantic tuples.

- **SPIDER** [Liu et al., 2017]: A metric combining SPICE and CIDEr to evaluate both semantic and consensus aspects of captions, given by SPIDER $= \alpha \cdot \text{SPICE} + (1 - \alpha) \cdot \text{CIDEr}$.

A.3.2 Distance Metrics

- **Hamming**: Counts the number of positions at which two strings of equal length differ, calculated as $D_H(x, y) = \sum_{i=1}^{n}[x_i \neq y_i]$.
- **Euclidean**: Measures the "straight line" distance between two points in Euclidean space, using the formula $D_E(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$.
- **Manhattan**: Computes the sum of the absolute differences of their Cartesian coordinates, expressed as $D_M(x, y) = \sum_{i=1}^{n} |x_i - y_i|$.
- **Cosine**: Determines the cosine of the angle between two vectors, indicating similarity, defined by Cosine Similarity $= \frac{x \cdot y}{\|x\|\|y\|}$.

A.4 CLIP Model and Contrastive Loss

In the CLIP (Contrastive Language–Image Pre-training) model, a dual-encoder architecture is employed where separate networks are used to encode images and texts into a common embedding space. Consider a mini-batch of N image-caption pairs $\{(x_i^I, x_i^C)\}_{i=1}^{N}$. The objective is to learn embeddings such that corresponding image and caption pairs have high cosine similarity compared to non-corresponding pairs.

Each image x_i^I and caption x_i^C are transformed by their respective encoders into embedding vectors $z_i^I = f_I(x_i^I)$ and $x_i^C = f_C(x_i^C)$, where $f_I(\cdot)$ and $f_C(\cdot)$ denote the image and text encoding functions respectively. Next, these embeddings are normalized in order to simplify the computation of cosine similarity between any given image and text embedding to a dot product, $x_i^I \cdot x_j^C$.

The cosine similarity scores are then scaled by a learned parameter τ (temperature), adjusting the distribution's sharpness and forming a matrix S of logits where each element $S_{ij} = \tau(x_i^I \cdot x_j^C)$ represents the logit for the i-th image with the j-th text.

The label vector $\mathbf{y} = [0, 1, 2, \ldots, N-1]$ indicates that the correct text for each image in the batch corresponds to the diagonal elements of the matrix S.

$$Loss_{CrossEntropy} = \frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp\left(S_{i,y_i}\right)}{\sum_{j=1}^{N}\exp\left(S_{i,j}\right)}\right) \tag{A.1}$$

Using this general definition of cross-entropy loss, we compute the loss for image-to-text alignment from matrix S and for text-to-image alignment from the transposed matrix S^T:

$$Loss_{CrossEntropy}|_{I \to C} = Loss_{CrossEntropy}(S)$$

$$Loss_{CrossEntropy}|_{C \to I} = Loss_{CrossEntropy}(S^T)$$

The final contrastive loss is then the average of these two losses:

$$Loss_{Contrastive} = \frac{1}{2}(Loss_{CrossEntropy}|_{I \to C} + Loss_{CrossEntropy}|_{C \to I}) \tag{A.2}$$

This symmetric approach ensures that the model learns to effectively map both images to texts and texts to images, maximizing the cosine similarity for correct pairs while minimizing it for incorrect pairs. This alignment optimizes the model's ability to perform tasks that require understanding and correlating content across visual and textual modalities.

Appendix B

Prompts

Here we will be listing all the prompts used in our system, along with details about their operation, the LLM used, temperature, maximum number of tokens, and input variables. The codes are adaptations of YAML files and can be readily utilized. Simply copy and paste into a YAML file and read directly in Python via a dictionary.

B.1 Sentence Cleaner

```
description:
  "Clean the original pathologist note. It receives 1 parameter:
    - sentence: The original pathologist note."
model: "gpt-4o"
temperature: 0.1
max_tokens: 200
template:
  "You are a pathologist annotator. Improve a pathologist's note by following these rules:
    1. Remove numbers and measurements (e.g., '8x7mm', '10%', '1 piece', '2 mm infarct', 'two pieces').
    2. Eliminate unnecessary punctuation and spaces.
    3. Exclude non-pathology-related comments (physical characteristics, location, and composition are not pathologies).
    4. Remove text inside parentheses unless relevant; if relevant, incorporate it.
    5. Improve clarity while retaining core medical information.
    6. If no pathology is observed, always state that there are no evident pathologies.

```

```
  Return a JSON with 'enhanced' (your enhanced version) and 'pathology' (boolean indicating if there is a pathology).

  sentence: {sentence}"
input_variables: - sentence
```

Listing B.1: Sentence Cleaner Prompt and Configurations

B.2 Questioner

```
model: "gpt-4o"
temperature: 0.1
max_tokens: 200
template: |
  "As a skilled pathologist, your task is to generate a series of detailed questions from a given diagnosis following these guidelines:
    1. Questions should focusing on the underlying pathology and diagnostic implications from the given diagnosis.
    2. Each question should delve into understanding the conditions mentioned in the diagnosis without re-asking about explicitly stated details.
    3. The questions should progressively explore the pathology, from basic definitions to more complex implications.
    4. Questions should explore the possibility of coexistence between conditions.
    5. Questions should explore gross and microscopic features of these conditions.
    6. Do not generate more than 10 questions, and make them as simple as possible.

  Finally, return these sub-questions a single list in a JSON format with the sub-questions listed under the key 'questions'.\n

  Diagnosis: \n\n{diagnosis}"
input_variables:
  - diagnosis
```

Listing B.2: Question Generation Prompt used to generate up to 10 questions given an input diagnosis.

B.3 Q&A RAG

This agent is more complex and therefore, its implementation was not as straightforward as with the others. It employs RAG and also operates with memory, accessing chat history to refine its responses. The prompt for this agent is divided into two parts: one for the memory, which enables it to incorporate relevant historical data, and another for the RAG, enhancing the accuracy and relevance of its outputs.

- Memory Prompt:

```
'''
Given a chat history and the latest user question which might reference context in the chat history, formulate a standalone question which can be understood without the chat history. Do NOT answer the question, just reformulate it if needed and otherwise return it as is.
'''

```

Listing B.3: Prompt used to evaluate the report regarding its classificatory aspec.

- RAG Prompt:

```
'''
You are a pathologist assistant.

Analyze the retrieved documents and use your medical expertise to answer the pathology-related question.

Focus on relevant medical findings, diagnostics, and pathological data in your response.

If the answer is not clearly supported by the documents provided, indicate that the information is inconclusive or unknown. Your responses should be precise, concise, and limited to three sentences.

If none of the documents provided are relevant, you may draw on your general medical knowledge to provide an informed guess, but clearly state that the response is based on general knowledge and not directly supported by the provided documents.

{context}
'''
```

Listing B.4: Prompt used to evaluate the report regarding its classificatory aspec.

B.4 Evaluator

B.4.1 Classification Aspect

```
model: "gpt-4o"
temperature: 0.001
max_tokens: 60
template: |
  '''
  You are an evaluator tasked with scoring the accuracy of a generated histopathology report by comparing its relevance and correctness against the specific pathologies provided in the original pathologist note.
  For this evaluation, forget demographic data and only focus on the pathologies presented.

  Scoring Scale for Pathologies:
    In comparison to the Correct Pathologist Note the report...
    - 4: Accurately identifies all main pathologies, without any errors.
    - 3: Correctly identifies the majority of the main pathologies, with minor errors or omissions of less critical pathologies. Here it is allowed to miss up to 1 pathology.
    - 2: Correctly identifies only some of the main pathologies, with significant omissions or incorrect identifications of several key pathologies.
    - 1: Completely unrelated; fails to mention any of the main pathologies or mentions entirely incorrect pathologies.

  Evaluate the generated note with a strict focus on how comprehensively and accurately it reflects the pathologies listed in the correct note. Assign a score based on this assessment and provide a brief (20 words) explanation of your reasoning.
  Return your evaluation in JSON format with <score> for the numerical rating and <explanation> for your justification. The explanation should highlight the clinical significance of capturing all primary pathologies accurately and the potential impact of any misses.

  Correct Pathologist Note: {correct_note}
```

```
  Generated Report: {generated_report}
  '''
input_variables:
  - correct_note
  - generated_report
```

Listing B.5: Prompt used to evaluate the report regarding its classificatory aspec.

B.4.2 Informative Aspect

```
model: "gpt-4o"
temperature: 0.001
max_tokens: 60
template: |
  '''
  You are an evaluator tasked with scoring the accuracy of a generated histopathology report by comparing the relevant and correctness of the information provided regarding the pathologies and conditions analysed.
  You will be provided with a sentence containing the pathologies/conditions and a full generated report. Forget about the age and smoker status of the patient, focus on the literature information provided.

  Scoring Scale:
    In comparison to the given pathologies/conditions, the report...
    Scoring Scale:
    - 4: Provides highly accurate and relevant details for all main pathologies/conditions mentioned, with no factual errors.
    - 3: Contains accurate and relevant information for most pathologies/conditions, with minor factual inaccuracies or slightly irrelevant details.
    - 2: Information about pathologies/conditions is partially accurate; includes notable factual errors or irrelevant details.
    - 1: Information is mostly inaccurate or irrelevant to the pathologies/conditions listed, showing a clear lack of understanding or incorrect facts.

  Evaluate the generated note with a strict focus on the truthfulness and relevance of the information it presents. Assign a score based on this assessment and provide a brief (20 words) explanation of your reasoning. Return your evaluation in JSON format with <score> for the numerical rating and <explanation> for your justification.

```

```
  Pathologies/Conditions: {correct_note}
  Generated Report: {generated_report}
  '''
input_variables:
  - correct_note
  - generated_report
```

Listing B.6: Prompt used to evaluate the report regarding its informative aspect.

B.5 Report Generator

```
model: "gpt-4o"
temperature: 0.1
max_tokens: 500
template: |
  """
  You are a clever Pathologist.
  You will be provided with two data sources: Literature References (LR) and
 Pathological Findings (PF).
  Your task is to generate a complete pathologist report about the PF.

  RULES:
    - Use the LR only to guide your analysis over PF, not to generate
 fictitious unseen information about PF.
    - If you face unlikely findings in the PF, do not try to make sense of
 them with a persuasive narrative. Instead explain the leak of probability
 in the Comments section using LR.
    - If you can't get useful info from LR, clearly state that it doesn't
 provide useful insights.
    - The generated report must contain the following sections and only them,
  nothing more:
      1. Gross Description: Visible traits of the specimen described as they
 appear in the PF, nothing more.
      2. Microscopic Description: If available in PF, document the cellular
 findings and tissue observations and improve the text.
      3. Comments: Additional insights or implications always based on the LR
 . If none, return empty.
      4. Summary: Brief recap of main findings and diagnosis.
    - Always adopt a probabilistic narrative to discuss the inferences that
 you make. E.g, use words like 'propprobably', 'likely', and 'might', for
 any kind of observation that is not explicitly written in PF.
```

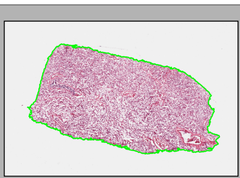
```
    - Finally, return yout response in json format under the keys 'gross', '
  micro', 'comments', and 'summary'.

  LR: {context}
  PF: {input}
  """
input_variables:
  - context
  - input
```

Listing B.7: Prompt used to generate the final reports.

Appendix C

Generated Reports

Here we present more examples of histopathology reports generated by our system.



ID: GTEX-XXXXX-XXXX
Age: 43 (36)
Smoker: True (Unknown)

Gross Analysis

The lung tissue specimens present no visible abnormalities.

Microscopic Analysis

Microscopic examination of the lung tissue reveals no abnormalities.

Comments

Although no abnormalities were observed in the lung tissue specimens, literature indicates that smokers often present with pathological changes such as emphysema, interstitial fibrosis, and increased numbers of T lymphocytes. Additionally, early lung pathology in smokers may include increased numbers of alveolar macrophages, eosinophils, and Langerhans cells, even in the absence of overt disease. It is important to consider that normal lung tissue histology can coexist with early stages of lung disease, particularly in smokers, according to studies.

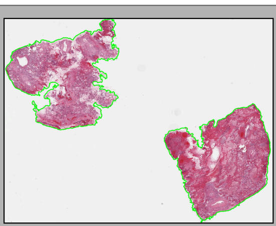
Summary

The lung tissue specimens from a 43-year-old patient show no visible or microscopic abnormalities.

True Diagnosis

No evident pathologies

Figure C.1: Example of a generated report. The ground truth references are highlighted in green. This example showcases a report accurately generated based on a correct prediction indicating the absence of pathologies. We classified the subject as a smoker when their original status was 'unknown', which led to comments about smoking patients being generated. This is considered a minor error that, as we see, did not impact the final diagnosis. The GPT-Score of this report is 4.



ID: GTEX-XXXXX-XXXX
Age: 51 (46)
Smoker: False (Non Smoker)

Gross Analysis

The lung tissue specimens present visible signs of congestion, edema, fibrosis, hemorrhage, and pneumonia. Literature indicates that congestion in lung tissue often appears as a reddish, heavy, and wet appearance due to increased blood content. Edema typically manifests as a frothy, fluid-filled appearance in the alveoli. Fibrosis usually presents as firm, thickened areas of tissue. Hemorrhage can be identified by the presence of blood within the lung parenchyma. Pneumonia often shows areas of consolidation and inflammation within the lung tissue.

Microscopic Analysis

Microscopic examination of the lung tissue reveals congestion, characterized by engorged blood vessels. Edema is evident by the presence of fluid within the alveolar spaces. Fibrosis is observed as thickened alveolar septa with increased collagen deposition. Hemorrhage is noted by the presence of red blood cells within the alveolar spaces. Pneumonia is identified by the presence of inflammatory cells, including neutrophils, within the alveoli and interstitial spaces.

Comments

According to literature, the congestion observed in the lung tissue may be due to inflammation and increased vascular permeability typically associated with pneumonia. Pulmonary edema, as seen in this case, is commonly associated with fluid accumulation in the alveoli due to increased vascular permeability and impaired fluid clearance mechanisms in pneumonia. The fibrosis observed could suggest a chronic inflammatory response and tissue repair process, leading to excessive deposition of collagen and other extracellular matrix components. This scarring can impair lung function by reducing lung compliance and gas exchange efficiency. Hemorrhage in the lung tissue may be related to conditions such as thrombocytopenia, sepsis, or severe congestion, as per literature. The coexistence of pneumonia and fibrosis in this patient may indicate an acute exacerbation of the fibrotic process, potentially worsening respiratory function.

Summary

The lung tissue specimens from a 51-year-old non-smoker present with congestion, edema, fibrosis, hemorrhage, and pneumonia. Literature describes that congestion and edema are typically associated with increased vascular permeability and fluid accumulation in pneumonia. Fibrosis is commonly seen as a result of chronic inflammation and tissue repair, leading to scarring and impaired lung function. Hemorrhage may be related to various conditions including sepsis and severe congestion. The presence of pneumonia in fibrotic lungs can lead to acute exacerbation of fibrosis, worsening respiratory function, as suggested by studies.

True Diagnosis

Extensive pneumonia and organizing thrombi. Interstitial fibrosis and hemorrhage. Main findings: fibrosis, hemorrhage, pneumonia.

Figure C.2: Example of a generated report with a GPT-Score of 2. The ground truth references are highlighted in green. This report was generated showing that the original diagnosis mentioned Fibrosis, Hemorrhage, and Pneumonia, three pathologies that were correctly identified in the report. However, we erroneously added two unmentioned pathologies, congestion and edema, which resulted in the GPT-Score dropping to 2.

Bibliography

A. K. Dewi and B. Purwanto, "Introduction of Histopathology," in Molecular Histopathology and Cytopathology. IntechOpen, 2023. [Online]. Available: https://www.intechopen.com/chapters/86530 [Cited on page 1.]

S. W. Jahn, M. Plass, and F. Moinfar, "Digital Pathology: Advantages, Limitations and Emerging Perspectives," Journal of Clinical Medicine, vol. 9, no. 11, p. 3697, Nov. 2020, number: 11 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2077-0383/9/11/3697 [Cited on pages 1, 2, and 29.]

N. Farahani, A. V. Parwani, and L. Pantanowitz, "Whole slide imaging in pathology: advantages, limitations, and emerging perspectives," Pathology and Laboratory Medicine International, vol. 7, pp. 23–33, Jun. 2015, publisher: Dove Press. [Online]. Available: https://www.dovepress.com/whole-slide-imaging-in-pathology-advantages-limitations-and-emerging-p-peer-reviewed-fulltext-article-PLMI [Cited on page 1.]

M. Cooper, Z. Ji, and R. G. Krishnan, "Machine learning in computational histopathology: Challenges and opportunities," Genes, Chromosomes and Cancer, vol. 62, no. 9, pp. 540–556, 2023, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gcc.23177. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/gcc.23177 [Cited on pages 1, 2, 28, and 29.]

J. van der Laak, G. Litjens, and F. Ciompi, "Deep learning in histopathology: the path to the clinic," Nature Medicine, vol. 27, no. 5, pp. 775–784, May 2021, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41591-021-01343-4 [Cited on pages 1 and 28.]

K. Kartasalo, W. Bulten, B. Delahunt, P.-H. C. Chen, H. Pinckaers, H. Olsson, X. Ji, N. Mulliqi, H. Samaratunga, T. Tsuzuki, J. Lindberg, M. Rantalainen, C. Wählby, G. Litjens, P. Ruusuvuori, L. Egevad, and M. Eklund, "Artificial

Intelligence for Diagnosis and Gleason Grading of Prostate Cancer in Biopsies—Current Status and Next Steps," European Urology Focus, vol. 7, no. 4, pp. 687–691, Jul. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405456921001814 [Cited on pages 2 and 28.]

E. Abels, L. Pantanowitz, F. Aeffner, M. D. Zarella, J. van der Laak, M. M. Bui, V. N. Vemuri, A. V. Parwani, J. Gibbs, E. Agosto-Arroyo, A. H. Beck, and C. Kozlowski, "Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association," The Journal of Pathology, vol. 249, no. 3, pp. 286–294, Nov. 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6852275/ [Cited on pages 2, 15, and 29.]

R. J. Chen, C. Chen, Y. Li, T. Y. Chen, A. D. Trister, R. G. Krishnan, and F. Mahmood, "Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning," Jun. 2022, arXiv:2206.02647 [cs]. [Online]. Available: http://arxiv.org/abs/2206.02647 [Cited on pages 2, 9, 20, 36, 49, 74, and 115.]

P. Osorio, G. Jimenez-Perez, J. Montalt-Tordera, J. Hooge, G. Duran-Ballester, S. Singh, M. Radbruch, U. Bach, S. Schroeder, K. Siudak, J. Vienenkoetter, B. Lawrenz, and S. Mohammadi, "Latent Diffusion Models with Image-Derived Annotations for Enhanced AI-Assisted Cancer Diagnosis in Histopathology," Dec. 2023, arXiv:2312.09792 [cs]. [Online]. Available: http://arxiv.org/abs/2312.09792 [Cited on page 2.]

W. O. Ikezogwo, M. S. Seyfioglu, F. Ghezloo, D. S. C. Geva, F. S. Mohammed, P. K. Anand, R. Krishna, and L. Shapiro, "Quilt-1M: One Million Image-Text Pairs for Histopathology," Oct. 2023, arXiv:2306.11207 [cs]. [Online]. Available: http://arxiv.org/abs/2306.11207 [Cited on pages 2, 20, and 21.]

S. Sengupta and D. E. Brown, "Automatic Report Generation for Histopathology images using pre-trained Vision Transformers," Nov. 2023, arXiv:2311.06176 [cs]. [Online]. Available: http://arxiv.org/abs/2311.06176 [Cited on page 2.]

H. Nori, Y. T. Lee, S. Zhang, D. Carignan, R. Edgar, N. Fusi, N. King, J. Larson, Y. Li, W. Liu, R. Luo, S. M. McKinney, R. O. Ness, H. Poon, T. Qin, N. Usuyama, C. White, and E. Horvitz, "Can Generalist Foundation Models Outcompete Special-Purpose

Tuning? Case Study in Medicine," Nov. 2023, arXiv:2311.16452 [cs]. [Online]. Available: http://arxiv.org/abs/2311.16452 [Cited on pages 3 and 24.]

Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-Augmented Generation for Large Language Models: A Survey," Mar. 2024, arXiv:2312.10997 [cs]. [Online]. Available: http://arxiv.org/abs/2312.10997 [Cited on pages 3 and 24.]

O. Unlu, J. Shin, C. J. Mailly, M. F. Oates, M. R. Tucci, M. Varugheese, K. Wagholikar, F. Wang, B. M. Scirica, A. J. Blood, and S. J. Aronson, "Retrieval Augmented Generation Enabled Generative Pre-Trained Transformer 4 (GPT-4) Performance for Clinical Trial Screening," Feb. 2024, iSSN: 2430-2376 Pages: 2024.02.08.24302376. [Online]. Available: https://www.medrxiv.org/content/10.1101/2024.02.08.24302376v1 [Cited on pages 3 and 25.]

A. Moraes, M. Moreno, R. Ribeiro, and P. G. Ferreira, "Predicting Age from Human Lung Tissue Through Multi-modal Data Integration," in Discovery Science, A. Bifet, A. C. Lorena, R. P. Ribeiro, J. Gama, and P. H. Abreu, Eds. Cham: Springer Nature Switzerland, 2023, pp. 644–658. [Cited on pages 7, 8, 10, 53, 77, and 105.]

R. Ribeiro, A. Moraes, M. Moreno, and P. G. Ferreira, "Integration of multi-modal datasets to estimate human aging," Machine Learning, Jul. 2024. [Online]. Available: https://link.springer.com/10.1007/s10994-024-06588-x [Cited on pages 7, 9, and 10.]

J. M. Ramirez, R. Ribeiro, O. Soldatkina, A. Moraes, R. Garcia-Perez, P. G. Ferreira, and M. Mele, "Cigarette smoking drives accelerated aging across human tissues," bioRxiv, pp. 2024–03, 2024, publisher: Cold Spring Harbor Laboratory. [Online]. Available: https://www.biorxiv.org/content/10.1101/2024.03.14.585016.abstract [Cited on pages 7, 11, 53, 82, and 106.]

Y. Yang, K. Zha, Y.-C. Chen, H. Wang, and D. Katabi, "Delving into Deep Imbalanced Regression," May 2021, arXiv:2102.09554 [cs]. [Online]. Available: http://arxiv.org/abs/2102.09554 [Cited on pages 9, 54, and 116.]

B. Ramamurthy, F. D. Coffman, and S. Cohen, "A perspective on digital and computational pathology," Journal of Pathology Informatics, vol. 6, no. 1, p. 29, Jan. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2153353922004849 [Cited on page 15.]

G. Neubig, “Neural Machine Translation and Sequence-to-sequence Models: A Tutorial,” Mar. 2017, arXiv:1703.01619 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1703.01619 [Cited on page 16.]

I. Hartsock and G. Rasool, “Vision-Language Models for Medical Report Generation and Visual Question Answering: A Review,” Apr. 2024, arXiv:2403.02469 [cs]. [Online]. Available: http://arxiv.org/abs/2403.02469 [Cited on page 16.]

J. Zhang, J. Huang, S. Jin, and S. Lu, “Vision-Language Models for Vision Tasks: A Survey,” IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–20, 2024, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10445007 [Cited on page 16.]

S. Sengupta and D. E. Brown, “Automatic Report Generation for Histopathology images using pre-trained Vision Transformers and BERT,” Mar. 2024, arXiv:2312.01435 [cs]. [Online]. Available: http://arxiv.org/abs/2312.01435 [Cited on pages 16, 17, and 47.]

M. Y. Lu, B. Chen, D. F. K. Williamson, R. J. Chen, I. Liang, T. Ding, G. Jaume, I. Odintsov, A. Zhang, L. P. Le, G. Gerber, A. V. Parwani, and F. Mahmood, “Towards a Visual-Language Foundation Model for Computational Pathology,” Jul. 2023, arXiv:2307.12914 [cs]. [Online]. Available: http://arxiv.org/abs/2307.12914 [Cited on pages 16, 18, 20, 21, 49, and 107.]

B. C. Guevara, N. Marini, S. Marchesin, W. Aswolinskiy, R.-J. Schlimbach, D. Podareanu, and F. Ciompi, “Caption generation from histopathology whole-slide images using pre-trained transformers,” Apr. 2023. [Online]. Available: https://openreview.net/forum?id=gpsfGAOUs58 [Cited on pages 16, 18, and 49.]

Q. Zhou, W. Zhong, Y. Guo, M. Xiao, H. Ma, and J. Huang, “PathM3: A Multimodal Multi-Task Multiple Instance Learning Framework for Whole Slide Image Classification and Captioning,” Mar. 2024, arXiv:2403.08967 [cs]. [Online]. Available: http://arxiv.org/abs/2403.08967 [Cited on pages 16 and 49.]

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, “HuggingFace’s Transformers: State-of-the-art Natural Language Processing,” Jul.

2020, arXiv:1910.03771 [cs]. [Online]. Available: http://arxiv.org/abs/1910.03771 [Cited on pages 16 and 59.]

Y. Ding, L. Liu, C. Tian, X. Zhang, and X. Tian, "Balanced image captioning with task-aware decoupled learning and fusion," Neurocomputing, vol. 538, p. 126159, Jun. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092523122300245X [Cited on pages 17 and 49.]

O. González-Chávez, G. Ruiz, D. Moctezuma, and T. Ramirez-delReal, "Are metrics measuring what they should? An evaluation of Image Captioning task metrics," Signal Processing: Image Communication, vol. 120, p. 117071, Jan. 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0923596523001534 [Cited on pages 17 and 108.]

P. Harzig, Y.-Y. Chen, F. Chen, and R. Lienhart, "Addressing Data Bias Problems for Chest X-ray Image Report Generation," Aug. 2019, arXiv:1908.02123 [cs]. [Online]. Available: http://arxiv.org/abs/1908.02123 [Cited on pages 18 and 108.]

G. Tennenholtz, Y. Chow, C.-W. Hsu, J. Jeong, L. Shani, A. Tulepbergenov, D. Ramachandran, M. Mladenov, and C. Boutilier, "Demystifying Embedding Spaces using Large Language Models," Mar. 2024, arXiv:2310.04475 [cs]. [Online]. Available: http://arxiv.org/abs/2310.04475 [Cited on page 19.]

C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He, H. Peng, J. Li, J. Wu, Z. Liu, P. Xie, C. Xiong, J. Pei, P. S. Yu, and L. Sun, "A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT," May 2023, arXiv:2302.09419 [cs]. [Online]. Available: http://arxiv.org/abs/2302.09419 [Cited on page 19.]

Y. Liang, H. Wen, Y. Nie, Y. Jiang, M. Jin, D. Song, S. Pan, and Q. Wen, "Foundation Models for Time Series Analysis: A Tutorial and Survey," Jun. 2024, arXiv:2403.14735 [cs]. [Online]. Available: http://arxiv.org/abs/2403.14735 [Cited on page 19.]

D. Juyal, H. Padigela, C. Shah, D. Shenker, N. Harguindeguy, Y. Liu, B. Martin, Y. Zhang, M. Nercessian, M. Markey, I. Finberg, K. Luu, D. Borders, S. A. Javed, E. Krause, R. Biju, A. Sood, A. Ma, J. Nyman, J. Shamshoian, G. Chhor, D. Sanghavi, M. Thibault, L. Yu, F. Najdawi, J. A. Hipp, D. Fahy, B. Glass,

E. Walk, J. Abel, H. Pokkalla, A. H. Beck, and S. Grullon, “PLUTO: Pathology-Universal Transformer,” May 2024, arXiv:2405.07905 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2405.07905 [Cited on page 20.]

S. Alfasly, P. Nejat, S. Hemati, J. Khan, I. Lahr, A. Alsaafin, A. Shafique, N. Comfere, D. Murphree, C. Meroueh, S. Yasir, A. Mangold, L. Boardman, V. H. Shah, J. J. Garcia, and H. R. Tizhoosh, “Foundation Models for Histopathology—Fanfare or Flair,” Mayo Clinic Proceedings: Digital Health, vol. 2, no. 1, pp. 165–174, Mar. 2024, publisher: Elsevier. [Online]. Available: https://www.mcpdigitalhealth.org/article/S2949-7612(24)00014-2/fulltext [Cited on pages 20 and 21.]

F. Li, L. Zhu, T. Wang, J. Li, Z. Zhang, and H. T. Shen, “Cross-Modal Retrieval: A Systematic Review of Methods and Future Directions,” Oct. 2023, arXiv:2308.14263 [cs]. [Online]. Available: http://arxiv.org/abs/2308.14263 [Cited on page 21.]

Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, and J. Zou, “A visual-language foundation model for pathology image analysis using medical Twitter,” Nature Medicine, vol. 29, no. 9, pp. 2307–2316, Sep. 2023. [Cited on page 21.]

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning Transferable Visual Models From Natural Language Supervision,” Feb. 2021, arXiv:2103.00020 [cs]. [Online]. Available: http://arxiv.org/abs/2103.00020 [Cited on pages 21, 59, and 107.]

R. Lin, L. Lenert, B. Middleton, and S. Shiffman, “A free-text processing system to capture physical findings: Canonical Phrase Identification System (CAPIS),” Proceedings. Symposium on Computer Applications in Medical Care, pp. 843–847, 1991. [Cited on page 22.]

M. L. Gundersen, P. J. Haug, T. A. Pryor, R. van Bree, S. Koehler, K. Bauer, and B. Clemons, “Development and Evaluation of a Computerized Admission Diagnoses Encoding System,” Computers and Biomedical Research, vol. 29, no. 5, pp. 351–372, Oct. 1996. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010480996900269 [Cited on page 22.]

S. Wu, K. Roberts, S. Datta, J. Du, Z. Ji, Y. Si, S. Soni, Q. Wang, Q. Wei, Y. Xiang, B. Zhao, and H. Xu, “Deep learning in clinical natural language processing: a methodical review," Journal of the American Medical Informatics Association, vol. 27, no. 3, pp. 457–470, Mar. 2020. [Online]. Available: https://doi.org/10.1093/jamia/ocz200 [Cited on pages 22 and 23.]

H. Wu, M. Wang, J. Wu, F. Francis, Y.-H. Chang, A. Shavick, H. Dong, M. T. C. Poon, N. Fitzpatrick, A. P. Levine, L. T. Slater, A. Handy, A. Karwath, G. V. Gkoutos, C. Chelala, A. D. Shah, R. Stewart, N. Collier, B. Alex, W. Whiteley, C. Sudlow, A. Roberts, and R. J. B. Dobson, "A survey on clinical natural language processing in the United Kingdom from 2007 to 2022," npj Digital Medicine, vol. 5, no. 1, pp. 1–15, Dec. 2022, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41746-022-00730-6 [Cited on page 23.]

M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, and S. Azam, "A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges," IEEE Access, vol. 12, pp. 26 839–26 874, 2024, conference Name: IEEE Access. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10433480 [Cited on page 23.]

T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, and V. Tseng, "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models," PLOS Digital Health, vol. 2, no. 2, p. e0000198, Feb. 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9931230/ [Cited on page 23.]

S. Biswas, "ChatGPT and the Future of Medical Writing," Radiology, vol. 307, no. 2, p. e223312, Apr. 2023. [Cited on page 23.]

F. Eggmann, R. Weiger, N. U. Zitzmann, and M. B. Blatz, "Implications of large language models such as ChatGPT for dental medicine," Journal of Esthetic and Restorative Dentistry, vol. 35, no. 7, pp. 1098–1102, 2023, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jerd.13046. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/jerd.13046 [Cited on page 23.]

K. Jeblick, B. Schachtner, J. Dexl, A. Mittermeier, A. T. Stüber, J. Topalis, T. Weber, P. Wesp, B. O. Sabel, J. Ricke, and M. Ingrisch, "ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports," European Radiology, vol. 34, no. 5, pp. 2817–2825, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11126432/ [Cited on page 23.]

J. Cheng, "Applications of Large Language Models in Pathology," Bioengineering, vol. 11, no. 4, p. 342, Apr. 2024, number: 4 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2306-5354/11/4/342 [Cited on page 23.]

H. Liu, W. Xue, Y. Chen, D. Chen, X. Zhao, K. Wang, L. Hou, R. Li, and W. Peng, "A Survey on Hallucination in Large Vision-Language Models," May 2024, arXiv:2402.00253 [cs]. [Online]. Available: http://arxiv.org/abs/2402.00253 [Cited on page 23.]

S.-Q. Yan, J.-C. Gu, Y. Zhu, and Z.-H. Ling, "Corrective Retrieval Augmented Generation," Feb. 2024, arXiv:2401.15884 [cs]. [Online]. Available: http://arxiv.org/abs/2401.15884 [Cited on page 23.]

K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, M. Schaekermann, A. Wang, M. Amin, S. Lachgar, P. Mansfield, S. Prakash, B. Green, E. Dominowska, B. A. y. Arcas, N. Tomasev, Y. Liu, R. Wong, C. Semturs, S. S. Mahdavi, J. Barral, D. Webster, G. S. Corrado, Y. Matias, S. Azizi, A. Karthikesalingam, and V. Natarajan, "Towards Expert-Level Medical Question Answering with Large Language Models," May 2023, arXiv:2305.09617 [cs]. [Online]. Available: http://arxiv.org/abs/2305.09617 [Cited on page 24.]

M. Zheng, J. Pei, and D. Jurgens, "Is "A Helpful Assistant" the Best Role for Large Language Models? A Systematic Evaluation of Social Roles in System Prompts," Nov. 2023, arXiv:2311.10054 [cs]. [Online]. Available: http://arxiv.org/abs/2311.10054 [Cited on page 24.]

J. Miao, C. Thongprayoon, S. Suppadungsuk, O. A. Garcia Valencia, and W. Cheungpasitporn, "Integrating Retrieval-Augmented Generation with Large Language Models in Nephrology: Advancing Practical Applications," Medicina, vol. 60, no. 3, p. 445, Mar. 2024, number: 3 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1648-9144/60/3/445 [Cited on page 25.]

M. A. Quidwai and A. Lagana, “A RAG Chatbot for Precision Medicine of Multiple Myeloma,” Mar. 2024, pages: 2024.03.14.24304293. [Online]. Available: https://www.medrxiv.org/content/10.1101/2024.03.14.24304293v1 [Cited on page 25.]

J. D. Bancroft, K. S. Suvarna, and C. Layton, Theory and Practice of Histological Techniques: Expert Consult: Online and Print. Elsevier, Nov. 2018. [Cited on page 27.]

J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, B. Foster, M. Moser, E. Karasik, B. Gillard, K. Ramsey, S. Sullivan, J. Bridge, H. Magazine, J. Syron, J. Fleming, L. Siminoff, H. Traino, M. Mosavel, L. Barker, S. Jewell, D. Rohrer, D. Maxim, D. Filkins, P. Harbach, E. Cortadillo, B. Berghuis, L. Turner, E. Hudson, K. Feenstra, L. Sobin, J. Robb, P. Branton, G. Korzeniewski, C. Shive, D. Tabor, L. Qi, K. Groch, S. Nampally, S. Buia, A. Zimmerman, A. Smith, R. Burges, K. Robinson, K. Valentino, D. Bradbury, M. Cosentino, N. Diaz-Mayoral, M. Kennedy, T. Engel, P. Williams, K. Erickson, K. Ardlie, W. Winckler, G. Getz, D. DeLuca, D. MacArthur, M. Kellis, A. Thomson, T. Young, E. Gelfand, M. Donovan, Y. Meng, G. Grant, D. Mash, Y. Marcus, M. Basile, J. Liu, J. Zhu, Z. Tu, N. J. Cox, D. L. Nicolae, E. R. Gamazon, H. K. Im, A. Konkashbaev, J. Pritchard, M. Stevens, T. Flutre, X. Wen, E. T. Dermitzakis, T. Lappalainen, R. Guigo, J. Monlong, M. Sammeth, D. Koller, A. Battle, S. Mostafavi, M. McCarthy, M. Rivas, J. Maller, I. Rusyn, A. Nobel, F. Wright, A. Shabalin, M. Feolo, N. Sharopova, A. Sturcke, J. Paschal, J. M. Anderson, E. L. Wilder, L. K. Derr, E. D. Green, J. P. Struewing, G. Temple, S. Volpi, J. T. Boyer, E. J. Thomson, M. S. Guyer, C. Ng, A. Abdallah, D. Colantuoni, T. R. Insel, S. E. Koester, A. R. Little, P. K. Bender, T. Lehner, Y. Yao, C. C. Compton, J. B. Vaught, S. Sawyer, N. C. Lockhart, J. Demchok, and H. F. Moore, “The Genotype-Tissue Expression (GTEx) project,” Nature Genetics, vol. 45, no. 6, pp. 580–585, Jun. 2013, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/ng.2653 [Cited on page 36.]

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040.pdf [Cited on page 117.]

R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2015/html/Vedantam_CIDEr_Consensus-Based_Image_2015_CVPR_paper.html [Cited on page 117.]

M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in Proceedings of the ninth workshop on statistical machine translation, 2014, pp. 376–380. [Online]. Available: https://aclanthology.org/W14-3348.pdf [Cited on page 117.]

C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in Text summarization branches out, 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013.pdf [Cited on page 117.]

P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic Propositional Image Caption Evaluation," in Computer Vision – ECCV 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, vol. 9909, pp. 382–398, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-319-46454-1_24 [Cited on page 117.]

S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of spider," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 873–881. [Online]. Available: http://openaccess.thecvf.com/content_iccv_2017/html/Liu_Improved_Image_Captioning_ICCV_2017_paper.html [Cited on page 117.]