POLITECNICO DI TORINO

Master's Degree in ICT for Smart Societies

Politecnico di Torino

Master's Degree Thesis

Uncertainty-aware Renal Cell Carcinoma Subtype Classification

Supervisors

Prof. Santa DI CATALDO

Prof. Francesco PONZIO

Prof. Xavier DESCOMBES

Candidate

Seyed Mohammad Mehdi HOSSEINI

September 2024

Summary

Kidney cancer requires accurate and timely diagnosis to guide effective treatment as a particularly prevalent cancer in older adults. Early detection and the precise identification of cancer subtypes and stages can significantly influence therapeutic decisions which leads to improved patient outcomes, preventing metastasis, and increasing survival rates. Renal cell carcinoma (RCC), the most common form of kidney cancer, is a heterogeneous type of cancer that comprises several subtypes, including clear cell RCC, papillary RCC, chromophobe RCC, and oncocytoma. Each of these subtypes comes with a unique biological behavior and treatment response. Currently, clinicians adopt a step-by-step approach to classify these subtypes based on the level of diagnostic uncertainty, starting with tumor morphology. While tumor morphology serves as the initial method for evaluation, its reliance on overlapping features, such as similar cellular structures and staining patterns, often introduces ambiguity and complicates accurate classification. This method is also a time-intensive process which requires significant expertise. When uncertainty persists during tumor morphology analysis, clinicians proceed to more advanced techniques like immunohistochemistry (IHC) profile analysis, which, although valuable, comes with its own set of challenges, including high costs and the need for specialized expertise. Together, these factors add complexity to clinical decision-making and extend the diagnostic timeline. Recent advancements in deep learning, particularly convolutional neural networks (CNNs), have opened new avenues for enhancing cancer diagnosis. CNNs are adept at processing histopathological images due to their ability to capture complex spatial patterns. This capability enables the differentiation of various cell structures and tissue textures, which are critical elements in the accurate classification of cancer subtypes. CNNs are less computationally intensive and work more effectively with smaller datasets, which are common in medical imaging. This makes them a better fit for tasks in this field than more complex models like Transformers. In our work, we leverage the strengths of supervised models, which benefit from well-labeled data, allowing for precise pattern recognition and reducing the margin for error. While self-supervised and weakly-supervised methods can extract valuable patterns from vast amounts of data with minimal labeling, their performance may fall short when distinguishing between

subtle differences required for accurate medical diagnoses that often demand highly detailed and specific labeling. Building on these advancements, we propose a hybrid model that incorporates deep learning for the initial detection of tumor regions and subtype classification. In cases where the model exhibits uncertainty in the initial classification, it triggers a secondary validation step using traditional machine learning techniques applied to IHC profile for more accurate confirmation. This integrated approach allows for a more comprehensive diagnostic framework, merging morphological insights from histopathology with molecular data from IHC. As a result, our model enhances RCC subtype classification accuracy while reducing processing time and cost, offering a promising solution for improving diagnostic precision in clinical practice.

Acknowledgements

As I reach the end of this long journey, I find myself deeply grateful for the many incredible people who have supported and guided me along the way. Completing this thesis has been one of the most challenging, yet exciting experiences of my life, and it is only with the encouragement, wisdom, and love of those around me that I have been able to persevere.

First and foremost, I would like to express my deepest appreciation to my supervisors, Prof. Di Cataldo, Prof. Ponzio and Prof. Descombes. Your insight, patience, and dedication have been an inspiration and I feel incredibly fortunate to have had the opportunity to learn from you. I would also like to extend my thanks to my incredible colleagues in the Morpheme Research Group, including Dr. Ambrosetti, Dr. Hannetel, and Mr. Mohamad. Your invaluable guidance, insightful feedback, and constant support have played a significant role in shaping this work.

Finally, to my family and friends, whose love and encouragement have been the foundation upon which I stand, no words can fully express my gratitude. To my parents and my brother, who have sacrificed more than I can ever repay and who have always believed in me, thank you for being my strength.

This work is as much a reflection of your guidance, encouragement, and love as it is of my own effort. I dedicate this thesis to you all, with all my heart.

Seyed Mohammad Mehdi Hosseini

Table of Contents

List of Tables

List of Figures

Acronyms

WHO
World Health Organization

AI
Artificial Intelligence

RCC
Renal Cell Carcinoma

ccRCC
Clear Cell Renal Cell Carcinoma

ccpRCC
Clear Cell Papillary Renal Cell Carcinoma

pRCC
Papillary Renal Cell Carcinoma

chRCC
Chromophobe Renal Cell Carcinoma

DL
Deep Learning

ML
Machine Learning

DNN
Deep Neural Networks

CNN

Convolutional Neural Networks

IHC

Immunohistochemistry

H&E

Hematoxylin and Eosin

HES

Hematoxylin, Eosin and Natural Saffron

CAIX

Carbonic Anhydrase IX

CK7

Cytokeratin 7

HMWCK

High-Molecular-Weight Cytokeratin

FISH

Fluorescence In Situ Hybridization

CGH

Comparative Genomic Hybridization

GANs

Generative Adversarial Networks

TPU

Tensor Processing Unit

RNN

Recurrent Neural Networks

WSI

Whole Slide Image

XAI

Explainable Artificial Intelligence

SL

Supervised Learning

DAG-SVM

Directed Acyclic Graph Support Vector Machine

WSL

Weakly-supervised Learning

MIL

Multiple Instance Learning

SSL

Self-supervised Learning

SOTA

state-of-the-art

ViT

Vision Transformer

TCGA

The Cancer Genome Atlas

mIF

multiplex immunofluorescence

PAS

Periodic Acid-Schiff

ASAP

Automated Slide Analysis Platform

HER2

Human Epidermal Growth Factor Receptor 2

IPA

Interactive Pointwise Attention

XML

eXtensible Markup Language

CD10

Cluster of Differentiation 10

PAX8

Paired Box Gene 8

P504

Alpha-methylacyl-CoA Racemase

ECAD

E-cadherin

VIM

Vimentin

ROI

Region of interest

XML-ROIs

XML-based regions of interest

WSI-ROIs

WSI-based regions of interest

SMOTE

(Synthetic Minority Over-sampling Technique)

KNN

K Nearest Neighbors

ENN

Edited Nearest Neighbors

IQR

Interquartile Range

ExpertDT

ExpertDeepTree

ExpertDUT

ExpertDeepUncertainTree

MC

Monte Carlo

MC-Root

Monte Carlo Root

Chapter 1

Introduction

1.1 Renal Cell Carcinoma

The kidneys are essential organs that maintain overall body homeostasis by filtering waste products from the blood, regulating electrolyte balance, controlling blood pressure, and producing hormones critical for red blood cell production and bone health. Each kidney contains about one million nephrons, which are the functional units responsible for filtering blood and forming urine [1, 2]. The kidneys are located retroperitoneally on either side of the spine, and their strategic placement and highly vascular nature make them susceptible to various pathological conditions, including malignancies [3]. The nephrons perform key processes such as glomerular filtration, tubular absorption, and secretion to form urine, while also regulating water, electrolytes, and blood pH [4]. Additionally, the kidneys play a critical role in the production of hormones like erythropoietin and the activation of vitamin D, which are vital for red blood cell production and calcium metabolism [5].

Kidney cancers represent a broad spectrum of malignancies originating from different tissues within the kidney, with renal cell carcinoma (RCC) being the most common and diverse type, in which the cancer cells are found in the lining of Renal Tubules as shown in Figure 1.1 [6]. RCC representing a significant health challenge globally. It accounts for approximately 90% of all kidney cancers, making it a primary concern in nephrological oncology. RCC is not a singular entity but rather a collective term encompassing various subtypes, each with unique histological and molecular characteristics, as defined by the World Health Organization (WHO) classification [6]. The most prevalent subtype is clear cell RCC, marked by its distinct clear cytoplasm, comprising about 75% of RCC cases. Other notable subtypes include papillary RCC, further categorized into type 1 and type 2, and chromophobe RCC, which is recognized for its pale, granular cells [7], as depicted in Figure 1.2. In addition to these, oncocytomas, though typically benign, share some

overlapping features with RCC, complicating diagnosis [8]. Rarer forms, such as collecting duct carcinoma and molecularly distinct variants like TFE3-rearranged RCC and hereditary leiomyomatosis-associated RCC, highlight the complexity and variability within the RCC spectrum [9].

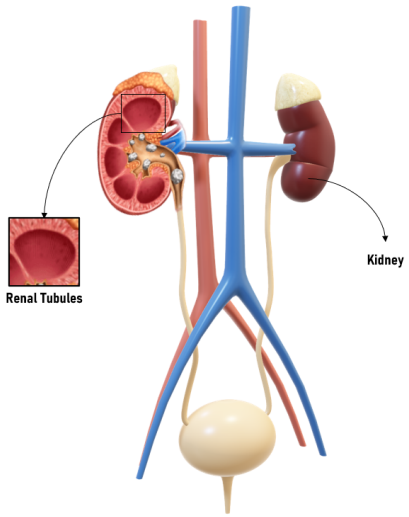


Figure 1.1: Anatomy of the Kidney Showing the Location of the Renal Tubules

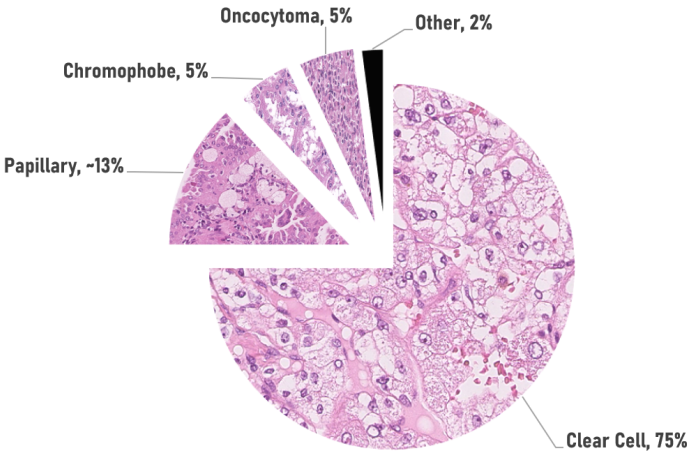
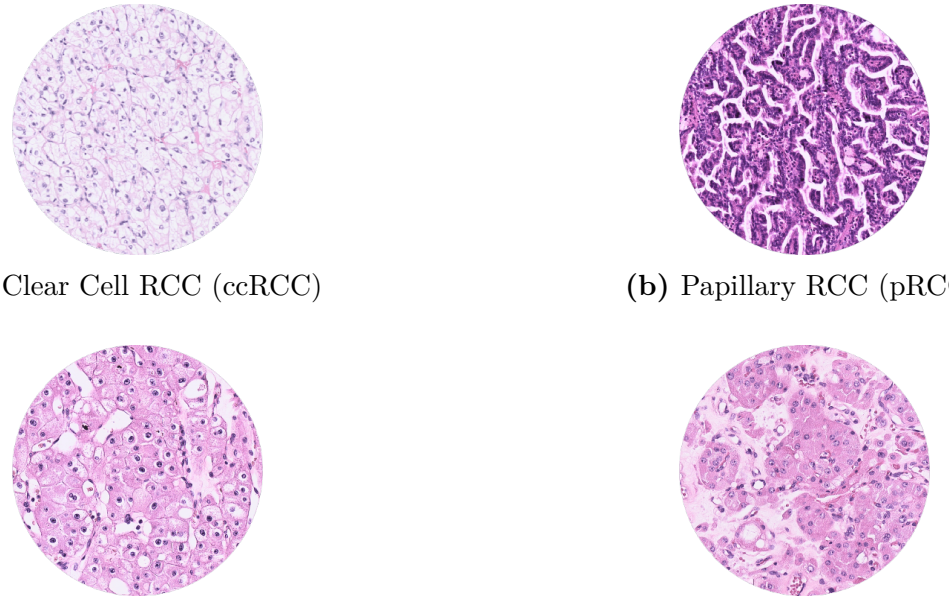


Figure 1.2: Histological Distribution of Renal Cell Carcinoma Subtypes

The incidence of RCC has been steadily increasing worldwide, with more than 76,000 new cases and over 13,780 deaths reported in the United States alone in

2021. This upward trend can be attributed to a combination of improved imaging techniques leading to incidental discoveries during imaging for other conditions, often at a localized stage where the disease is confined to the kidney. However, despite early detection, approximately one-third of patients will experience disease progression to advanced stages, including regional or distant metastases. This progression reduces survival rates significantly and presents a huge clinical challenge. Patients diagnosed with localized RCC have a five-year survival rate of around 70%, however, this rate drops to a mere 13% for those with distant metastases. It highlights the aggressive nature of the disease once it spreads beyond the kidney. The accurate classification of these subtypes is very important, as it extremely impacts prognosis and informs treatment strategies tailored to each tumor's specific biology [10].



(a) Clear Cell RCC (ccRCC)

(b) Papillary RCC (pRCC)

(c) Chromophobe RCC (chRCC)

(d) Oncocytoma

Figure 1.3: Histopathology slides of different RCC subtypes with H&E staining

1.1.1 Microscopic Morphology

The microscopic morphology of RCC subtypes reveals distinct cellular patterns that are critical for diagnosis. The most common subtype, ccRCC, is characterized by cells with clear cytoplasm due to glycogen and lipid content [11]. In pRCC, RCC shows papillary formations with foamy macrophages [12]. Chromophobe RCC exhibits pale cytoplasm with prominent cell membranes and perinuclear halos [12]. Finally, oncocytomas are composed of densely packed cells with abundant

granular, eosinophilic cytoplasm, and round, uniform nuclei with small nucleoli. The cells are arranged in nests or sheets, and the stroma is often minimal [13]. Histopathologically, RCC subtypes display distinct architectural patterns that help in their identification under a microscope. The majority of diagnoses begin with microscopic morphology analysis, either through direct examination under microscopes or using Whole Slide Images (WSIs) at various magnification levels. This approach is standard unless cases present complexities and uncertainties that cannot be adequately assessed through this examination alone.

1.1.2 Immunohistochemistry Profile

In challenging cases, immunohistochemistry is essential for confirming the diagnosis and distinguishing between RCC subtypes stated by microscopic morphology. Specific immunohistochemical markers aid in identifying the subtypes. For instance, Carbonic Anhydrase IX (CAIX) is frequently used for clear cell RCC [11], while Cytokeratin 7 (CK7) is commonly used to identify papillary RCC [12]. Chromophobe RCC can be identified by the combination of CK7 positivity and S100-A1 expression [12] while oncoctyoma can be identidied with the absence of CK7. Immunohistochemistry not only improves diagnostic accuracy but also supports prognostic assessments by highlighting specific molecular pathways associated with each RCC subtype. These markers underscore the complexity and precision required in the histopathological diagnosis of RCC. After establishing the immunohistochemistry profile, further molecular analyses are often pursued to deepen the understanding of the tumor's genetic landscape. Techniques such as karyotyping, fluorescence in situ hybridization (FISH), and comparative genomic hybridization (CGH) arrays are commonly used to detect chromosomal abnormalities and genetic mutations [14].

1.2 Artificial Intelligence

Artificial Intelligence (AI) refers to the simulation of human intelligence processes by machines, particularly computer systems. These processes include learning, reasoning, problem-solving, perception, and language understanding. AI has innovated numerous fields by enabling machines to perform tasks that typically require human intelligence, such as decision-making, visual perception, speech recognition, and language translation. AI systems can be broadly categorized into rule-based systems, expert systems, and machine learning-based systems, the latter being the most transformative in recent years due to the development of more sophisticated models and greater availability of data.

Machine Learning (ML) and Deep Learning (DL) are subsets of AI that have transformed various industries, particularly in fields like computer vision, natural

language processing, and medical diagnosis. Machine Learning refers to algorithms that allow computers to learn patterns from data without being explicitly programmed for every task. Traditional ML techniques, such as decision trees, support vector machines (SVM), and random forests, rely on statistical methods and involve feature engineering, where domain experts manually identify and select features relevant to a particular problem. This requires significant prior knowledge and domain expertise, but has been successfully applied to tasks such as classification, regression, and clustering across a variety of domains.

Deep Learning, a subfield of ML, employs neural networks with multiple layers, known as deep neural networks (DNN). These networks can automatically learn hierarchical representations of data, reducing the need for manual feature extraction. Convolutional Neural Networks (CNN), a type of DNN, have become fundamental in computer vision, due to their ability to learn spatial hierarchies of features through convolutional layers. CNNs are used in a wide range of applications, from image classification and object detection to image segmentation [15]. In medical image analysis, CNNs are particularly prominent, excelling at tasks such as detecting tumors in radiological and histopathological images, diagnosing retinal diseases, and segmenting organs in MRI scans [16]. Famous CNN architectures like AlexNet, VGGNet, ResNet, and U-Net and their improved architectures have paved the way for advancements in medical image analysis.

The application of DL models in medical diagnosis has shown remarkable success, particularly in analyzing medical images like CT scans, X-rays, MRIs, and histopathology slides. CNNs have been applied to detect diseases such as lung cancer, brain tumors, breast cancer and kidney cancer achieving diagnostic accuracy that often rivals or exceeds human experts. In histopathology, CNN models have been used to classify tissue images, such as in the case of lung cancer where histopathological images are analyzed to distinguish between different cancer types like adenocarcinoma and squamous cell carcinoma [17]. CNNs have also been used to accurately classify subtypes of renal cell carcinoma (RCC), including ccRCC, pRCC, chRCC and oncocytoma, based on histopathological features [18]. These models not only detect and classify diseases but also assist in early diagnosis and treatment planning by identifying subtle details that human experts may overlook. Recently, Generative Adversarial Networks (GANs) have emerged as powerful tools for generating synthetic medical images, which can augment training data and improve model performance in scenarios where labeled data is scarce [19].

1.2.1 AI in Medical Diagnosis

AI has rapidly emerged as a transformative tool in medical diagnosis, particularly through its ability to analyze vast amounts of patient data with speed and accuracy.

Recent advancements in machine learning algorithms, alongside increased computational power provided by modern GPUs, have accelerated the development of AI applications in healthcare. These technologies are being integrated into clinical settings to assist in diagnostics, improving efficiency and potentially reducing human error. In particular, AI excels at analyzing complex data from sources like electronic health records, genomic data, and medical images. Among the various AI learning paradigms, supervised learning is one of the dominant approaches in medical image analysis. Supervised models rely on labeled datasets, where each image is associated with a known diagnosis or classification. The use of large annotated datasets, such as ImageNet and the CAMELYON dataset, has played a crucial role in advancing medical imaging tasks by providing the necessary data to train AI models effectively [20].

However, the need for extensive labeled data poses a significant challenge. As annotating medical images is time-consuming and requires expert knowledge, alternative learning paradigms are gaining interest. Semi-supervised learning, which uses a small amount of labeled data and a large amount of unlabeled data, aims to bridge this gap. Similarly, self-supervised learning and weakly supervised learning seek to reduce the dependency on large labeled datasets by leveraging pretext tasks or partial labels [15, 21]. These approaches are becoming increasingly popular in medical imaging, where high-quality labeled data can be scarce.

While these alternative paradigms hold promise, supervised learning continues to be the most reliable method in medical image analysis, thanks to its proven accuracy and robustness in diagnostic tasks. Supervised learning offers clear advantages when sufficient labeled data is available, particularly in clinical environments where precision and interpretability are paramount.

Chapter 2

State of the Art

2.1 AI in Histopathology

The integration of computer analysis into histopathology began in the early 1960s when image analysis techniques were applied to digitized slides of cells, allowing pathologists to quantify characteristics such as cell size and chromatin distribution [22]. Initially, these methods relied on manual feature engineering, where specific attributes were selected based on biological insights. Traditional histopathological analysis, despite being the gold standard, faces challenges such as observer variability, labor-intensive processes, and complexity in discerning cancer subtypes, which has led to the exploration of more intelligent and smart solutions [23]. As technology advanced, machine learning algorithms gradually automated these processes, particularly in the early 2000s, when radiomics and pathomics approaches allowed for the analysis of texture, shape, and density features without the need for manually defined characteristics [24]. These machine learning methods played a crucial role in early diagnostic tasks such as the classification of tumor types, biomarker prediction, and the grading of histological features, offering a more efficient alternative to manual evaluation [25]. While early machine learning methods improved efficiency, they still required manual feature selection.

The real transformation, however, occurred with the introduction of deep learning in the last decade, particularly through the application of CNNs to WSIs. One of the key turning points in this field was the development of models trained on large open-source datasets such as CAMELYON, which focused on detecting breast cancer metastases in lymph nodes [20]. In addition to these widely utilized open-source datasets, numerous valuable private datasets are also playing a crucial role in advancing current research efforts. The success of these models in achieving high accuracy motivated researchers to apply deep learning to histopathology across various cancers. This focus has been recently extended to RCC, where advanced

deep learning models are being developed to improve detection, classification, and segmentation by analyzing histopathological slides, offering the potential for more accurate diagnoses and better patient outcomes. Depending on the availability and quality of annotations in the input data, these models can be generally classified into supervised learning models, which rely on well-labeled datasets; weakly-supervised models, which handle incomplete or noisy labels; and self-supervised models, which generate their own supervisory signals from the data. The following sections will explore these approaches in greater detail.

2.1.1 Supervised Learning Models

Supervised learning (SL) models have shown promising advancements in classifying RCC subtypes. These approaches, while highly accurate, differ in their methodologies and real-world applicability, especially in the clinical setting. Each model contributes unique strengths and presents challenges that highlight both the potential and limitations of AI-driven histopathological analysis in RCC.

Zhu M. et al. [26] developed a deep neural network specifically designed to classify RCC subtypes—ccRCC, pRCC, and chRCC and oncocytoma. They evaluated their model performance on both surgical resection and biopsy slides. Their model achieved 95% accuracy on surgical resection considering also normal patients and 93% across different RCC subtypes. They have also validated their results with external data from TCGA open-source database, demonstrating their model generalizability, however this validation doesn't include oncocytoma subtype. Generally, including oncocytoma as a benign entity in the classification task makes the models become particularly relevant for clinical settings, where distinguishing between benign and malignant tumors is crucial.

Abdeltawab et al. [27] developed an automated classification model to distinguish between ccRCC and ccpRCC, achieving an accuracy of 91% in identifying ccpRCC. Their method employed multiple CNNs trained on patches of different sizes to recognize features at varying scales, enhancing the model's ability to capture diverse histological patterns. By combining the outputs of these CNNs, the model was able to achieve robust performance in both internal and external datasets. The use of multiple CNNs at different scales allowed for improved feature recognition, particularly for subtle differences between similar subtypes like ccRCC and ccpRCC. Although their study did not focus on other RCC subtypes, it remains highly valuable as it addresses the most common subtype (ccRCC) versus one of the rarer subtypes (ccpRCC), which often presents clinical challenges and can significantly impact treatment approaches. They also highlighted their model's ability to serve as a clinical tool to assist pathologists by pre-screening slides and offering a second opinion, thus improving diagnostic accuracy and reducing false negatives.

Fenstermaker et al. [28] introduced a deep learning model capable of classifying

RCC subtypes—ccRCC, pRCC, and chRCC—with an impressive accuracy exceeding 99% from small biopsy samples. One major advantage of this model is its ability to reduce the need for repeat biopsies by improving diagnostic consistency, which is particularly beneficial in cases where renal mass biopsies (RMBs) are insufficient. However, its limitation lies in its narrow focus: the model was developed for specific RCC subtypes and did not address benign tumors such as oncocytomas, which often pose a diagnostic challenge when distinguishing them from RCCs, particularly chRCC. Tabibu et al. [29] applied deep learning techniques to classify RCC subtypes similar to Fenstermaker et al. [28], but they have also included normal patients into their study. They used two pre-trained CNNs and fine-tuned the models for RCC data. In addition, their approach involved using a directed acyclic graph support vector machine (DAG-SVM) on top of the CNN architecture to improve subtype classification. Their model showed good accuracy in distinguishing between different subtypes but with the same limitation of Fenstermaker et al. [28]. They also didn't include oncocytoma into their study and they didn't validate their model with external datasets.

Chen et al. [30] developed a model to diagnose and predict survival rates for ccRCC patients compared to non-ccRCC patients, including non-tumor cases and other RCC subtypes like pRCC and chRCC. They manually extracted key features from microscopic images using a third-party toolbox and combined these with clinical data to train a machine learning model, using LASSO regression to identify the most important factors. This manual feature extraction process allows for greater interpretability compared to deep learning models, where features are often automatically learned and can act as a "black box". However, manual extraction can be time-consuming and may miss subtle image patterns that deep learning models could capture. While the model offers valuable prognostic insights and incorporates clinical factors like tumor stage and nuclei grading, it lacks the automatic feature discovery of deep learning models, which limits its ability to generalize across diverse datasets.

Ponzio et al. [18] introduced a hybrid approach known as ExpertDeepTree (ExpertDT), combining CNN-based deep learning with pathologist-driven expertise. Their study demonstrated that integrating human knowledge into the model improved its ability to classify RCC subtypes. ExpertDT's hierarchical tree structure, guided by pathologists, significantly outperformed traditional CNNs, achieving 95% accuracy across four RCC aforementioned subtypes. This model addresses the common issue of misclassifications in cases with overlapping morphological features especially between chRCC and oncocytoma. They illustrated that the use of expert knowledge to enhance AI decision-making marks a significant step forward.

Overall, supervised learning models for RCC subtype classification are rapidly advancing, but challenges remain. While they show high accuracy, generalizability across diverse clinical populations and the ability to handle benign mimickers like

oncocytomas or rare RCC variants are areas needing further research. Moreover, integrating these models into clinical workflows, especially for predicting patient outcomes, will require ongoing refinement and validation.

2.1.2 Weakly-supervised Learning Models

While SL models have achieved remarkable accuracy, their dependency on large, annotated datasets presents a significant bottleneck among many clinical and research settings. This challenge has led to the exploration of learning approaches, which offer less detailed annotations. Weakly-supervised Learning (WSL) models uses less detailed labels for training that are easier and faster to collect. These labels may be noisy and need less fine-grained supervision to achieve, but they allow for creating large datasets quickly. One approach uses general labels like slide-level tags, e.g. in RCC, ccRCC, pRCC, etc. In histopathology applications, while this approach reduces annotation costs, it can introduce more errors compared to fully supervised models. However, despite the potential trade-offs in accuracy, weak supervision significantly lightens the annotation burden on pathologists while still supporting effective RCC subtype classification.

Brendel and Bethge. [31] introduced a Bag-of-Local-Features model that demonstrated impressive performance on ImageNet, suggesting the utility of this approach for medical imaging. In RCC, this framework could be adapted to process histopathological patches to improve feature extraction and classification accuracy. Similarly, Gadermayr and Tschuchnig [32] provided a comprehensive review on possible limitations on Multiple Instance Learning (MIL) in digital pathology, emphasizing the challenge of extracting discriminative features from non-annotated regions and highlighting the need for more advanced attention mechanisms. Hou et al. [33] addressed these concerns with a patch-based CNN that effectively handled the large-scale nature of WSIs, laying a foundation for RCC subtype classification.

Moreover, Jia et al. [34] proposed a constrained deep weak supervision model for histopathology image segmentation, focusing on incorporating prior knowledge to improve segmentation accuracy. This technique is particularly useful in the classification of RCC subtypes, as it reduces false positives in ambiguous regions. Furthermore, Ilse et al. [35] introduced an attention-based MIL framework that dynamically adjusts focus on relevant regions, significantly enhancing interpretability and performance in WSL models.

In a novel study, Abu Haeyeh et al. [36] developed a deep-learning-based framework for the classification of renal histopathology images based on multiscale CNN model. Their framework demonstrated high accuracy in distinguishing RCC subtypes such as ccRCC and ccpRCC along with normal tissues including parenchyma and fat tissues, outperforming standard models like ResNet-50. Furthermore, Zheng et al. [37] proposed a human-machine fusion model for RCC grading. Their

SSL-CLAM model utilized histopathology WSIs to successfully diagnose ccRCC grades, showing the potential of deep learning to enhance diagnostic precision through a human-machine collaborative approach.

Despite these advancements, WSL models still face limitations. For instance, while attention mechanisms improve interpretability, they can introduce biases if trained on limited datasets. Additionally, WSL algorithms typically require large amounts of data, which is often not readily available in the context of RCC studies. Moreover, domain adaptation remains a challenge, as models trained on specific datasets (e.g., TCGA) may not generalize well to others. Collaborative approaches combining human expertise with AI, as demonstrated by Zheng et al. [37], offer promising results but require further validation in clinical settings.

2.1.3 Self-supervised Learning Models

To advance beyond the need for labeled data, Self-supervised Learning (SSL) enables models to discover meaningful patterns from raw images without requiring annotations. In histopathology, SSL techniques have emerged as a promising approach for reducing the dependency on large, annotated datasets. SSL methodologies allow models to learn inherent patterns from unlabeled data by solving auxiliary tasks, called pretext tasks, designed to extract meaningful features from the data. In digital pathology, the application of standard SSL techniques, often developed for natural images, presents specific challenges due to the homogeneity of tissue structures and the symmetrical shapes of cells and nuclei, making tasks like patch localization or colorization less effective without significant adaptation [38].

Researchers have adapted these techniques to the histopathology domain, such as magnification prediction and ordering tasks, which involve presenting CNNs with patches from different magnifications and requiring the model to predict their relationships. Additionally, novel approaches like jigmag, where networks must correctly identify the magnification sequence of multiple patches, have outperformed traditional methods by focusing on histology-specific features [38].

Mohamad et al. [21] proposed a self-supervised learning task that enhances histopathological image analysis by interconnecting different magnification levels. Their model localizes a high-resolution tile within a global patch, addressing a key limitation of existing methods that overlook the relationships between magnification levels in WSIs. The proposed approach outperforms SOTA pretext tasks and models pre-trained on ImageNet. In addition, Wessels et al. [39] used a well-known self-supervised vision transformer (ViT) model, called DINO (self-distillation with no labels), introduced by Caron et al. [40], to predict overall survival and disease-specific survival DSS in patients with ccRCC based on histopathological images. Their model identified significant features from nuclei and peritumoural stroma, demonstrating the power of SSL in extracting clinically relevant information for

personalized treatment.

Other significant contributions include Srinidhi et al.'s consistency training approach, which integrates SSL with semi-supervised learning to enhance model generalization on pathology images [41]. By incorporating magnification-based tasks and leveraging teacher-student training, these methods have shown impressive results in histology-specific domains such as cancer classification, often surpassing traditional SSL approaches tailored to natural images [42]. These advances in SSL for histopathology promise greater accuracy in areas where annotations are limited, ultimately reducing the reliance on pathologists for manual labeling.

Lately, Chen et al. [43] have presented a novel self-supervised model, UNI, aimed at addressing the challenges of generalization and transferability in computational pathology tasks. One of the primary strengths of the model is its extensive pretraining on over 100 million images from more than 100,000 WSIs, making it highly robust for diverse tissue types and numerous pathology tasks. By leveraging the DINOv2 self-supervised learning algorithm, the model has shown strong generalization capabilities across 34 clinical tasks, including cancer subtyping and tissue classification, outperforming previous (SOTA) models such as CTransPath [44] and REMEDIS [45]. However, the model has certain limitations. For instance, while it performs exceptionally well on classification tasks, its performance on dense prediction tasks like cell segmentation is less drastic, primarily due to the lack of vision-specific biases in its architecture. Additionally, the model's computational demands, particularly for large ViT-based architectures, could hinder its accessibility for institutions with limited resources. The authors also highlight the risk of data contamination, particularly when models are trained on overlapping datasets like TCGA, potentially skewing the evaluation results. Despite these drawbacks, the UNI model's ability to excel in few-shot learning tasks and its adaptability across multiple pathology domains underscore its potential as a foundational model for computational pathology applications.

2.2 AI in Immunohistochemistry

Recent advancements in ML and DL have now made their way into IHC, particularly in the domains of pathology and cancer research. The application of AI algorithms has proven to significantly enhance diagnostic precision by automating the interpretation of IHC markers across a variety of cancers, including breast, prostate, and lung, facilitating targeted therapeutic strategies and prognostic stratification [46]. Lems et al. [47] exemplified this through their development of a novel color-agnostic DL model tailored for IHC WSI analysis. Their model demonstrated robust performance across various staining modalities, achieving notable improvements in color deconvolution and multiplex immunofluorescence (mIF)

image analysis. Bannier et al. [48] further contributed to this field by creating a DL model designed to automate the quantification of human epidermal growth factor receptor 2 (HER2) expression in invasive breast cancers, addressing discrepancies in the identification of HER2-low tumors and achieving high levels of sensitivity and specificity across multiple datasets. The integration of multiplex IHC with ML tools such as ImmuNet has allowed for a more detailed spatial analysis of immune cells within the tumor microenvironment, providing valuable insights for biomarker discovery and prognostic evaluation in cancer [49]. Moreover, advanced deep CNNs, like the interactive pointwise attention (IPA) network, have demonstrated significant improvements in the accuracy of IHC image classification by mitigating issues such as gradient dispersion and noise through innovative attention mechanisms [50].

In addition to breast and other common cancers, these advancements have also had a profound impact on the RCC analysis. For instance, Panwoon et al. [51] identified novel biomarkers that differentiate ccRCC from non-ccRCC subtypes (pRCC and chRCC) using bioinformatics and ML techniques. Their work achieved a remarkable accuracy of 98.89%, leveraging gene expression profiles and validating these findings through IHC techniques. Noureddine et al. [52] expanded on this by combining multiplexed IHC with DL to map the tumor microenvironment, offering deeper insights into immune-tumor cell interactions and the complex heterogeneity within tumors. Acosta et al. [53] also made significant contributions by applying DL models to infer genetic heterogeneity within ccRCC, successfully predicting mutation statuses and correlating these findings with critical clinical outcomes, such as disease-specific survival. These efforts underline the transformative potential of ML and DL technologies in improving diagnostic accuracy and enhancing our understanding of the biological complexity within RCC.

Chapter 3

Data

Our work integrates two crucial types of data to address the RCC classification task: Haematoxylin and Eosin (H&E) and Haematoxylin, Eosin and Natual Saffron (HES) stained WSI and quantified IHC profile data. The rationale for this dual approach lies in the complementary nature of these data sources. As discussed earlier, microscopic morphology plays a foundational role in diagnosing RCC subtypes, with distinct cellular patterns serving as the basis for subtype identification. WSIs allow for a detailed visual examination of these morphological features across varying magnifications, forming the primary mode of analysis. However, in complex or ambiguous cases where morphological assessment alone is insufficient, IHC analysis becomes crucial. By quantifying specific molecular markers, IHC provides additional insights that enhance diagnostic accuracy and assist in differentiating RCC subtypes. This sequence ensures a more robust and comprehensive classification model, addressing both the visual and molecular dimensions of RCC diagnosis. In the subsequent sections, we will provide more detailed information on the different types of data involved in our study, their definitions, characteristics, and the critical role they play in enhancing the accuracy and reliability of our model.

3.1 Tissue Staining: Purpose and Mechanisms

Tissue staining is a fundamental technique in histology used to enhance the visibility of tissue components under a microscope. Tissues are inherently transparent, making it challenging to differentiate between various cellular structures. Staining provides contrast by selectively binding to specific molecules or structures within the cells and tissues, thereby making these features more discernible. Different staining methods exploit chemical interactions between the stain and tissue, allowing researchers and pathologists to study tissue morphology, identify disease states, or assess the distribution of certain biomolecules.

3.1.1 Staining Methods

There are numerous staining methods available in histology, each designed to highlight specific cellular structures or molecular features depending on the diagnostic or research needs. Below, we explore three key staining techniques that we used during our study.

Hematoxylin and Eosin (H&E) Staining

Hematoxylin and eosin (H&E) staining is the most widely used histological stain, providing a basic but detailed overview of tissue architecture. Hematoxylin binds to acidic components in the cell, such as nucleic acids, and stains the nuclei dark blue or purple. Eosin, an acidic dye, binds to basic components, such as cytoplasmic proteins, staining them pink. This dual staining approach highlights the general layout of tissues, with nuclei standing out from the surrounding cytoplasm and extracellular matrix. H&E is highly versatile and offers insight into both normal tissue organization and pathological changes such as inflammation or cancer. In RCC, H&E staining is crucial for examining cellular morphology and identifying key features such as clear cells and papillary structures, which are essential for RCC subtype classification.

Hematoxylin, Eosin and Natural Saffron (HES) Staining

Hematoxylin, Eosin and Natural Saffron (HES) staining is an enhancement of the H&E stain by incorporating saffron, which imparts a yellow hue to collagen fibers. This triple-staining method offers greater specificity, particularly for connective tissues, as saffron highlights extracellular matrix components more effectively than eosin alone. The HES stain is especially useful in distinguishing between different types of tissues or identifying the extent of fibrosis in pathological samples. Nuclei are stained blue or purple (Hematoxylin), cytoplasm and muscle fibers are pink (Eosin), and collagen or connective tissues appear yellow (Saffron). In RCC, HES staining can help highlight the degree of fibrosis and changes in the extracellular matrix, which are important for assessing tumor progression and stromal response.

Immunohistochemistry (IHC) Staining

Immunohistochemistry (IHC) is a powerful technique used to detect specific antigens in tissues using antibodies. This method differs from H&E and HES as it is based on antigen-antibody interactions rather than simple chemical dyes. In IHC, an antibody specific to a target molecule, such as a protein, binds to its antigen within the tissue. A secondary antibody, conjugated to a detection system like an enzyme or fluorescent marker, binds to the primary antibody, making the antigen visible

under a microscope. IHC is widely used in diagnostic pathology, especially in cancer, to identify molecular markers such as hormone receptors or oncogenes, offering highly specific information about the tissue's molecular composition. In RCC, IHC plays a important role in detecting markers such as Carbonic Anhydrase IX (CAIX), Vimentin, and CD10, which aid in the accurate diagnosis, subclassification, and prognostic assessment of RCC subtypes.

3.1.2 Staining Procedure

The process of tissue staining is a multi-step procedure designed to make cellular structures within tissue samples visible under a microscope by selectively coloring specific components. This process leads to identify key features of cells and tissues, such as nuclei, cytoplasm, and extracellular matrix, and to diagnose diseases like cancer. Here is the step-by-step process involved in tissue staining, which ensures that tissue structures are preserved, sectioned, and stained to highlight their microscopic details:

1. **Tissue Fixation:** Before staining can begin, tissue samples must be preserved, typically by fixation. Fixation stabilizes the tissue, preventing degradation and maintaining the structure of cells and molecules. Formalin (formaldehyde solution) is commonly used for this purpose as it cross-links proteins and preserves tissue morphology.

2. **Tissue Processing and Embedding:** Fixed tissue samples are processed by dehydrating them in a series of alcohol baths, followed by clearing with a solvent like xylene, which removes any remaining water and prepares the tissue for embedding. The tissue is then embedded in paraffin wax, which hardens to provide support for thin sectioning.

3. **Sectioning:** After embedding, thin slices of the tissue (typically 3-5 micrometers thick) are cut using a microtome. These thin sections allow for the staining to effectively penetrate and highlight cellular structures. The sections are then placed on microscope slides.

4. **Deparaffinization and Rehydration**: The tissue sections are deparaffinized by immersing them in xylene or a similar solvent, followed by rehydration through graded alcohols. This step is crucial because staining agents, which are often water-based, cannot interact with tissue still embedded in paraffin.

5. **Application of Stain:** Once the tissue is rehydrated, the chosen stain is applied. For example, in H&E staining, hematoxylin is applied first to stain cell nuclei, followed by eosin to stain the cytoplasm and extracellular structures. Each stain is applied for a specific amount of time to achieve optimal contrast.

6. **Differentiation and Washing**: After the stain is applied, excess stain is washed away using water or alcohol. In some cases, differentiation steps are required, where a solution is used to selectively remove excess stain from certain areas of the tissue, enhancing contrast between structures.

7. **Dehydration and Mounting:** Once staining is complete, the tissue sections are dehydrated again using graded alcohols, followed by a clearing agent such as xylene. Finally, a cover slip is placed over the tissue with a mounting medium to protect the stained section and allow for long-term preservation.

This general staining process applies to most histological stains, though specific stains, such as IHC or special stains like periodic acid-Schiff (PAS), may have additional or modified steps based on the nature of the stain and the target tissue structures. Figure 3.1 shows an example of H&E stained tissue samples on glass slides.

Figure 3.1: H&E stained tissue samples on glass slides. [54]

3.2 Whole Slide Image

As discussed earlier, in pathology, the analysis of tissue begins by placing a thin section of tissue on a glass slide. The tissue is then stained to highlight different

cellular components, making it easier for pathologists to observe under a microscope. Traditionally, these glass slides have been used for diagnosis, but this method has several limitations. It is time-consuming, requires large physical storage space, has a degradation period and more importantly makes remote collaboration difficult.

Whole Slide Imaging (WSI) is a modern digital alternative that addresses many of these issues. WSI also improves the precision of diagnoses by offering high repeatability, especially in large-scale studies. However, it comes with its own challenges. It can be expensive, and some types of slides, like those used for cytology, are difficult to scan. Additionally, WSI systems may not handle large volumes of slides efficiently [55]. Despite these limitations, WSI is becoming more popular in pathology, especially with the help of emerging AI tools designed to improve diagnostic accuracy and workflow [56]. As technology advances, WSI is expected to replace traditional slides in many aspects of pathology.

WSIs could be extremely large digital files and high resolution images, often comprising millions of pixels. For instance, when scanned at 40x magnification, they can achieve a resolution of 0.25 micrometers per pixel. As a result, they require significant storage space, with uncompressed files occupying around 48 megabytes for every square millimeter of tissue [57]. To efficiently manage and navigate this vast amount of data, WSIs are structured in a multi-resolution pyramid format. This pyramid begins with a full-slide thumbnail at the top, providing a low-resolution overview of the entire tissue sample. As one moves down the pyramid, each level offers increasingly higher resolutions, revealing more detailed information like tissue architecture, cellular structures, and nuclei as shown in Figure 3.2. This hierarchical organization allows users to smoothly zoom in and out during analysis, accessing only specific patches at the required magnification levels rather than loading the entire high-resolution image, which often results in running our of memory in usual systems and resources.

To successfully obtain WSIs, several key steps must be carefully integrated into the practical workflow:

1. **Slide Conditioning and Preparation:**
    - Proper preparation of the slide is essential. The sample must be correctly stained and mounted with high-quality coverslips to ensure clear imaging.
    - Ensure the tissue is flat and fully covers the region of interest, as uneven samples can lead to focusing issues during scanning [55].
2. **Image Scanning:**
    - A specialized WSI scanner captures the slide at multiple magnifications (e.g., 20x, 40x). It is important to choose an optimal scanning resolution for diagnostic purposes [58].

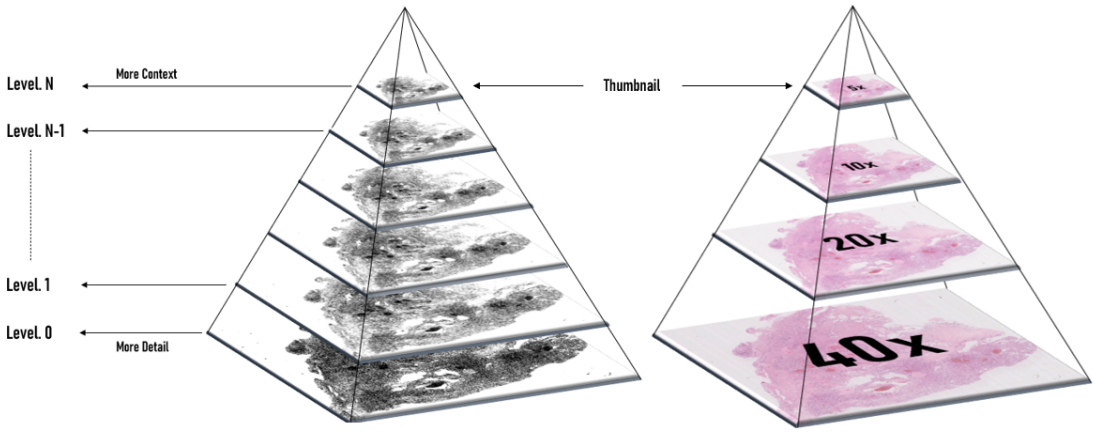


Figure 3.2: Pyramid representation of whole-slide images (WSIs) showing different resolution levels. Higher levels offer more context with less detail, while lower levels (e.g., 40x) provide higher magnification for finer details.

- Scanning involves capturing numerous image tiles, which must be accurately positioned and focused [59].

3. **Image Assembly:**
   - The individual image tiles from the scanner are algorithmically stitched together to form a continuous whole-slide image.
   - Ensuring minimal distortion and precise alignment of tiles is critical to maintaining the diagnostic value of the image [60].
4. **Verification and Quality Control:**
   - Quality control is essential for WSI to ensure there are no artifacts, stitching errors, or image distortions.
   - A review process is often performed by pathologists or automated software to verify the quality of the digitized slide before it can be used for diagnostics [61]. Figure 3.3 shows some WSI examples in which they have been failed in terms of quality control.

3.2.1 Softwares and Libraries

There are a variety of tools available for visualizing, editing, and annotating WSIs, each offering unique capabilities for handling the large and complex datasets typical

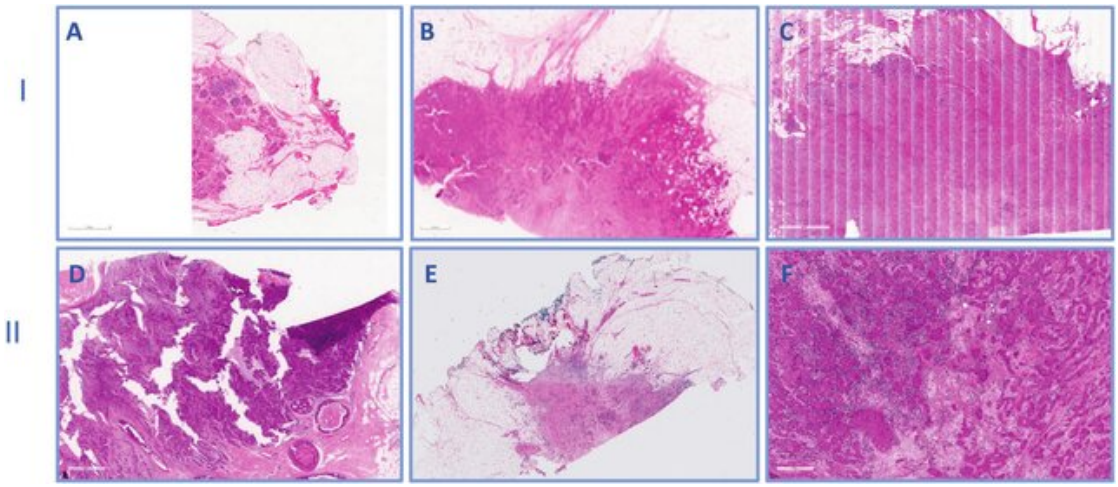


Figure 3.3: Examples of failed WSIs considering quality control. A) Incomplete slide scanning, B) Out of Focus image, C) Improper line stitching. D) Thick sections with tissue cracking and folding, E) Uneven H&E Stain distribution, F) Air bubbles on slide [62]

in digital pathology. Some of the widely used tools include ASAP (Automated Slide Analysis Platform), OpenSlide, QuPath, PathoZoom, and SlideRunner, among others. These platforms allow researchers and clinicians to explore histopathological images in detail, make annotations, and even integrate image data into machine learning workflows for advanced analysis. Each tool has its strengths, ranging from basic visualization to advanced image processing and analysis, catering to diverse research needs.

Among these tools, ASAP stands out as an open-source platform specifically designed for the visualization, annotation, and automatic analysis of WSIs. Built on top of several powerful open-source libraries such as OpenSlide, Qt, and OpenCV, ASAP offers a modular architecture where its key components—slide input/output, image processing, and a viewer—can be used independently or together. In addition to its powerful image handling capabilities, ASAP allows users to write multi-resolution tiled TIFF files for ARGB, RGB, Indexed, and monochrome images, with support for different data types such as float. Its Python bindings are particularly valuable for researchers, enabling consistent access to multi-resolution WSIs as Numpy arrays, which is crucial for integrating image data into machine learning models and custom image analysis pipelines. ASAP also provides basic image primitives, like "patches," which can be processed through its connection with OpenCV to perform tasks such as feature extraction and image enhancement.

ASAP's Qt-based viewer is one of its standout features, offering fast and smooth navigation through WSIs, even at high magnification levels. This capability allows

users to examine fine details in large images without running out of memory, making it highly efficient for working with massive datasets. Researchers can zoom in on specific regions of interest seamlessly, ensuring an uninterrupted analysis workflow even with large, complex slides. It also includes robust annotation tools such as point, polygonal, and spline annotations, which allow precise marking of regions of interest. These annotations are saved in a simple, human-readable XML format, ensuring compatibility with other software tools and simplifying data sharing and integration. The platform is also highly extensible, supporting plugins for tools, filters, extensions, and additional file formats, which enhances its flexibility for diverse research applications. ASAP further incorporates on-the-fly image processing during viewing, such as color deconvolution and nuclei detection, making it a powerful tool for real-time analysis.

At the core of ASAP's image reading capabilities is OpenSlide, a foundational open-source C library. OpenSlide, currently at version 4.0.0 (released in October 2023), supports a wide range of slide formats from vendors like Aperio (`.svs, .tif`), DICOM (`.dcm`), Hamamatsu (`.vms, .vmu, .ndpi`), Leica (`.scn`), MIRAX (`.mrxs`), Philips (`.tiff`), Sakura (`.svslide`), Trestle (`.tif`), Ventana (`.bif, .tif`), Zeiss (`.czi`), and generic tiled TIFF files (`.tif`). This extensive format compatibility makes OpenSlide a valuable tool and library for researchers working with diverse slide data across the digital pathology landscape. In addition to its core functionality, OpenSlide offers bindings for Python and Java, which further extend its usability in the research community. The Python binding includes a Deep Zoom generator and a simple web-based viewer, enabling remote viewing and analysis of WSIs.

Together, ASAP and OpenSlide provide a comprehensive toolkit for working with WSIs. While OpenSlide offers a technical backbone for reading and accessing slide data across numerous formats, ASAP provides a high-level, user-friendly platform for visualization, annotation, and image processing. This combination allows researchers and clinicians to conduct manual annotations, perform sophisticated image analysis, and integrate the data into machine learning pipelines, all while efficiently handling the massive datasets that are common in digital pathology without running into memory limitations.

3.3 Our Dataset

In this study, we have utilized multiple datasets from training to test in order to ensure the generalizability and robustness of our findings. By incorporating a diverse range of data sources, we aim to validate our approach across different conditions and variations. Table 3.1 reports the total number of patients presents for different centers in our study.

	Subtypes			
Cohort	**ccRCC**	**pRCC**	**chRCC**	**Oncocytoma**
Nice A	64 (141)	30 (62)	15 (28)	15 (28)
Nice B	4 (8)	3 (6)	2 (4)	7 (14)
Lyon	5 (15)	5 (15)	5 (15)	5 (15)
Paris Cochin	5 (15)	5 (15)	5 (15)	5 (15)

Table 3.1: Total number of patients (slides) for different centers and cohorts.

3.3.1 Nice Cohort

The main dataset utilized in this study originates from the Central Laboratory of Pathological Anatomy and Cytology (LCAP) of Centre Hospitalier Universitaire de Nice (Hôpital Pasteur). This center is a renowned medical institution in France, providing high-quality care and conducting advanced medical research. The dataset acquired from this hospital comprises a collection of H&E stained WSIs plus the IHC profile quantifications for four different RCC subtypes including: ccRCC, pRCC, chRCC and oncocytoma. All the slides where scanned using Aperio AT2 (Leica Biosystems, Wetzlar, Germany) scanner. These data were annotated by expert pathologists at the Pathology Department of this center, which have been instrumental in the training and validation of our models. The IHC profile quantification includes the intensity level and percentage of each biomarker presence. This dataset is especially valuable due to its diversity in patient case complexity, enabling us to test the generalizability of our approach in real-world clinical scenarios.

WSIs Annotations

For all patients in this cohort, we have corresponding annotations, or regions of interest (ROIs), that delineate homogeneous tissue areas. These annotations were generated using two distinct methods. In some cases, only homogeneous cancer regions were cropped and saved as independent WSIs, which we will refer to as WSI-ROIs. For other patients, annotations were made for various tissue types, including cancer regions, non-cancerous areas such as necrosis and fibrosis, and normal tissue. However, it should be noted that not all tissue types were annotated for every patient. These annotations were initially marked on WSI thumbnails and subsequently converted into XML format by defining splines and polygons that correspond to specific tissue types, providing greater detail at higher magnifications. We will refer to this type of annotation as XML-ROIs in the literature. An illustrative example of this annotation format can be found in Appendix A.1. Figure 3.4 shows examples of both XML-ROIs and WSI-ROIs, along with the corresponding tissue types.

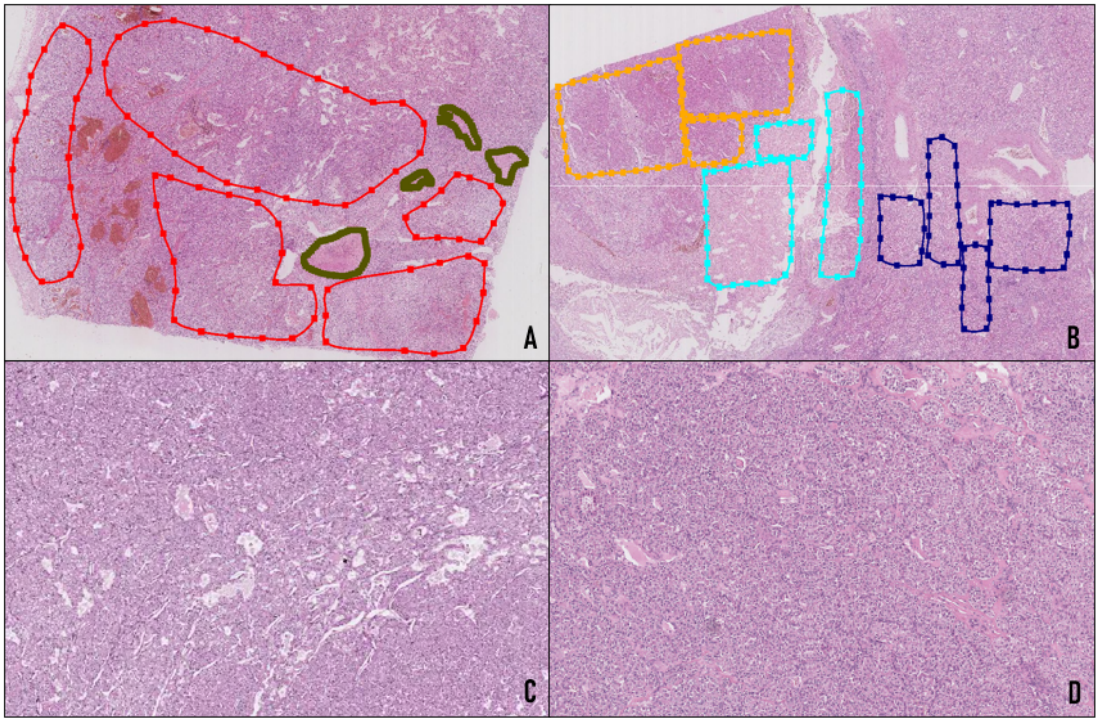


Figure 3.4: Examples of region of interest (ROI) annotations in whole slide images (WSIs). Panels A and B represent XML-ROIs, where different tissue types are annotated with splines at high magnification. In these images, red denotes clear cell renal cell carcinoma (ccRCC), green denotes fiber, orange represents papillary renal cell carcinoma (pRCC), cyan indicates necrotic regions, and blue highlights normal tissue. Panels C and D display WSI-ROIs, where homogeneous tumor regions are shown, with panel C representing chromophobe renal cell carcinoma (chRCC) and panel D showing oncocytoma.

Immunohistochemistry Profile Quantification

In this study, we applied a standard method for analyzing immunohistochemistry (IHC) profiles, which involves assessing both the intensity of staining and the percentage of positively stained cells, as described by the method in the literature. The intensity of staining is scored on a scale from 0 to 3, with 0 representing no staining, 1 indicating weak staining, 2 for moderate staining, and 3 for strong staining. Simultaneously, the percentage of positive cells is evaluated in intervals, ranging from 0% to 100%, and categorized into quantizied groups such as <5%, 5%–25%, 26%–50%, 51%–75%, and >75%. In our dataset, the IHC quantification includes key biomarkers such as Cluster of Differentiation 10 (CD10), Paired Box Gene 8 (PAX8), Alpha-methylacyl-CoA Racemase (P504), Cytokeratin 7 (CK7),

E-cadherin (ECAD), Carbonic anhydrase IX (CaIX), and Vimentin (VIM). This comprehensive profiling of both the presence and intensity of biomarkers provides critical insights into tumor classification and behavior, allowing for more precise distinctions between the various RCC subtypes.

3.3.2 External Cohort: Lyon and Paris Cochin Cohorts

In addition to the previous cohort, to validate our findings, we obtained additional slides from the Department of Pathology at Hospices Civils de Lyon, Lyon, France, and the Department of Pathology at Cochin Hospital, Paris, France. These external datasets include These two dataset comprises WSIs stained with HES from all RCC subtypes introduced in our study. The slides processed using proprietary automated staining protocols unique to each of the laboratories.

Chapter 4

Methodology

In this chapter, we present the methodology used in this research, outlining the key steps and processes undertaken to achieve the study's objectives. The methodology encompasses the data collection, preprocessing techniques, model development, and evaluation strategies applied throughout the study. A systematic approach was followed to ensure the integrity and reliability of the results, starting from initial data handling to the final stages of analysis. Each step in the methodology was designed to address the specific research questions while adhering to established scientific principles. This section also provides an overview of the tools, algorithms, and validation methods used to ensure accuracy and robustness in the findings.

In this study, our objective is to closely replicate the clinical approach used for RCC subtype classification. We start with a morphological analysis, followed by IHC analysis if the initial results are uncertain. Figure 4.1 shows the overall workflow of our model, highlighting its various stages. Our hybrid model first performs the morphological analysis, providing both the predicted subtype and an associated confidence score. If this confidence score falls below a predefined uncertainty threshold, the model proceeds to IHC analysis to either confirm or adjust the final subtype.

The initial phase to have this hybrid model involves training the models and optimizing their hyperparameters. Once the optimal model is achieved, it is employed for testing. In the following sections, we will explore each component of the workflow in detail, explaining their individual roles and functions.

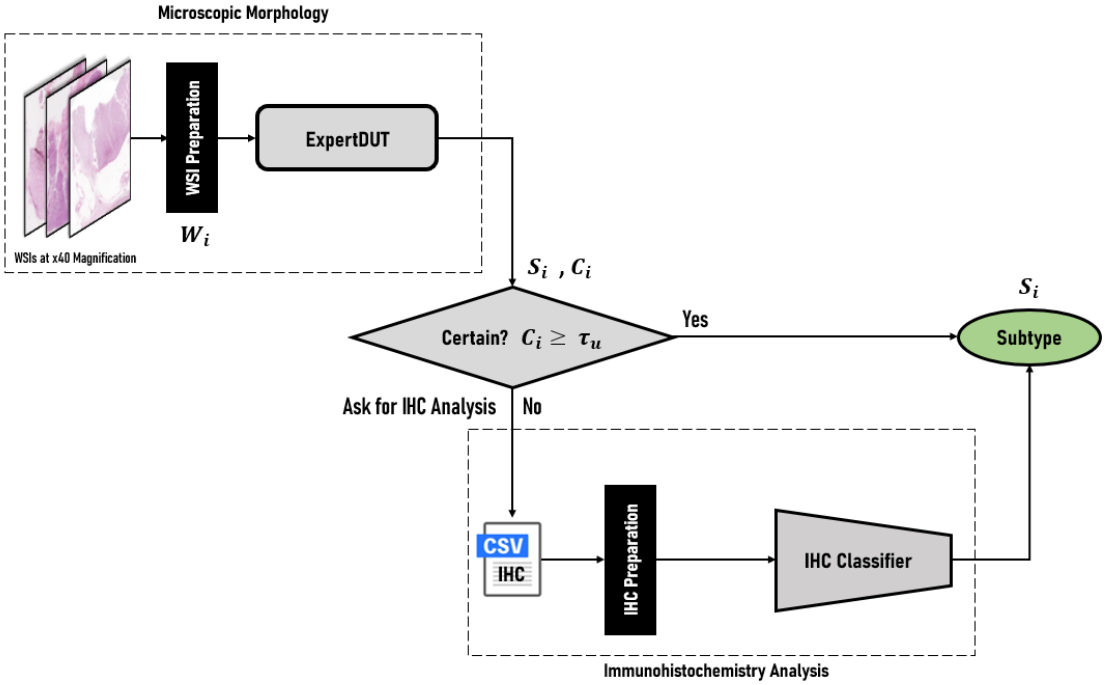


Figure 4.1: General diagram of the proposed hybrid model to classify RCC subtypes

4.1 Data Preprocessing

4.1.1 Whole Slide Images

The first critical step in this study is data processing, which is fundamental for both training and testing the model. Preparing the input data properly ensures robust model performance. We begin by processing whole slide images (WSIs) during the training phase. As we discussed earlier, we are employing a supervised learning approach and our dataset (see Section. 3.3.1) contains two types of annotations: XML-based regions of interest (XML-ROIs) and WSI-based regions of interest (WSI-ROIs).

As illustrated in Figure 4.2, slides associated with each annotation type are directed through distinct processing pipelines. For WSIs with XML-ROIs, the corresponding XML annotation file is initially parsed. Patches are then extracted from the highest magnification level of the WSI (×40 in this study) with a patch size of 1000x1000 pixels. Each patch is cross-referenced with the XML annotations to determine if it falls within any annotated region, defined by splines or polygons. If a patch lies within an annotated region, it is labeled accordingly and passed to the next processing stage. Otherwise, it is discarded.

In contrast, WSIs with WSI-ROIs annotations bypass the XML parsing step, and patches are directly extracted from the ROI. Subsequently, all patches, regardless of the annotation type, undergo a quality check to filter out background or artifact patches, i.e., those lacking sufficient tissue content to provide meaningful information for the model

During the testing phase, as shown in Figure 4.3, the same procedure is followed: patches are first extracted, and then background patches are removed in subsequent steps.

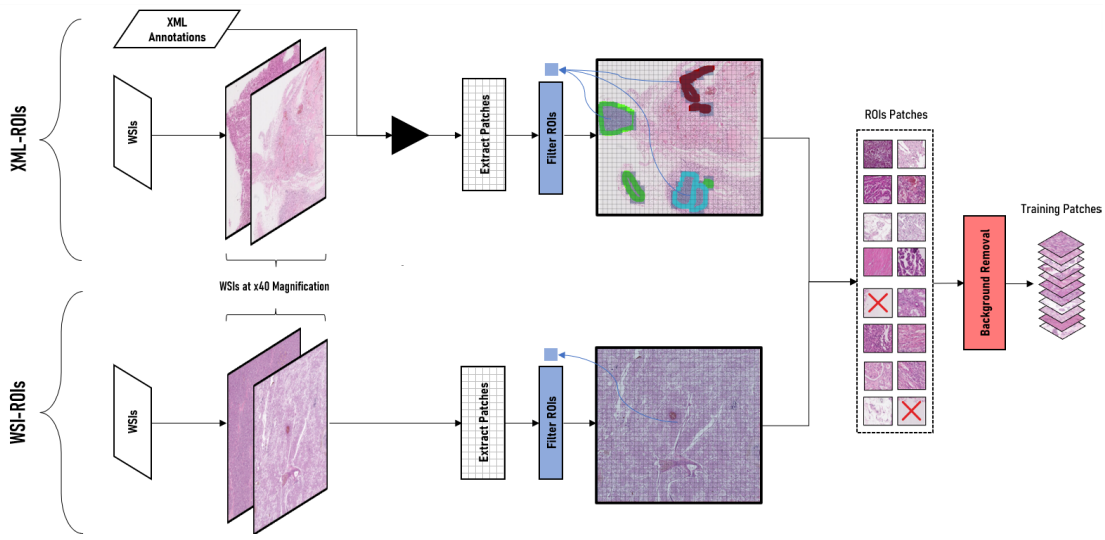


Figure 4.2: Training data preparation diagram

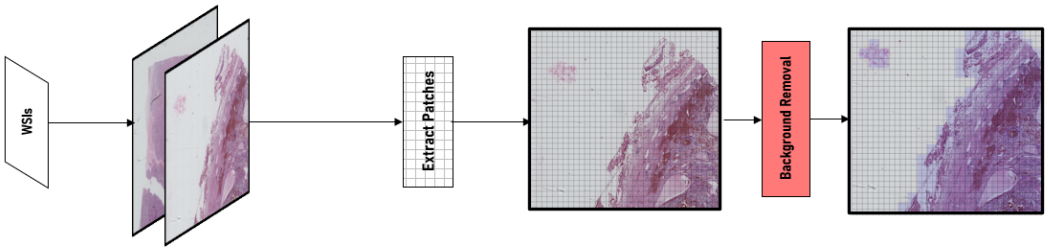


Figure 4.3: Test data preparation diagram

Background Removal

In order to identify the patches that are backgrounds, most of the previous models like Ponzio et al. [18] used a simple approach. The background removal has

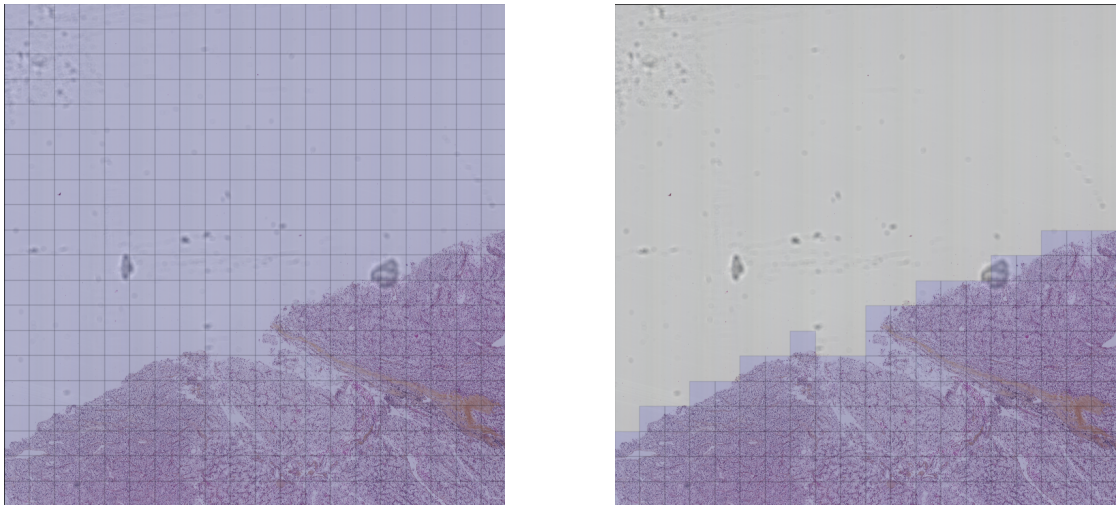
been carried out by simply defining an average value threshold of patch pixels to eliminate empty areas, namely where the tissue is almost absent. The corresponding threshold on the mean pixel value was empirically set on the training set. However we introduced a new method to identify both background and also the artifacts presents into background patches that could be missconsidered as tissue and non-background patches by previous methods.

Our method utilizes a gradient-based approach to identify background patches, which is highly effective in distinguishing between the well-structured tissue regions and the more homogeneous background in histopathological images. By applying Sobel operators to the grayscale version of the image, we compute the gradient magnitude that captures changes in pixel intensity, particularly around the tissue boundaries where significant structural variations occur. This gradient information, combined with thresholds for gradient magnitude and variance, allows for robust detection of background areas, even in complex histopathological images where there are high variability in texture, staining, and contrast. The use of gradient-based methods in this context is beneficial as it capitalizes on the sharp transitions between tissue and background, especially along the contours of cells and tissue structures [63]. By setting thresholds on the mean gradient and its variance, our method effectively suppresses low-texture areas, which are often indicative of background regions, while preserving key tissue details for further analysis. Additionally, incorporating a dark pixel threshold helps handle regions with low illumination or uneven staining, which is a common issue in histopathology [64].

The primary advantage of this gradient-based approach for histopathology images is its computational efficiency and adaptability to various staining protocols and imaging conditions. It requires no prior training or complex models, making it suitable for real-time processing of large-scale histopathological datasets. Studies have shown that combining gradient-based detection with intensity-based criteria improves the accuracy of background identification in medical images, leading to better downstream analysis, such as tissue segmentation or classification [65]. Furthermore, its ability to adapt to different types of histological stains, H&E, is crucial for versatile histopathological analysis [66]. Figure. 4.4 illustrates how the mean method fails to accurately detect background patches, primarily due to contrast variations that fall outside the threshold settings applied during training.

Patches Imbalance Handling

After completing the background removal process, our patches are now primed and ready for the next step. As discussed earlier in Section 3.3.1, there is an imbalance in the number of patches in our dataset. Training the model on this imbalanced data could lead to biased learning, which in turn results in poor model performance, especially for the minority class, as suggested by multiple



(a) Mean Method **(b)** Gradient Method

Figure 4.4: Comparison of Mean and Gradient Methods for Background Detection

studies on class imbalance and its effects on model accuracy [67]. Techniques such as oversampling and undersampling are commonly used to address this issue. Oversampling techniques like SMOTE generate synthetic samples to balance the classes, while undersampling reduces the majority class size to match that of the minority class. Since our dataset contains a sufficient number of patches, we opted for the undersampling approach to mitigate the imbalance. In this method, patches are chosen randomly from the other classes, ensuring that the selected samples maintain the integrity of the minority class distribution while preventing overfitting on the majority class. Studies have shown that undersampling, especially when used in combination with appropriate classifiers, can yield improved model performance in cases where data abundance allows such a strategy [68].

4.1.2 Immunohistochemical Profile

In this section, we will outline the pipeline for preparing data for the immunohistochemical subtype classification model. Our dataset consists of raw data derived from pathologists' interpretations, which contains several issues that must be addressed before it can be effectively used for training.

Data Loading and Cleaning

The data loading step involves importing subtype-specific IHC data from '.csv' (Comma-Separated Values) files, ensuring that only the relevant features needed for further analysis are included. After loading, the data undergoes a cleaning

process to identify and address any invalid entries. These invalid values can include missing data, out-of-range entries, or uncertainties in pathologist annotations, such as ambiguous values like '1 (faible) ou 0' and '??' found in some patients' biomarker data. These are replaced with NaN (Not a Number) to signal missing or uncertain values in the dataset. Rows with excessive missing values are removed, but the threshold for dropping data is carefully selected to retain as much useful information as possible. This data cleaning step is essential to ensure that the machine learning model is not affected by noise or errors, which could lead to poor performance or inaccurate results. Properly cleaned data allows for more reliable analysis and model training [69].

Feature Extraction and Enhancement

The next step involves converting the features from string format into suitable numeric data types. After this conversion, we need to decide whether to retain the existing features or explore options for enhancing or manipulating them. Several methods have been proposed to improve our analysis and insights.

The `baseline` method preserves the dataset in its original form, providing a baseline for comparison against enhanced versions. The `percentage_only` method isolates features representing the percentage of stained cells by various biomarkers, excluding those related to intensity. This method emphasizes the percentage of positive staining as a standalone metric, which is often critical in assessing the distribution of biomarker expression. Conversely, the `intensity_only` method focuses solely on intensity-related features, capturing the strength of the staining, which reflects the level of biomarker expression in tissue samples.

The `combined` method takes a more comprehensive approach by combining both the percentage of stained cells and the intensity into a single enhanced feature as shown in Equation 4.1:

$$\text{Combined Feature} = \text{Percentage of Stained Cells} \times 10^{\text{Intensity}} \tag{4.1}$$

This transformation effectively integrates both aspects of IHC data—prevalence and intensity of staining—into one cohesive feature. By using Equation 4.1, we can represent both dimensions of each biomarker expression in a single feature, allowing the model to better capture complex relationships across different subtypes.

Lastly, the `hscore` method creates a feature that mimics the H-score system, which is widely used in pathology to assess biomarker expression.

$$\text{H-Score} = \text{Percentage of Stained Cells} \times \text{Intensity} \tag{4.2}$$

This semi-quantitative scoring system, as shown in Equation 4.2, combines the proportion of positive cells and staining intensity into a single feature for each

biomarker, offering a clinically relevant representation of biomarker expression [70]. Implementing this enhancement ensures that the model benefits from a well-established scoring method commonly applied in clinical practice for IHC grading.

Overall, these feature enhancement methods offer flexibility in processing IHC data, allowing machine learning models to learn from different dimensions of biomarker expression efficiently. From raw percentages and intensity values to sophisticated combinations that reflect real-world clinical scoring systems like the H-score, these techniques significantly enhance our ability to analyze and interpret IHC data [70].

Data Imputation

The first method employed was the KNNImputer, which estimates missing values based on the values of the three nearest neighbors (with $k = 3$). Specifically, missing values are imputed by averaging the corresponding feature values from these nearest neighbors. We also utilized simpler imputation approaches that fill in missing values using a mean or median strategy for each feature. These imputation techniques are essential for maintaining the integrity of the dataset, especially when missing data is widespread [71]. The choice of imputation method depends on the distribution of the data and the amount of missing information. We will later discuss in Section 5.2 about how we have chosen the proper method to use.

Data Scaling

Once the missing values are addressed, the dataset undergoes scaling. StandardScaler is applied to standardize the features, ensuring that each feature has a mean of zero and a unit variance. This process is represented by Equation 4.3, which transforms the original feature values. Scaling is particularly important when working with machine learning algorithms that are sensitive to the scale of input features. [72].

$$z = \frac{x - \mu}{\sigma} \tag{4.3}$$

where:

- z is the standardized value,
- x is the original value of the feature,
- μ is the mean of the feature across the dataset, and
- σ is the standard deviation of the feature.

By applying Equation 4.3, each feature is transformed so that it has a mean of zero and a standard deviation of one, making it easier for machine learning algorithms to handle features on different scales and leading to improved model performance.

Oversampling/Undersampling

To handle imbalanced data, various resampling techniques are applied to ensure that the model can generalize well across all classes, particularly the minority class. One common technique is SMOTE (Synthetic Minority Over-sampling Technique), which generates synthetic examples by interpolating between existing minority class samples. This technique helps by effectively increasing the representation of the minority class without simply duplicating the existing data points, thus avoiding overfitting. SMOTE is particularly useful when the dataset is highly imbalanced, as it ensures that the classifier has enough minority class examples to learn from during training [73].

In addition to SMOTE, TomekLinks is another technique that can be used to address imbalanced data. TomekLinks identifies pairs of samples from different classes that are very close to each other and removes the majority class sample from each pair. This process helps to create a cleaner decision boundary by removing ambiguous samples that might otherwise confuse the classifier. By removing noisy or borderline majority class examples, TomekLinks ensures that the model focuses on clearer examples, thus improving overall classification performance [74].

SMOTEENN is a hybrid approach that combines both SMOTE and Edited Nearest Neighbors (ENN). After SMOTE oversampling is applied to increase the minority class representation, ENN is used to clean the dataset by removing noisy or misclassified examples, primarily from the majority class. This combination allows for not only balancing the dataset but also improving the quality of the training data by reducing the influence of noisy samples. This dual approach of oversampling and cleaning the data helps to improve model performance by allowing better discrimination between classes [75].

These resampling techniques are vital for ensuring that the machine learning model does not become biased toward the majority class, which is a common issue in imbalanced datasets. By balancing the classes, these techniques improve the model's ability to correctly classify minority class samples, leading to more robust and fair predictions. In Section 5.2, we will dive deeper into how we selected the optimal method from the these options.

4.2 ExpertDeepUncertainTree (ExpertDUT)

In this study, we introduce a novel deep learning architecture termed *ExpertDeepUncertainTree (ExpertDUT)*, developed as a slight modification of the ExpertDeepTree (ExpertDT) architecture from Ponzio et al. [18]. It was specifically designed for the subtyping of RCC by incorporating both CNNs and pathologists' expert-based knowledge. This hybrid model enhances traditional CNN performance by embedding pathologist expertise into the model's decision-making process and incorporates uncertainty-based refinement mechanisms, making it particularly suitable for resolving complex classification tasks. In forward, we will discuss different aspects of the architecture and the key modifications with respect to ExpertDT.

4.2.1 Overall Structure

ExpertDUT is composed of a tree-style architecture as shown in 4.5, with binary CNN classifiers placed at different levels of the tree (root, node, and leaves) to progressively differentiate RCC subtypes from WSI inputs from each patient ($WSI(i)$). The architecture has been refined to handle uncertainty in classification, particularly for cases where the model cannot confidently differentiate between subtypes. The model provides the corresponding subtype for the patient, denoted as S_i, along with the model's confidence score for that patient, represented as $C(i)$. This uncertainty will later help the the model to identify cases that could be potentially errors in the mophological analysis. The organization of the tree is informed by pathologist-defined expertise, structuring the decision hierarchy based on histopathological insights into RCC subtypes. This approach is designed to address the challenge of differentiating between RCC subtypes that exhibit overlapping morphological characteristics, such as RCC subtypes.

4.2.2 Binary Classifier Configuration

Each level within ExpertDUT is composed of binary classifiers, where each classifier is tasked with distinguishing between two specific classes. The CNN models selected for this purpose are based on the VGG16 architecture, pre-trained on the ImageNet dataset [76]. This architecture was chosen as the backbone due to its demonstrated effectiveness in various medical image analysis tasks, particularly in cases involving small datasets and complex classification problems. Pre-trained models have been shown to generalize well, even with limited training data, such as in the case of RCC samples. Ponzio et al. [18] further validated this by comparing the VGG16 binary classifier to several state-of-the-art architectures, including ResNet50 [77], ResNet101 [77], DenseNet121 [78], Inception [79], Xception [80], and ConvNeXt [81]. Their analysis, which assessed both training from scratch and transfer learning

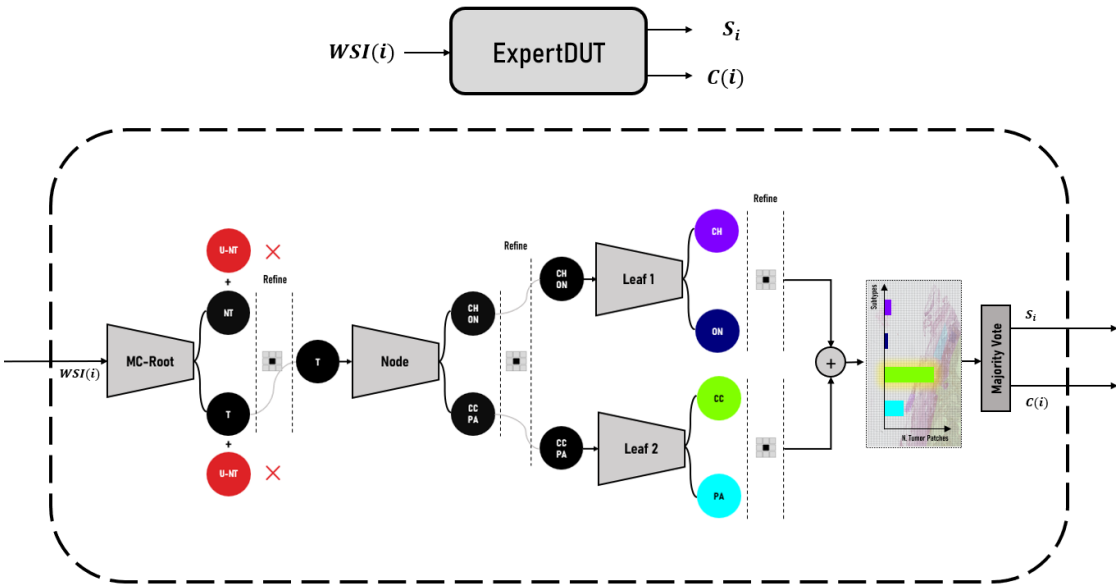


Figure 4.5: Overall structure of ExpertDUT, illustrating the process from input to final model output. NT refers to Non-Tumor, T refers to Tumor, U-NT refers to Uncertain Non-Tumor, and U-T refers to Uncertain Tumor. CC, PA, CH, and ON represent Clear-Cell, Papillary, Chromophobe, and Oncocytoma, respectively

with ImageNet weights, underscored the consistent performance gains provided by transfer learning across these models, reaffirming the value of using pre-trained networks for medical image classification. As illustrated in Table 4.1 and Figure 4.6, the architecture of the binary classifier includes multiple convolutional layers, pooling layers, and a dense layer. While the original VGG16 model accepts input images of size 224×224, our modified version processes images of size 112×112, enhancing computational efficiency without compromising performance.

4.2.3 Tree Structure

The detailed binary classification levels presented are as follows:

- **Root Level**: The root level discriminates between tumor and non-tumor regions, a fundamental initial classification step. The non-tumor regions include tissue identified as healthy or necrotic, while the tumor class contains all RCC subtypes. Building on prior work, Root-level classifier in ExpertDUT also integrates Monte Carlo sampling via Monte Carlo dropout [82], which facilitates the identification of uncertain predicted patches during the classification process. Specifically, the root classifier plays a crucial role by

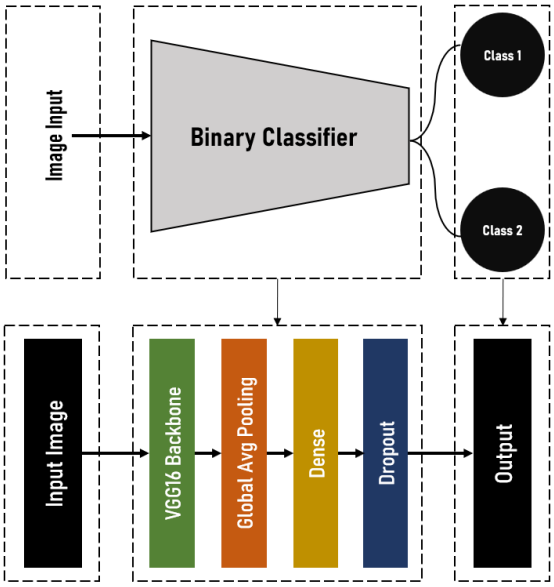


Figure 4.6: Binary classifier architecture with VGG16 backbone

discriminating between tumor and non-tumor regions, serving as a fundamental initial classification step. Any errors in this initial classification step are likely to propagate through the entire decision tree, potentially compounding inaccuracies in subsequent layers. Therefore, minimizing errors at this level is critical, as it significantly reduces the risk of misclassification downstream, enhancing the overall reliability and performance of the model in accurately distinguishing between tumor and non-tumor regions. These patches will be discarded through the whole tree even in the refinement process (See Section 4.2.4) to avoid error propagation to the leaves. Figure 4.7 presents a comparison between a Monte Carlo Root (MC-Root) and a normal Root, which was also used in the original ExpertDT implementation. Further details on this concept will be discussed later in Sections 4.2.7 and 5.1.1.

- **Node Level**: The second level (referred to as the *Node*) is designed to differentiate between the composite classes of ccRCC + papRCC and chrRCC + ONCO. This grouping, informed by expert knowledge, reflects the morphological similarities within each composite group and allows for easier discrimination.

- **Leaf Level**: The leaves of the tree perform fine-grained classification. The first leaf (*Leaf 1*) distinguishes between chRCC and Oncocytoma, while the second leaf branch (*Leaf 2*) differentiates between ccRCC and pRCC.

After each level, a refinement process is applied to reduce errors in classifier

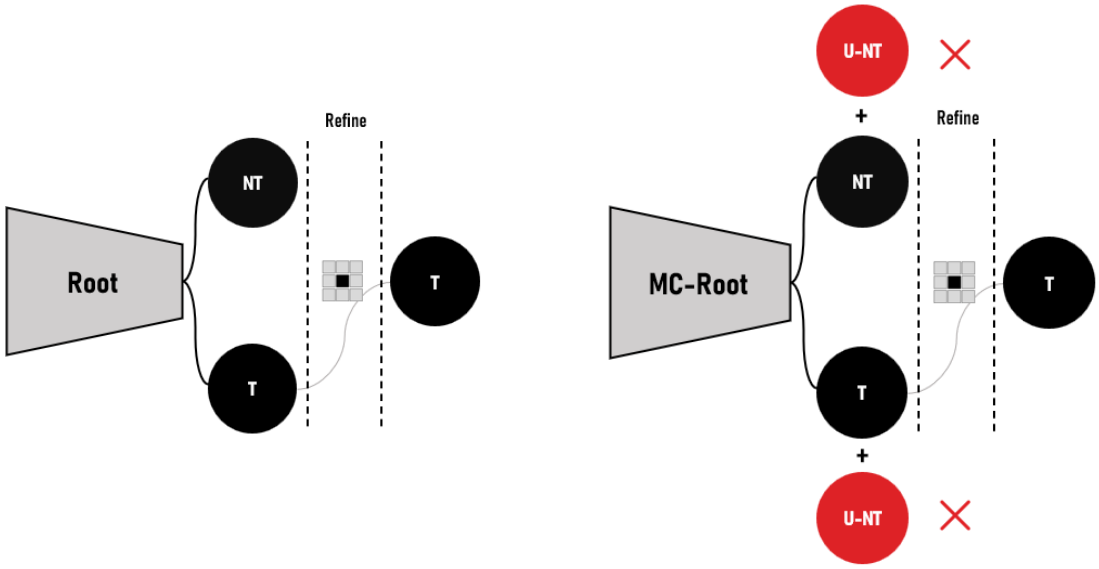
Layer	Channels	Kernel Size	Stride	Output Size
Input	3	-	-	112×112
2x Conv	64	3×3	1	112×112
	64	3×3	1	112×112
Max Pool	64	2×2	2	56×56
2x Conv	128	3×3	1	56×56
	128	3×3	1	56×56
Max Pool	128	2×2	2	28×28
3x Conv	256	3×3	1	28×28
	256	3×3	1	28×28
	256	3×3	1	28×28
Max Pool	256	2×2	2	14×14
3x Conv	512	3×3	1	14×14
	512	3×3	1	14×14
	512	3×3	1	14×14
Max Pool	512	2×2	2	7×7
3x Conv	512	3×3	1	7×7
	512	3×3	1	7×7
	512	3×3	1	7×7
Max Pool	512	2×2	2	3×3
Avg Pool	512	-	-	1×1
Dense	1024	-	-	1×1
Dropout	-	-	-	-
Output	2	-	-	1×1

Table 4.1: Overview of the layers in the binary classifier architecture. While the original VGG16 model presented with input images of size 224×224, we used input images of size 112×112

predictions across all patches of the WSI except those identified as uncertain.

4.2.4 Refine Mechanism

A key innovation in the ExpertDT architecture was the incorporation of an *refine mechanism*. The refine mechanism functions as a low-pass filter, removing noise by correcting misclassified tiles. It operates by evaluating the classification output for each tile in the context of its surrounding tiles. If a tile's classification differs from the majority of its neighboring tiles or the model registers high uncertainty, it is reassigned based on the majority label, reducing the likelihood of isolated misclassifications. ExpertDUT utilizes the exact same refinement mechanism used in ExpertDT by a filter window of 3×3, consider all the 9-connected neighborhood patches including the patch itself.



(a) Normal Root **(b)** MC-Root level

Figure 4.7: Comparison of Normal and MC-Root. Uncertain patches will be totally ignored during the next level classification.

4.2.5 Selective Pruning

ExpertDT also incorporated a *pruning strategy* at the node level. In cases where the model cannot confidently distinguish between composite classes (e.g., ccRCC + papRCC vs. chRCC + oncocytoma), based on the proportion of tiles classified into each group and the associated uncertainty scores, the model will prune the node and directly connect the root to the leaf classifiers. The authors empirically set this threshold to 30% on training set. This selective pruning mechanism ensures that the most challenging classification decisions are deferred to a more detailed classification stage, effectively mitigating the risk of misclassification in ambiguous cases. In our study, we have removed this selective pruning and feed the predicted patch-map of node level after refinement mechanism to the corresponding leaf. By analyzing the final unpruned patch-level results, we will find out how much the whole tree is confident about the final presented subtype through out the whole tree.

4.2.6 Tree Uncertainty

After obtaining the final results, our model generates a map of the whole slide image patches so called segmentation map, where each patch is classified according to its

respective RCC subtype through the whole tree. By aggregating all the patch-level results for each patient, the corresponding subtype is determined using a majority voting approach. In addition to assigning the most likely subtype, we also quantify the majority voted subtype confidence regarding its output predictions. To achieve this, we calculate the probability of the final voted subtype relative to the other subtypes for each patient i. To generalize this formula for any subtype s for patient i, the probability $C(i)$ is expressed as:

$$C(i) = \frac{T^i_{\text{majority}}}{\sum_s T^i_s} \tag{4.4}$$

where:

- T^i_{majority} is the total aggregated number of patches classified as the majority subtype for patient i across multiple slides,
- $\sum_s T^i_s$ is the total aggregated number of patches across all subtypes s, i.e. tumor regions, for patient i across multiple slides.

$C(i)$ is a probablitity bounded within the interval [0,1]. However, their specific ranges are influenced by the number of subtypes N in the classification task.

The minimum value of $C(i)$ occurs when the classifications of patches are evenly distributed among all subtypes. In this scenario, the majority subtype has the smallest possible proportion of patches, which is $\frac{1}{N}$. The maximum value is 1, occurring when all patches are classified as the same subtype. Therefore, the interval for $C(i)$ is:

$$\frac{1}{N} \leq C(i) \leq 1 \quad \text{and for } N = 4: \quad 0.25 \leq C(i) \leq 1 \tag{4.5}$$

A low $C(i)$ reflects greater disagreement among patch classifications, signaling higher uncertainty in the model's prediction. This suggests that additional clinical investigation may be necessary to ensure an accurate diagnosis. Conversely, high $C(i)$ indicates that most patches agree on the subtype, leading to lower uncertainty and higher confidence in the classification. Understanding these probability values allows for a nuanced assessment of the model's performance and helps identify cases where further diagnostic procedures might be warranted due to the model's lower confidence.

4.2.7 Monte Carlo Dropout as an Bayesian Approximation

Monte Carlo (MC) dropout is a technique used to estimate CNN binary classifiers uncertainty by applying dropout at both training and inference stages. By incorporating dropout during CNN binary classifier prediction, we approximate a

Bayesian model, allowing us to sample from the model's predictive distribution. This approach enables the model to estimate uncertainty by generating multiple stochastic forward passes for each patch, each one representing a different sampled model configuration.

Mathematically, for a given input x, the model produces T stochastic forward passes, with each pass generating a prediction $\hat{y}^{(t)}$. The mean prediction $\hat{y}$ is given by:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} \hat{y}^{(t)} \tag{4.6}$$

The model uncertainty can then be derived by computing different methods such as the variance of these predictions. A high variation (4.7) suggests significant variation in predictions across the Monte Carlo samples, indicating model uncertainty.

$$\mathrm{Var}(\hat{y}) = \frac{1}{T} \sum_{t=1}^{T} \left(\hat{y}^{(t)} - \hat{y} \right)^2 \tag{4.7}$$

This process provides a Bayesian approximation without the need for complex probabilistic inference techniques. Monte Carlo dropout was popularized by Gal and Ghahramani [82] as a practical and scalable method for incorporating Bayesian principles in deep learning models.

Measure Patch Classification Uncertainty

As we aimed before, we sought to understand the statistical behavior during different forward passes with Monte Carlo dropout in order to identify when the model makes an incorrect patch-level classification based on the generated statistics. To achieve this, we compute several methods to construct a distribution that provides information about each of correct and incorrect predictions. Several entropy-based and variation-based methods, such as Total Variance (4.7), are used to quantify the uncertainty during different forward passes.

Predictive Entropy: The predictive entropy quantifies the total uncertainty in the model's prediction.

$$H(p) = -\sum_{i=1}^{C} p_i \log_2(p_i) \tag{4.8}$$

In predictive entropy, p_i is the predicted probability for class i, and C is the total number of classes. A high predictive entropy (4.8) indicates that the model is uncertain about the class label.

Rényi Entropy: Rényi entropy generalizes the Shannon entropy by introducing a parameter α that controls the sensitivity to tail probabilities. It is defined as:

$$H_\alpha(p) = \frac{1}{1-\alpha} \log \left(\sum_{i=1}^{C} p_i^\alpha \right) \tag{4.9}$$

For $\alpha = 1$, Rényi entropy (4.9) reduces to the standard Shannon entropy. This metric helps assess uncertainty in cases where certain classes have significantly higher probabilities than others.

Mutual Information: Mutual information measures the reduction in uncertainty about one random variable given knowledge of another. In the context of Monte Carlo dropout, it is calculated as the difference between the predictive entropy and the expected entropy across the stochastic forward passes:

$$I(p) = H(p) - \mathbb{E}[H(p|\hat{y}^{(t)})] \tag{4.10}$$

Mutual information (4.10) helps in understanding how much uncertainty arises from the model's parameters rather than the input data.

Margin of Confidence: The margin of confidence measures the difference between the top two predicted class probabilities, averaged across all Monte Carlo samples:

$$\text{MoC}(p) = \frac{1}{T} \sum_{t=1}^{T} \left(p_{\text{max}}^{(t)} - p_{\text{second\_max}}^{(t)} \right) \tag{4.11}$$

This metric (4.11) is useful for assessing how confident the model is in its most probable prediction compared to the next most probable class.

Bhattacharyya Distance for Distribution Comparison

To further understand the statistical behavior of the model's predictions, we employ the Bhattacharyya distance to compare the distributions generated by different uncertainty measures, following the approach discussed by Milanes et al. [83]. The Bhattacharyya distance is useful for comparing two probability distributions by quantifying the amount of overlap between them. In this context, we compare the distributions of uncertainty scores for correct and incorrect predictions to assess how well these distributions separate [84].

Given two distributions $p(x)$ and $q(x)$, the Bhattacharyya distance $D_B(p, q)$ is defined as:

$$D_B(p, q) = -\ln\left(\sum_x \sqrt{p(x)q(x)}\right) \tag{4.12}$$

The term inside the logarithm, known as the Bhattacharyya coefficient, measures the overlap between the two distributions:

$$BC(p, q) = \sum_x \sqrt{p(x)q(x)} \tag{4.13}$$

By applying this metric to the distributions generated by different uncertainty measures (e.g., predictive entropy, mutual information, total variance), we can calculate the Bhattacharyya distance for both correct and incorrect predictions. This allows us to assess the distance between these distributions, providing valuable insights into the separability of correct and incorrect predictions.

We use this approach to evaluate the overlap between the distributions of correct and incorrect predictions based on several uncertainty metrics. A higher Bhattacharyya distance indicates less overlap between the distributions, suggesting that the model's uncertainty measures are better at distinguishing between correct and incorrect predictions.

$$D_B(p_{\text{correct}}, p_{\text{incorrect}}) = -\ln\left(\sum_x \sqrt{p_{\text{correct}}(x)p_{\text{incorrect}}(x)}\right) \tag{4.14}$$

This analysis provides a clearer understanding of the model's behavior when making incorrect predictions, helping to identify the conditions under which the model's uncertainty scores can be trusted to differentiate correct from incorrect classifications. By employing this approach, we can assess whether the model is likely to make an error. We then flag these uncertain predictions to prevent them from influencing the final results and minimize the risk of inaccuracies, as earlier discussed in Section 4.2.3.

4.3 Immunohistochemical (IHC) Subtype Classification Model

In the development of the IHC subtype classifier, we employed four advanced machine learning algorithms: RandomForest, GradientBoosting, ExtraTrees, and XGBoost. Each of these models represents a class of ensemble learning methods, which are particularly well-suited for complex, high-dimensional biological datasets. Below, we present a detailed overview of each model, focusing on their respective strengths and weaknesses in the context of IHC subtype classification.

4.3.1 Random Forest

The *Random Forest* algorithm, introduced by Breiman [85], is a widely-used ensemble method that constructs multiple decision trees during training and outputs the class that is the mode of the classes predicted by individual trees. One of the key strengths of RandomForest is its ability to handle large datasets with high-dimensional features, as it is resistant to overfitting and can capture complex interactions within the data. Additionally, RandomForest provides a measure of feature importance, which is useful for identifying key biomarkers in IHC data. However, the algorithm has limitations in terms of computational efficiency, especially when a large number of trees are required for stable performance, and interpretability becomes challenging when the model comprises hundreds or thousands of trees [85].

4.3.2 Gradient Boosting

The *GradientBoostingClassifier* algorithm, developed by Friedman [86], builds trees sequentially, with each new tree correcting the errors of the previous one. This method tends to outperform RandomForest in terms of accuracy, particularly for imbalanced and noisy datasets, which are common in biological classification tasks. Gradient Boosting is particularly advantageous in terms of reducing bias and achieving high accuracy with fewer trees than RandomForest. However, it is computationally more expensive, as trees are built iteratively, and the model is prone to overfitting if the hyperparameters, such as learning rate and tree depth, are not carefully tuned [86].

4.3.3 Extra Trees

The *ExtraTreesClassifier* (Extremely Randomized Trees) algorithm, introduced by Geurts et al. [87], is a variation of RandomForest that selects split points for decision trees entirely at random, as opposed to RandomForest's approach of selecting optimal splits from a random subset of features. This added randomness tends to reduce variance and computational time, making ExtraTrees more efficient for large-scale datasets. However, this can sometimes result in reduced accuracy compared to other ensemble methods, particularly if the dataset requires more finely tuned decision boundaries. Nonetheless, ExtraTrees remains an attractive option when computational speed is a critical factor [87].

4.3.4 XGBoost

XGBoost, developed by Chen and Guestrin [88], is an optimized version of Gradient Boosting that introduces several innovations, including regularization techniques

that prevent overfitting, advanced tree-pruning strategies, and efficient handling of missing data. XGBoost has become one of the leading algorithms for structured data classification tasks and is known for achieving state-of-the-art performance on many benchmark datasets. In our study, XGBoost's ability to handle sparse and imbalanced data makes it particularly effective. However, XGBoost's complexity requires careful tuning of multiple hyperparameters, which can increase the computational burden and make the model harder to interpret compared to simpler methods like RandomForest [88].

Chapter 5

Experiments and Results

This section presents the experiments conducted to evaluate the effectiveness of our proposed approach, along with a detailed analysis of the results obtained. We begin by outlining the experimental setup, including the datasets used, the preprocessing steps, and the parameters selected for our models. Subsequently, we describe the evaluation metrics employed to assess performance. Finally, we discuss the results in comparison with baseline methods, highlighting the improvements and insights gained from our study.

For the data preparation in training the ExpertDUT and IHC Classifier models, we employed a 3-fold cross-validation approach to ensure robust evaluation. Although nested cross-validation could provide a more unbiased performance estimate by separating hyperparameter tuning and model evaluation, as discussed in current research methodologies, we opted to continue using standard cross-validation. This approach provides a balance between computational efficiency, and effective model assessment, especially given the additional external test data that will further evaluate the model's generalization [89]. As discussed earlier in Section 3.3.1, we used a dataset (Nice A) comprising a total of 124 patients with 259 WSIs and their corresponding IHC Profile analysis for cross-validation, that were diagnosed with one of the four RCC subtypes: ccRCC, pRCC, chRCC and Oncocytoma. Given the imbalanced nature of our dataset, we adopted a stratified cross-validation strategy to preserve class distributions within each fold. The dataset was partitioned into three stratified subsets, with each fold serving iteratively as the test set while the remaining two were used for training. Importantly, the same patient data was consistently employed across both models during training and testing phases, ensuring a fair comparison of performance considering the ExpertDUT and IHC Classifier as a hybrid model. Figure 5.1 illustrates the distribution of our patients across the different folds.

Fold	Dataset	ccRCC	pRCC	chRCC	Oncocytoma	Total
Fold 1	Training	42	20	10	9	81
	Test	22	10	5	5	42
Fold 2	Training	43	20	10	9	82
	Test	21	10	5	5	41
Fold 3	Training	43	20	10	10	83
	Test	21	10	5	4	40

Table 5.1: Fold distribution of training and validation sets for each RCC subtype in Nice A cohort.

5.1 ExpertDUT Training

The ExpertDUT architecture consists of multiple levels of binary classification, arranged in a tree structure. Each binary CNN classifier was trained independently using the specific subset of data relevant to its classification task. The WSIs were stained using the H&E method and scanned with ×40 magnification as the highest magnification level.

In order to prepare the data for training ExpertDUT at different levels of the tree, the patients' WSIs annotations, so called ROIs (See Section 3.3.1) were cropped into smaller, equally sized patches of 1000×1000 pixels, subsequently downscaled to 112x112 pixels for input into the CNN binary classifiers. The total number of training patches for each level of ExpertDUT has been reported in Table 5.2. Due to imbalanace nature of our datasets, we used downsampling approach to have equal number of patches among all normal and cancerous tissue types. We have reported the total number of patches related to each level of ExpertDUT across different folds in Table 5.3, 5.4 and 5.5.

Type	Number of Patches
ccRCC	175,471
pRCC	94,773
CHROMO	42,540
ONCOCYTOMA	39,899
Normal	61,707
Necrosis	26,779
Fiber	24,948

Table 5.2: Total Number of Patches (1000×1000 pixels) by Tissue in ×40 Magnification

Category	Root	Node	Leaf 1	Leaf 2
normal	42277	-	-	-
fibrosis	16456	-	-	-
necrosis	17785	-	-	-
ccRCC	10569	26825	-	50057
pRCC	10569	26825	-	50057
CHROMO	10569	26825	26825	-
ONCOCYTOMA	10569	26825	26825	-
Total	118794	107300	53650	100114

Table 5.3: Patches label counts for fold 1 across different levels

Category	Root	Node	Leaf 1	Leaf 2
normal	46495	-	-	-
fibrosis	21405	-	-	-
necrosis	13996	-	-	-
ccRCC	11623	25538	-	74137
pRCC	11623	25538	-	74137
CHROMO	11623	25538	25538	-
ONCOCYTOMA	11623	25538	25538	-
Total	128388	102152	51076	148274

Table 5.4: Patches label counts for fold 2 across different levels

After extracting the image patches, each CNN binary classifier was trained individually on its corresponding dataset. The training protocol closely followed the methodology proposed by Ponzio et al. [18]. Specifically, all classifiers utilized the VGG16 architecture initialized with weights pre-trained on the ImageNet dataset. To retain essential learned features and prevent overfitting, the first 11 layers of the VGG16 network were kept frozen during training. The models were trained over 150 epochs using the Adaptive Moment Estimation (ADAM) optimizer, with a learning rate of $1e-5$ and a batch size of 128. To further mitigate the risk of overfitting, we implemented an early stopping strategy: the training process was terminated if the training loss did not show improvement over 20 consecutive epochs.

5.1.1 Monte Carlo Root Effect

After training all the CNN binary classifiers, to enhance the model's ability to flag incorrect patch predictions as uncertain at the Root level of ExpertDUT, we began by thoroughly analyzing the distribution histograms of various uncertainty estimation methodologies applied to the softmax probabilities obtained from the stochastic forward passes during the Monte Carlo dropout process. We found that 10 stochastic forward passes were sufficient to achieve the approximation

Category	Root	Node	Leaf 1	Leaf 2
normal	34642	-	-	-
fibrosis	12035	-	-	-
necrosis	21777	-	-	-
ccRCC	8660	21475	-	65352
pRCC	8660	21475	-	65352
CHROMO	8660	21475	21475	-
ONCOCYTOMA	8660	21475	21475	-
Total	103094	85900	42950	130704

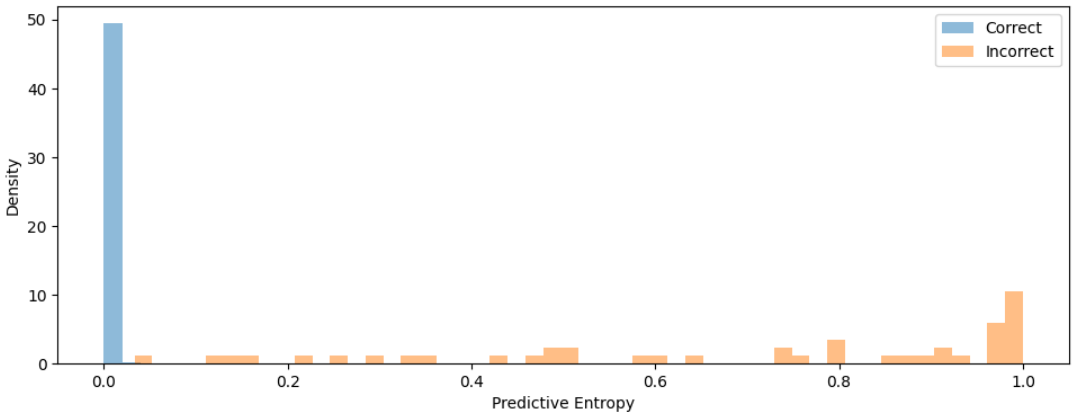
Table 5.5: Patches label counts for fold 3 across different levels

required for our analysis. Although we evaluated additional forward passes (e.g. 20, 30), they did not yield significant improvements, but instead led to increased computational costs. We calculated the distance between the distribution of correct and incorrect predictions for each method and we found that Predictive entropy and rényi entropy yielded larger Bhattacharyya distances. This larger distance indicates a better separability between the distributions of uncertainty scores for correct and incorrect predictions, which is crucial for reliable uncertainty quantification.

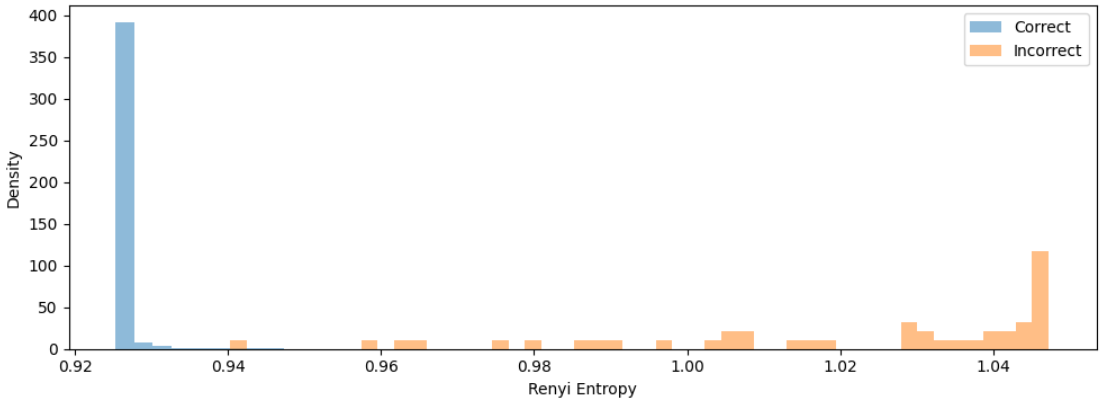
To optimize the effectiveness of rényi entropy, we investigated different values of the parameter α that controls its sensitivity to the tail probabilities of the distribution. Smaller α values give more weight to smaller probabilities, making the entropy more sensitive to rare events, while larger values give more importance to higher probabilities, making it less sensitive to rare events. When $\alpha = 1$ rényi entropy reduces to Shannon entropy, capturing the traditional concept of information entropy. By systematically adjusting α, we aimed to maximize the Bhattacharyya distance between the distributions of correct and incorrect predictions in rényi entropy.

Our experiments revealed that an optimized value in the rényi entropy outperformed the predictive entropy. This optimization enhances the model's ability to distinguish between correct and incorrect predictions more effectively than using the standard predictive entropy. Figure 5.1 shows the the rényi and predictive entropy histogram between correct and incorrect predictions. Other histograms related to other uncertainty measures have been reported in appendix (See Figure A.1).

Subsequently, we addressed the challenge of setting an appropriate threshold to classify incorrect predictions as uncertain predictions based on their uncertainty scores. We initially identified a potential threshold visually by examining the distributions of uncertainty histograms. To validate and refine this threshold, we followed the approach by Milanes et al. [83]. We try to find the highest uncertainty accuracy while having less number of correct predictions to be considered as uncertain. Figure A.2 compares different thresholds with different uncertainty



(a) Predictive entropy, Bhattacharyya distance = 3.00758



(b) rényi entropy with $\alpha = 0.5591$, Bhattacharyya distance = 3.07275

Figure 5.1: Comparison between a) predictive entropy and b) optimized ényi entropy

metrics. Here is the definition to the uncertainty metrics:

- N_{cc} is the number of correct and certain patch predictions,
- N_{ic} is the number of incorrect but certain patch predictions,
- N_{iu} is the number of incorrect and uncertain patch predictions,
- N_{cu} is the number of correct but uncertain patch predictions,
- N is the total number of Root level training patches,

Based on a comprehensive evaluation, we selected an rényi entropy threshold of ‘0.9402‘ with $\alpha = 0.5591$ to balance the trade-off between correctly identifying

incorrect patch predictions as uncertain patches and minimizing the uncertain flags for correct predictions. This threshold effectively flags predictions with high uncertainty, preventing them from adversely affecting the final results. Table 5.6 compare the training patch-level classification accuracy performance between normal Root level and MC-Root level. Also, we also report the patient-level training confusion matrices with normal Root level and MC-Root level in Figure 5.2.

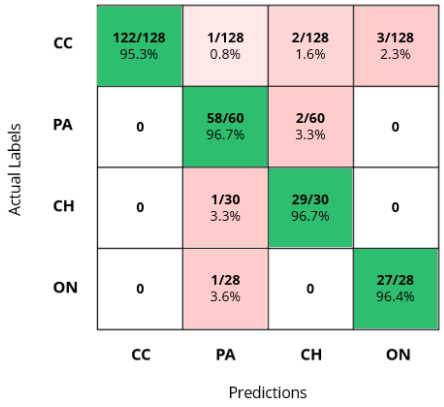
	Training	**Validation**
Normal Root	99.96%	96.52%
MCRoot	100%	98.28%

Table 5.6: Training and validation patch-level accuracy for normal Root and MC-Root binary classifiers

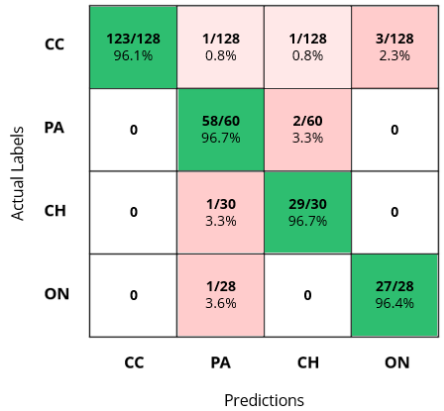
5.2 IHC Classifier Training

To assess the effectiveness of our mutli-class IHC Classifier model, we conducted a series of experiments using experts quantitive interprested data from different biomarkers expression. We have respected the exact same cross-validation setup distribution of patients used to train and test the ExpertDUT. To identify the optimal methodologies for different parts of the classifier training pipeline, as discussed previously in Section 4.1.2, we employed a grid search to optimize the pipeline. In addition to that, we have also applied a Recursive Feature Elimination (RFE) method at the beginning of our pipeline to identify and select the most significant biomarkers contributing to our study. Table 5.7 summarizes all the possible options for different components of the pipeline from data preprocessing (See Section 4.1.2) to model training.

The grid search experiments, shown in Table 5.8 demonstrated that incorporating the H-Score as a feature enhancer significantly boosted model performance across various configurations. Notably, Models with 5 and 6 biomarkers, which utilized the GradientBoosting classifier alongside the H-Score feature enhancer and Median imputation, achieved an impressive test accuracy of 97.5%. This exceptional performance remained consistent regardless of the resampling technique employed; both SMOTE and Tomek Links produced similar outcomes, suggesting that the choice of resampler had minimal impact when optimal feature enhancement and classification methods were applied. Furthermore, the GradientBoosting classifier outperformed both the RandomForest and ExtraTrees classifiers, underscoring its superior ability to capture the underlying patterns in the data. The Median imputer also proved to be more effective than the KNN imputer, further contributing to the model's elevated accuracy. Overall, the grid search process enabled us to identify



(a) ExpertDT with Normal Root (Ponzio et al.) [18], Weighted Train Accuracy = 96.27%



(b) ExpertDUT with MC-Root, Weighted Train Accuracy = 96.46%

Figure 5.2: Comparison between normal Root and MC-Root in ExpertDUT; CC: ccRCC, PA: pRCC, CH: chRCC, ON: oncocytoma

the optimal combination of pipeline components, emphasizing the critical role of feature enhancement and classifier selection in achieving accurate predictions.

In conjunction with Figure 5.3, which represents the decision tree pathologists use to classify RCC subtypes based on IHC analysis, the model's biomarker selection strategy becomes particularly insightful. Beginning with just two biomarkers, the model demonstrates a clear, structured approach: it first selects a biomarker to guide the classification down to the leaf nodes. Subsequently, it strategically picks another biomarker that is effective across both branches, allowing for the accurate distinction between different subtypes at the terminal leaves. This hierarchical decision-making mirrors the clinical diagnostic process, reinforcing the model's ability to emulate expert-level decisions by utilizing minimal yet highly informative

Step	Method
Feature Selector	1 Biomarker
	2 Biomarker
	3 Biomarker
	4 Biomarker
	5 Biomarker
	6 Biomarker
Feature Enhancer	Baseline
	Percentage Only
	Intensity Only
	Combined
	H-Score
Imputer	KNN Imputer
	Mean Imputer
	Median Imputer
Resampler	SMOTE
	Tomek Links
	SMOTEENN
Classifier	RandomForrest
	GradientBoosting
	ExtraTrees
	XGBoost

Table 5.7: IHC training pipeline components for grid search

biomarkers. Starting from three biomarkers onward, where we observe a significant boost in model accuracy, the model's selection strategy becomes even more refined. Initially, it prioritizes biomarkers that ensure precise classification at the leaf level that improves the distinctions between subtypes. Then, it selects additional biomarkers at the node level to further enhance classification accuracy. This approach demonstrates the model's capacity to optimize both early and late-stage decisions, leveraging key biomarkers not just for initial differentiation but also for fine-tuning accuracy as it progresses through the decision tree. It also provides the flexibility to select the number of biomarkers used in the IHC classifier based on clinical requirements, while considering the accuracy of each model.

No. Biomarker	Model	Resampler	Imputer	Features	Train/Test Accuracy
1	ExtraTrees	TomekLinks	Median	Combined	70.3%/72.3%
2	RandomForest	TomekLinks	KNN	Combined	95.0%/92.6%
3	ExtraTrees	SMOTE	KNN	Combined	99.4%/92.6%
4	RandomForest	TomekLinks	Median	H-Score	100.0%/95.1%
5	GradientBoosting	TomekLinks	Median	H-Score	100.0%/97.5%
6	GradientBoosting	SMOTE	Median	H-Score	100.0%/97.5%

Table 5.8: Results of the Grid Search Experiment

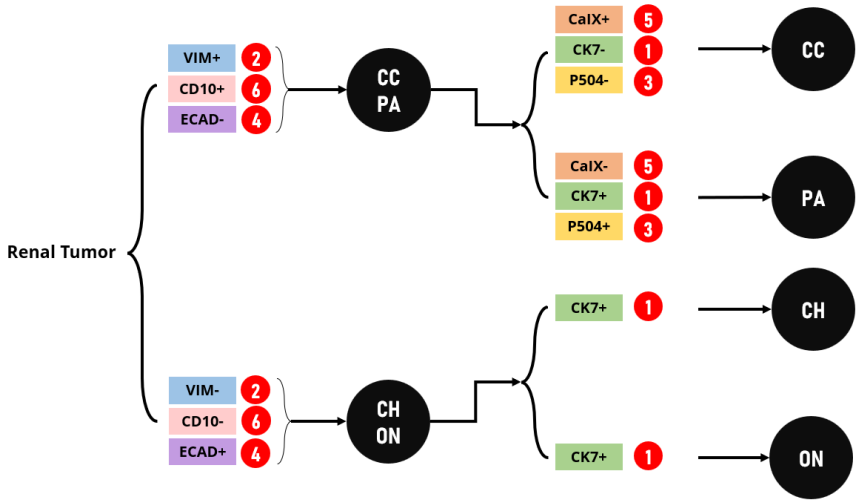


Figure 5.3: Comparison of the IHC model's biomarker selection with the pathologists' decision tree for identifying RCC subtypes. CC: ccRCC, PA: pRCC, CH: chRCC, ON: oncocytoma. The numbers in front of each biomarker indicate the order in which the model selects them, starting with CK7 as the first and CD10 as the final biomarker chosen.

5.3 Hybrid Model: ExpertDUT + IHC Classifier

At the outset of our work, we proposed developing an uncertainty-aware pipeline where the model would initially classify subtypes based on microscopic morphological features, a common practice in clinical settings. In cases of uncertainty, the pipeline would then consult the IHC model to confirm or revise the initial classification. So far, we have demonstrated that both methods, when applied independently, yield promising results. Now, we aim to integrate them into a more robust decision-making pipeline.

The first step is identifying the point of convergence between these models and determining the threshold for transitioning from ExpertDUT to IHC analysis. As previously mentioned, ExpertDUT provides confidence scores for its subtype classifications. Our next objective is to analyze the distribution of these confidence scores on the training set, distinguishing between correct and incorrect subtype classifications. This will help inform the threshold for invoking the IHC model. Figure 5.4 illustrates the distribution of confidence scores for correct and incorrect predictions made by the model. Correct predictions are tightly clustered around high confidence values, with a median close to 1 and a narrow interquartile range (IQR), indicating the model is consistently confident in its correct classifications except some outliers that the model correctly classified the subtype but not very

confident. In contrast, the incorrect predictions exhibit a much broader distribution of confidence scores, with a median around 0.8 and an IQR spanning from approximately 55% to 85%. Some incorrect predictions also extend below 50%, highlighting the model's uncertainty in these cases. Based on this clear separation, an 82% confidence threshold was chosen. This threshold corresponds to the lower bound of the IQR for correct predictions, indicating that the majority of correct classifications have confidence scores exceeding 82%, while a substantial portion of incorrect classifications falls below this threshold. Increasing the threshold would result in more patients being flagged as uncertain, subsequently requiring additional classification via the IHC classifier. However, to balance the trade-off between accuracy and cost, we opted for this threshold to minimize the number of uncertain patients while still maintaining high classification accuracy (see Figure A.2). This approach was specifically demonstrated for fold 2 of the training scheme, though the process for selecting the threshold is consistent across all folds.

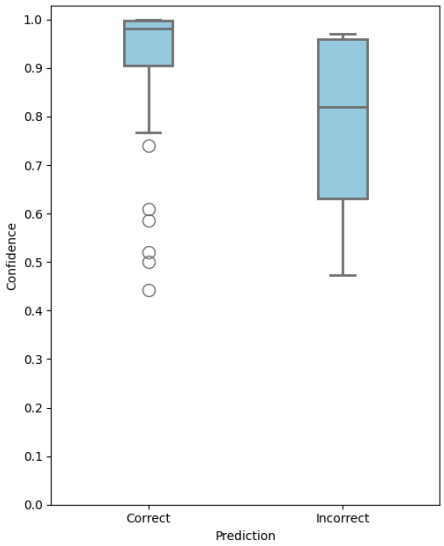


Figure 5.4: Box plot of correct and incorrect subtype classification made by ExpertDUT with respect to their confidence scores

Figure 5.5 demonstrates the accuracy achieved with different number of biomarkers in IHC Classifier, while Figure 5.6 presents the hybrid model subtype classification confusion matrix, which includes ExpertDUT and the IHC classifier using 4, 5, or 6 biomarkers, achieving an accuracy of 99.39%. We also illustrate how varying threshold settings can influence the accuracy in our training set. For instance, we demonstrate that our current threshold configuration maintains a high accuracy while minimizing the number of IHC analyses, thereby reducing associated costs. The cost of IHC analysis was calculated based on the price of biomarker vials

used in tests with the Dako Omnis automated immunohistochemistry and in situ hybridization platform, along with additional clinical costs for quantifying IHC signals. Figure 5.7 shows how the cost of IHC analysis increases with the number of biomarkers while also correlating to the accuracy achieved. Similarly, Figure A.3 and Figure A.4, present the same analysis for the IHC classifier using 3 and 2 biomarkers, respectively. These results demonstrate that with fewer biomarkers, the IHC model introduces errors, leading to misclassification of subtypes that were correctly identified by ExpertDUT, especially when analyzing a large number of patients.

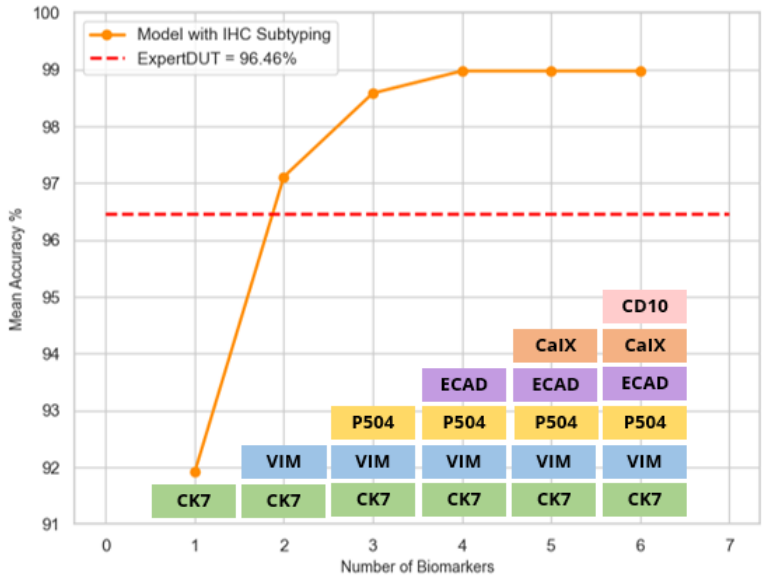
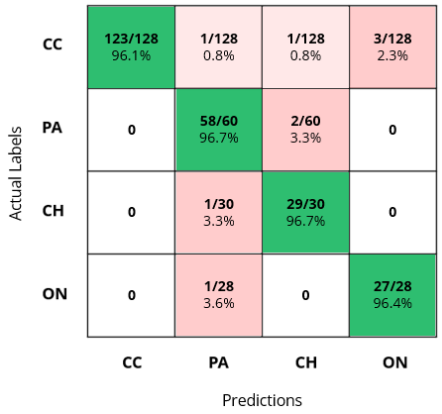
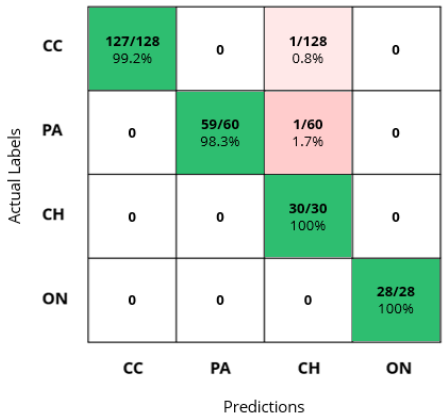


Figure 5.5: Training accuracy vs Number of patients analyzed through IHC Classifier with 4, 5 or 6 biomarkers



(a) ExpertDUT, Weighted Accuracy = 96.46%



(b) Hybrid Model (ExpertDUT + IHC Classifier), Weighted Accuracy = 99.39%

Figure 5.6: Comparison between ExpertDUT and Hybrid Model (ExpertDUT + IHC Classifier); CC: ccRCC, PA: pRCC, CH: chRCC, ON: oncocytoma

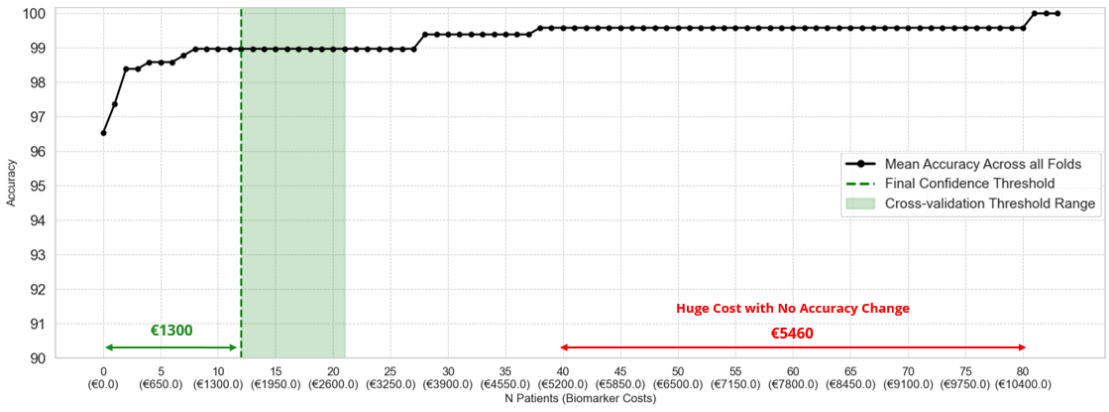
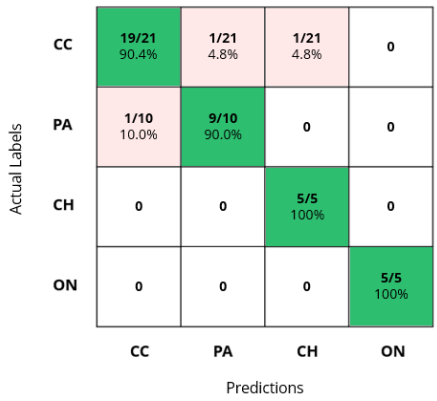
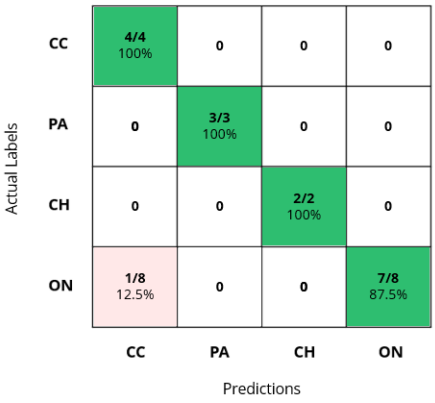


Figure 5.7: Training accuracy vs Number of patients analyzed through IHC Classifier with 4, 5 or 6 biomarkers

Finally, we present the results from our validation and test sets in Figure 5.8 and Figure 5.9. Since only the Nice cohort datasets included the full data (WSIs and IHC profile data), we report the hybrid model's performance on these datasets. For the external test cohorts, we report only the results of the ExpertDUT model. As shown, the hybrid model performs well on both the Nice A and B datasets. In the external test cohorts, ExpertDUT demonstrates strong performance on the Lyon dataset, despite the use of different staining techniques. However, in the Paris Cochin dataset, the model struggles with pRCC classification, frequently misclassifying it as chRCC but overall we see almost the same good performance among other subtypes.

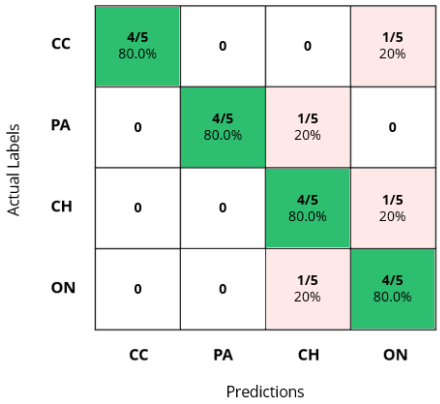


(a) Validation Result on Nice A Dataset, Weighted Accuracy = 95.11%

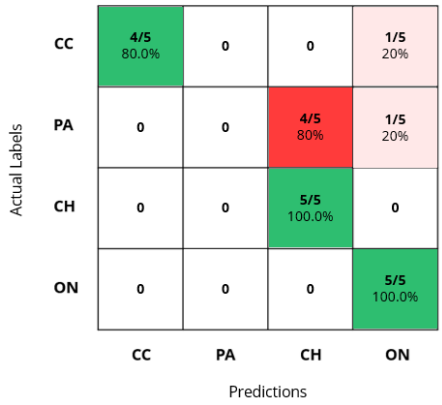


(b) Test Result on Nice B Dataset, Weighted Accuracy = 96.88%

Figure 5.8: Result on Validation and Test set on Nice cohort using Hybrid Model: ExpertDUT + IHC Classifier; CC: ccRCC, PA: pRCC, CH: chRCC, ON: oncocytoma



(a) Test Result on Lyon Dataset, , Weighted Accuracy = 80.00%



(b) Test Result on Paris Cochin Dataset, Weighted Accuracy = 70.00%

Figure 5.9: Result External test set on Lyon and Paris Cochin cohort using ExpertDUT; CC: ccRCC, PA: pRCC, CH: chRCC, ON: oncocytoma

Chapter 6

Conclusion and Future Works

In this work, we presented a hybrid model integrating deep learning and traditional machine learning techniques to RCC subtypes. By leveraging CNNs for the initial tumor detection and incorporating an IHC profile validation for uncertain cases, we enhanced both the accuracy and confidence of the classification process. Our results demonstrate that the hybrid model improves diagnostic precision, offering a comprehensive and efficient solution for RCC subtype classification with an interpretable approach for uncertain and difficult cases that leads to a more reliable and clinically applicable system. Additionally, our model's performance was validated on both internal and external cohorts, showing strong generalizability across diverse clinical settings.

Despite the promising results, several potential enhancements remain for future research. Incorporating genomic data alongside histopathological and IHC data could lead to even higher accuracy in RCC subtype classification and enable more tailored, patient-specific treatment recommendations. Additionally, automating IHC quantification as part of the classification pipeline could be an interesting area of exploration. It is also worth investigating the direct inclusion of IHC images into the model, bypassing the need for manual profile quantification, which could simplify the workflow. Another avenue for future improvement would involve the inclusion of additional RCC subtypes or maybe other type of cancers to assess how the model performs across a wider range of classifications. This would not only expand the model's applicability but also test its ability to generalize further across complex, heterogeneous data. Moreover, advancing towards real-time applications of this hybrid model in clinical settings would be of great value, enabling fast and reliable diagnosis while reducing the diagnostic time for pathologists.

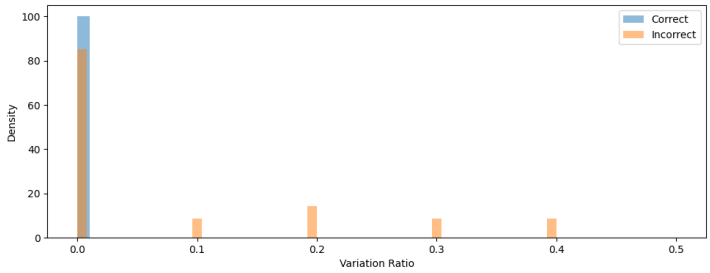
In conclusion, this work presents a comprehensive, uncertainty-aware framework

that integrates CNN-based deep learning with IHC profile analysis, offering a highly accurate and efficient solution for RCC subtype classification. As AI continues to push the boundaries in medical diagnostics, the proposed hybrid approach represents a promising step towards more precise, automated, and scalable cancer diagnosis methods.

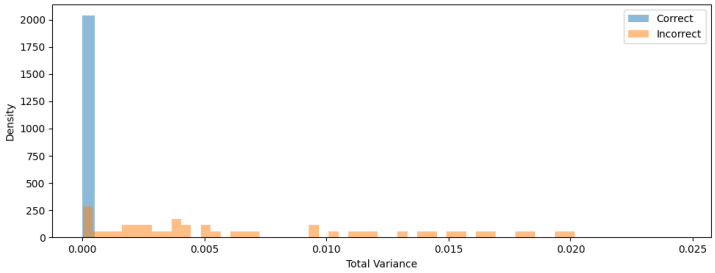
Appendix A

Listing A.1: Example of XML annotations in ASAP format

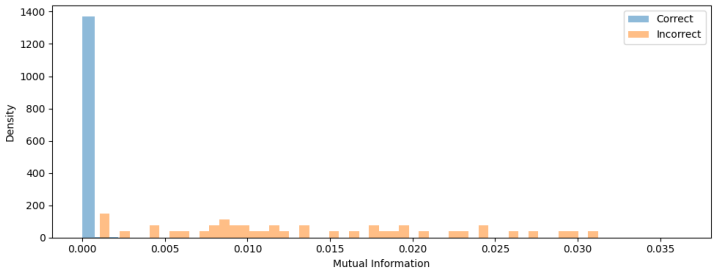
```
<?xml version="1.0"?>
<ASAP_Annotations>
    <Annotations>
    <Annotation Name="Annotation 0" Type="Polygon" PartOfGroup="
fiber" Color="#F4FA58">
        <Coordinates>
        <Coordinate Order="0" X="61283.0898" Y="65948.7344" />
        <Coordinate Order="1" X="61655.1953" Y="67065.0547" />
        <Coordinate Order="2" X="61941.4336" Y="68210" />
        <Coordinate Order="3" X="62113.1758" Y="69354.9375" />
        <Coordinate Order="4" X="62170.4219" Y="70499.8828" />
        <Coordinate Order="5" X="62227.668" Y="71673.4531" />
        </Coordinates>
    </Annotation>
    </Annotations>
    <AnnotationGroups>
    <Group Name="fiber" PartOfGroup="None" Color="#64FE2E">
        <Attributes />
    </Group>
    </AnnotationGroups>
</ASAP_Annotations>

```



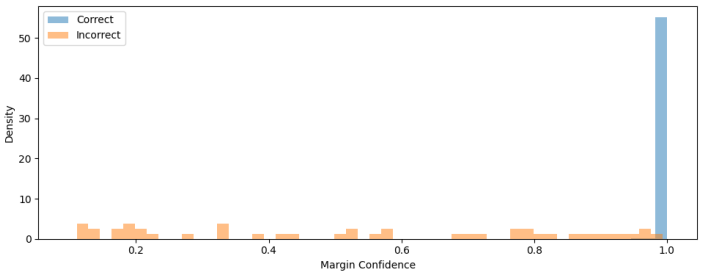
(a) Variation Ratio, Bhattacharyya distance = 0.18562



(b) Total Variation, Bhattacharyya distance = 0.92450



(c) Mutual Information, Bhattacharyya distance = 2.96869



(d) Margin of Confidence, Bhattacharyya distance = 1.63617

Figure A.1: Histogram of Correct/Incorrect predictions of MC-Root with a) Variation Ratio, b) Total Variation, c) Mutual Information and d) Margin of Confidence uncertainty measurements

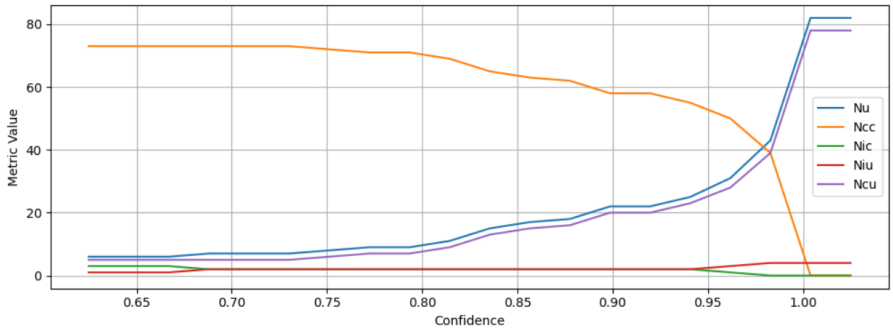


Figure A.2: Comparing different confidence threshold with different uncertainty metrics. The uncertainty metrics demonstrate that as the threshold increases, a greater number of patients are classified as correct but uncertain, highlighting the trade-off between confidence and classification accuracy

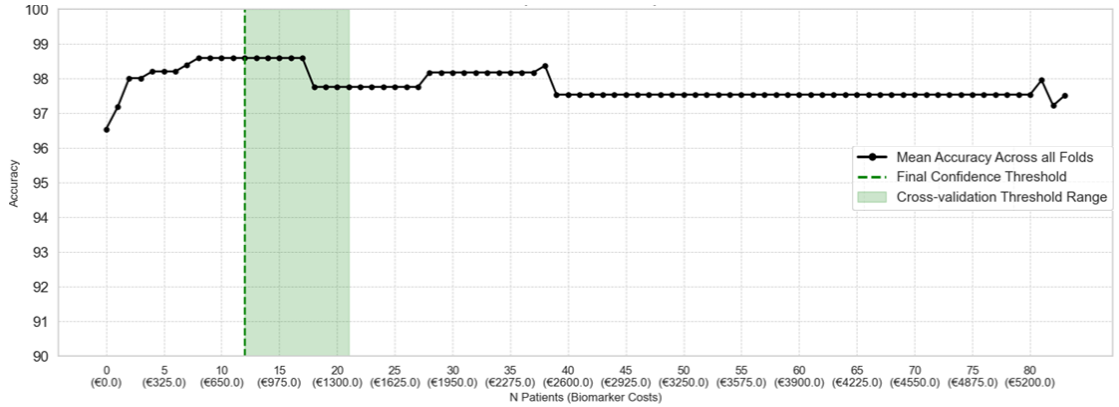


Figure A.3: Training accuracy vs Number of patients analyzed through IHC Classifier with 3 biomarkers

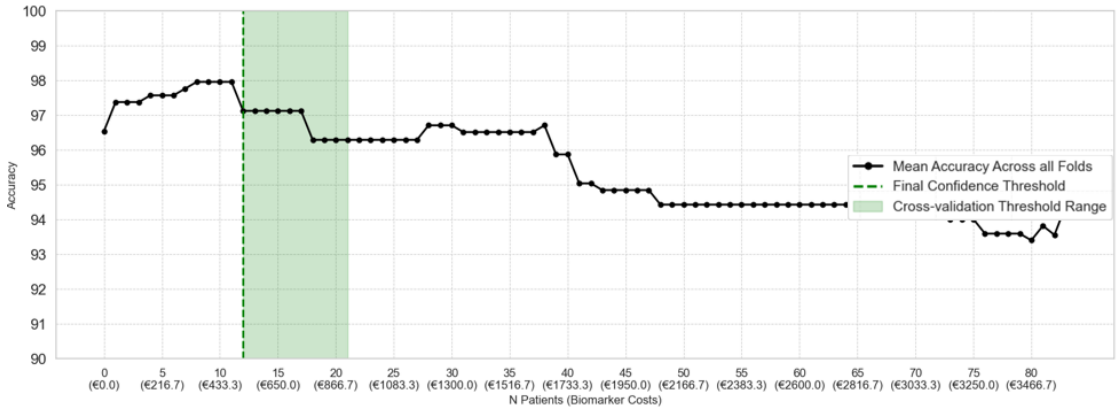


Figure A.4: Training accuracy vs Number of patients analyzed through IHC Classifier with 2 biomarkers

Bibliography

[1] Debbie Fortnum. «Introduction to Kidney Functions and Pathophysiology». In: *Principles of Specialty Nursing*. Springer, 2024, pp. 1–20. DOI: 10.1007/978-3-031-30320-3_1 (cit. on p. 1).

[2] Harry G. Preuss. «Basics of Renal Anatomy and Physiology». In: *Clinics in Laboratory Medicine* 13.1 (1993), pp. 1–20. DOI: 10.1016/S0272-2712(18)30456-6 (cit. on p. 1).

[3] Mitchell H. Rosner. «An Overview of Renal Physiology». In: *Comprehensive Clinical Nephrology*. Elsevier, 2011, pp. 150–170. DOI: 10.1007/978-1-84882-034-0_7 (cit. on p. 1).

[4] «Renal Physiology Theory and Functional Assessment». In: *Nephrology Nursing*. Elsevier, 2023, pp. 50–75. DOI: 10.1016/b978-0-323-95884-4.00009-3 (cit. on p. 1).

[5] Khalid Bashir Robert S. McMahon Dana Penfold. «Anatomy, Abdomen and Pelvis, Kidney Collecting Ducts». In: *Open Access Anatomy* 1.1 (2019), pp. 1–5 (cit. on p. 1).

[6] Kiril Trpkov Reza Alaghehbandan Farshid Siadat. «What's new in the WHO 2022 classification of kidney tumours?» In: *Pathologica* 115.1 (2023), pp. 1–10. DOI: 10.32074/1591-951x-818 (cit. on p. 1).

[7] Jonathan Kanakaraj, Justin Chang, Lance J. Hampton, and Steven Christopher Smith. «The New WHO Category of Molecularly Defined Renal Carcinomas: Clinical and Diagnostic Features and Management Implications». In: *Urologic Oncology* 40.2 (2024), pp. 1–5. DOI: 10.1016/j.urolonc.2024.02.003 (cit. on p. 1).

[8] Ying-Bei Chen. «Update on Selected High-grade Renal Cell Carcinomas of the Kidney: FH-deficient, ALK-rearranged, and Medullary Carcinomas». In: *Advances in Anatomic Pathology* 31 (Dec. 2023). DOI: 10.1097/PAP.0000000000000426 (cit. on p. 2).

[9] Marta Amann-Arévalo, Pablo Ballestín Martínez, Natalia Vidal Cassinello, Ignacio Moreno Perez, Montserrat de la Torre-Serrano, and Javier Puente. «Molecularly Defined Renal Carcinomas». In: *Kidney Cancer* 11.2 (2024), pp. 150–165. DOI: 10.3233/kca-230015 (cit. on p. 2).

[10] B. Escudier et al. «Renal cell carcinoma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up». In: *Annals of Oncology* 30.5 (2019). Electronic address: clinicalguidelines@esmo.org, pp. 706–720. DOI: 10.1093/annonc/mdz056 (cit. on p. 3).

[11] Andrew N. Young, Mahul B. Amin, John A. Petros, M.J. Natan, Shuming Nie, and May D. Wang. «Nanomolecular Histopathology for Renal Tumor Classification». In: *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society*. 2005, pp. 1618–1621. DOI: 10.1109/IEMBS.2005.1616516 (cit. on pp. 3, 4).

[12] Naima Tariq, Nadira Mamoon, Asna Haroon, Zafar Ali, and Imran Ahmad. «Role Of Immunohistochemistry In Subtyping Renal Cell Carcinomas With Overlapping Morphological Features». In: *Journal of Ayub Medical College Abbottabad* 30.3 (2018), pp. 389–394 (cit. on pp. 3, 4).

[13] D D'Souza, D Bell, T Fahrenhorst-Jones, et al. «Renal oncocytoma». In: *Radiopaedia.org* (2024). Accessed on 10 Sep 2024. URL: https://doi.org/10.53347/rID-1969 (cit. on p. 4).

[14] Christopher J. Ricketts and Marston Linehan. «Molecular Genetics of Renal Cell Carcinoma». In: *Annual Review of Medicine* 69 (2018), pp. 207–219. DOI: 10.1146/annurev-med-062016-090533 (cit. on p. 4).

[15] Yanjie Han. «Deep learning methods and corresponding applications in medical imaging». In: *Applied and Computational Engineering* 46 (2024), pp. 161–172. DOI: 10.54254/2755-2721/46/20241106 (cit. on pp. 5, 6).

[16] Ruaa Jasim Al Gharrawi and Alyaa Abdulhussein Al-Joda. «A Survey of Medical Image Analysis Based on Machine Learning Techniques». In: *Journal of Al-Qadisiya for Computer Science and Mathematics* 15.1 (2023), pp. 389–394. DOI: 10.29304/jqcm.2023.15.1.1139 (cit. on p. 5).

[17] Neha Baranwal, Preethi Doravari, and Renu Kachhoria. «Classification of Histopathology Images of Lung Cancer Using Convolutional Neural Network (CNN)». In: *arXiv preprint arXiv:2112.13553* (2021). URL: https://arxiv.org/abs/2112.13553 (cit. on p. 5).

[18] Francesco Ponzio, Xavier Descombes, and Damien Ambrosetti. «Improving CNNs classification with pathologist-based expertise: the renal cell carcinoma case study». In: *Scientific Reports* 13 (2023), p. 15887. DOI: 10.1038/s41598-023-42847-y. URL: https://www.nature.com/articles/s41598-023-42847-y (cit. on pp. 5, 9, 27, 33, 46, 50).

[19] A. Manju, M.C. Arivukarasi, and M. Mahasree. «AEDAMIDL: An Enhanced and Discriminant Analysis of Medical Images using Deep Learning». In: *Proceedings of the International Conference on Science and Technology*. 2022. DOI: 10.1109/ICSTCEE56972.2022.10100240 (cit. on p. 5).

[20] Geert Litjens et al. «1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset». In: *GigaScience* 7.6 (2018), giy065. DOI: 10.1093/gigascience/giy065 (cit. on pp. 6, 7).

[21] Mohamad Mohamad. «Self-Supervision for Renal Cell Carcinoma Subtyping». Relators: Santa Di Cataldo, Francesco Ponzio. Master's Thesis. Politecnico di Torino, 2023. URL: https://webthesis.biblio.polito.it/26847/ (cit. on pp. 6, 11).

[22] Jeroen Van der Laak, Geert Litjens, and Francesco Ciompi. «Deep learning in histopathology: the path to the clinic». In: *Nature medicine* 27.5 (2021), pp. 775–784 (cit. on p. 7).

[23] Alfredo Distante et al. «Artificial Intelligence in Renal Cell Carcinoma Histopathology: Current Applications and Future Perspectives». In: *Diagnostics* 13.2294 (2023), pp. 1–25. DOI: 10.3390/diagnostics13132294 (cit. on p. 7).

[24] Michaela Unger and J. N. Kather. «Deep learning in cancer genomics and histopathology». In: *Genome Medicine* 16 (2024), pp. 1–13 (cit. on p. 7).

[25] Md. Shamim Hossain, Leisa Armstrong, David M. Cook, and Pauline Zaenker. «Application of Histopathology Image Analysis Using Deep Learning Networks». In: *Human-centric Intelligent Systems* 4 (2024), pp. 95–106 (cit. on p. 7).

[26] Mengdan Zhu, Bing Ren, Ryland Richards, Matthew Suriawinata, Naofumi Tomita, and Saeed Hassanpour. «Development and evaluation of a deep neural network for histologic classification of renal cell carcinoma on biopsy and surgical resection slides». In: *Scientific Reports* 11 (2021), p. 7080. DOI: 10.1038/s41598-021-86540-4. URL: https://www.nature.com/articles/s41598-021-86540-4 (cit. on p. 8).

[27] Hisham A. Abdeltawab, Fahmi A. Khalifa, Mohammed A. Ghazal, Liang Cheng, Ayman S. El-Baz, and Dibson D. Gondim. «A deep learning framework for automated classification of histopathological kidney whole-slide images». In: *Journal of Pathology Informatics* 13 (2022). n/a, p. 100093 (cit. on p. 8).

[28] Matthew Fenstermaker, John Doe, and Jane Roe. «Deep Learning for RCC Diagnosis and Subtyping in Renal Mass Biopsies». In: *Journal of Urology* 210.5 (2023), pp. 1132–1140. DOI: 10.1016/j.juro.2023.05.001 (cit. on pp. 8, 9).

[29] Samuel Tabibu, Prabhav Vinod, and C V Jawahar. «Pan-Renal Cell Carcinoma Classification and Survival Prediction from Histopathology Images Using Deep Learning». In: *JCO Clinical Cancer Informatics* 3 (2019), pp. 1–14. DOI: 10.1200/CCI.18.00069. URL: https://pubmed.ncbi.nlm.nih.gov/31324828/ (cit. on p. 9).

[30] Jing Chen, Hong Wei, and Rui Zhang. «RCC Subtyping and Prognosis Prediction Using LASSO and CNN Models». In: *The Lancet Oncology* 24.6 (2023), pp. 763–775. DOI: 10.1016/S1470-2045(23)30123-9 (cit. on p. 9).

[31] Wieland Brendel and Matthias Bethge. «Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet». In: *arXiv preprint arXiv:1904.00760* (2019). DOI: 10.48550/ARXIV.1904.00760. URL: https://arxiv.org/abs/1904.00760 (cit. on p. 10).

[32] Michael Gadermayr and Maximilian Tschuchnig. «Multiple Instance Learning for Digital Pathology: A Review on the State-of-the-Art, Limitations & Future Potential». In: *arXiv preprint arXiv:2206.04425* (2022). DOI: 10.48550/ARXIV.2206.04425. URL: https://arxiv.org/abs/2206.04425 (cit. on p. 10).

[33] Le Hou, Dimitris Samaras, Tahsin M Kurc, Yi Gao, James E Davis, and Joel H Saltz. «Patch-based Convolutional Neural Network for Whole Slide Tissue Image Classification». In: *arXiv preprint arXiv:1504.07947* (2015). DOI: 10.48550/ARXIV.1504.07947. URL: https://arxiv.org/abs/1504.07947 (cit. on p. 10).

[34] Zhipeng Jia, Xingyi Huang, Eric I-Chao Chang, and Yan Xu. «Constrained Deep Weak Supervision for Histopathology Image Segmentation». In: *IEEE Transactions on Medical Imaging* 36.11 (2017), pp. 2376–2388. DOI: 10.1109/TMI.2017.2724070. URL: https://doi.org/10.1109/tmi.2017.2724070 (cit. on p. 10).

[35] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. «Attention-based Deep Multiple Instance Learning». In: *arXiv preprint arXiv:1802.04712* (2018). DOI: 10.48550/ARXIV.1802.04712. URL: https://arxiv.org/abs/1802.04712 (cit. on p. 10).

[36] Yasmine Abu Haeyeh, Mohammed Ghazal, Ayman El-Baz, and Iman M. Talaat. «Development and Evaluation of a Novel Deep-Learning-Based Framework for the Classification of Renal Histopathology Images». In: *Bioengineering* 9.9 (2022), p. 423. DOI: 10.3390/bioengineering9090423 (cit. on p. 10).

[37] Qingyuan Zheng, Rui Yang, Panpan Jiao, Xinmiao Ni, Jingping Yuan, Liang Wang, Zhi-Yuan Chen, and Xiuheng Liu. «A Weakly Supervised Deep Learning Model and Human-Machine Fusion for Accurate Grading of Renal Cell Carcinoma from Histopathology Slides». In: *Cancers* 15.12 (2023), p. 3198. DOI: 10.3390/cancers15123198 (cit. on pp. 10, 11).

[38] Navid Alemi Koohbanani, Balagopal Unnikrishnan, Syed Ali Khurram, Pavitra Krishnaswamy, and Nasir Rajpoot. «Self-Path: Self-supervision for Classification of Pathology Images with Limited Annotations». In: *arXiv preprint arXiv:2008.05571* (2020) (cit. on p. 11).

[39] Frederik Wessels, Max Schmitt, Eva Krieghoff Henning, et al. «A self-supervised vision transformer to predict survival from histopathology in renal cell carcinoma». In: *World Journal of Urology* 41 (2023), pp. 769–778. DOI: 10.1007/s00345-023-04489-7 (cit. on p. 11).

[40] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. «Emerging properties in self-supervised vision transformers». In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2021), pp. 9650–9660 (cit. on p. 11).

[41] Chetan L. Srinidhi, Seung Wook Kim, Fu-Der Chen, and Anne L. Martel. «Self-supervised driven consistency training for annotation efficient histopathology image analysis». In: *Medical Image Analysis* 75 (2022), p. 102256 (cit. on p. 12).

[42] Joseph Boyd, Mykola Liashuha, Eric Deutsch, Nikos Paragios, Stergios Christodoulidis, and Maria Vakalopoulou. «Self-Supervised Representation Learning using Visual Field Expansion on Digital Pathology». In: *arXiv preprint arXiv:2109.03299* (2021) (cit. on p. 12).

[43] Richard J. Chen, Tong Ding, Ming Y. Lu, et al. «Towards a General-Purpose Foundation Model for Computational Pathology». In: *Nature Medicine* 30.3 (2024), pp. 850–862. DOI: 10.1038/s41591-024-02857-3 (cit. on p. 12).

[44] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. «Transformer-based unsupervised contrastive learning for histopathological image classification». In: *Medical Image Analysis* 81 (2022), p. 102559. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2022.102559. URL: https://www.sciencedirect.com/science/article/pii/S1361841522002043 (cit. on p. 12).

[45] Shekoofeh Azizi et al. «Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging». In: *Nature Biomedical Engineering* 7 (2023), pp. 756–779. DOI: 10.1038/s41551-023-01049-7. URL: https://www.nature.com/articles/s41551-023-01049-7 (cit. on p. 12).

[46] Diana Gina Poalelungi, Anca Neagu, Ana Fulga, Marius Neagu, Dana Tutunaru, Aurel Nechita, and Iuliu Fulga. «Revolutionizing Pathology with Artificial Intelligence: Innovations in Immunohistochemistry». In: *Journal of Personalized Medicine* 14.7 (2024), p. 693. DOI: 10.3390/jpm14070693 (cit. on p. 12).

[47] Carlijn M. Lems, Daan Geijs, John-Melle Bokhorst, Maxime Sülter, Leander van Eekelen, and Francesco Ciompi. «Color Deconvolution for Color-Agnostic and Cross-Modality Analysis of Immunohistochemistry Whole-Slide Images with Deep Learning». In: *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI)* (2024). DOI: 10.1109/isbi56570.2024.10635258 (cit. on p. 12).

[48] Pierre-Antoine Bannier, Loïc Herpin, Rémy Dubois, Lydwine Van Praet, Charles Maussion, et al. «Deep Learning Model for Automated Quantification of HER2 Expression in Invasive Breast Cancers from Immunohistochemical Whole Slide Images». In: *Cancer Research* 84.9 (2024), Abstract nr PO2-07–05. DOI: 10.1158/1538-7445.sabcs23-po2-07-05 (cit. on p. 13).

[49] Mark A.J. Gorris, Evgenia Martynova, M. W. D. Sweep, et al. «Multiplex Immunohistochemical Analysis of the Spatial Immune Cell Landscape of the Tumor Microenvironment». In: *Journal of Visualized Experiments* 65717 (2023). DOI: 10.3791/65717 (cit. on p. 13).

[50] Kai Zou, Suwan Zhu, Si-chuan Wang, Ziqian Wang, and Fan Yang. «A Pixel-Enhanced Deep Convolutional Neural Network for Classifying Immunohistochemistry Images». In: *Proceedings of SPIE Medical Imaging* (2023), Paper 2681112. DOI: 10.1117/12.2681112 (cit. on p. 13).

[51] Chanita Panwoon, Wunchana Seubwai, Malinee Thanee, and Sakkarn Sangkhamanon. «Identification of Novel Biomarkers to Distinguish Clear Cell and Non-Clear Cell Renal Cell Carcinoma Using Bioinformatics and Machine Learning». In: *PLOS ONE* TBD (2024). DOI: 10.1371/journal.pone.0305252 (cit. on p. 13).

[52] Kouther Noureddine, Paul Gallagher, Martial Guillaud, and Calum MacAulay. «Investigating Intra-Tumor Heterogeneity Using Multiplexed Immunohistochemistry & Deep Learning: A New Approach to Spatially Map the Tumor Microenvironment». In: *Proceedings of the American Association for Cancer Research Annual Meeting*. Vol. 82. 12 Suppl. 2022, Abstract nr 1719. DOI: 10.1158/1538-7445.am2022-1719 (cit. on p. 13).

[53] Paul Acosta, Vandana Panwar, Vipul Nataraj Jarmale, et al. «Intratumoral Resolution of Driver Gene Mutation Heterogeneity in Renal Cancer Using Deep Learning». In: *Cancer Research* 82 (2022). DOI: 10.1158/0008-5472.CAN-21-2318 (cit. on p. 13).

[54] University of Washington News. *In New Nature Paper, Allen School Researchers Slide Into the Future of Cancer Diagnosis with Groundbreaking AI Model for Digital Pathology*. https://news.cs.washington.edu/2024/05/29/in-new-nature-paper-allen-school-researchers-slide-into-the-future-of-cancer-diagnosis-with-groundbreaking-ai-model-for-digital-pathology/. Accessed: 30 September 2024. 2024 (cit. on p. 17).

[55] Navid Farahani, Anil V Parwani, and Liron Pantanowitz. «Whole slide imaging in pathology: advantages, limitations, and emerging perspectives». In: *Pathology and Laboratory Medicine International* 7 (2015), pp. 23–33 (cit. on p. 18).

[56] Nikki S Vyas et al. «Comparing whole slide digital images versus traditional glass slides in the detection of common microscopic features seen in dermatitis». In: *Journal of Pathology Informatics* 7 (2016), pp. 38–45 (cit. on p. 18).

[57] Mark Zarella, Douglas Bowman, Famke Aeffner, Navid Farahani, Albert Xthona, Syeda Absar, Anil Parwani, Marilyn Bui, and Douglas Hartman. «A Practical Guide to Whole Slide Imaging: A White Paper From the Digital Pathology Association». In: *Archives of pathology and laboratory medicine* 143 (Oct. 2018). DOI: 10.5858/arpa.2018\-0343\-RA (cit. on p. 18).

[58] John Griffin and Darren Treanor. «Digital pathology in clinical use: Where are we now and what is holding us back?» In: *Histopathology* 67.2 (2015), pp. 169–181 (cit. on p. 18).

[59] Michael G. Hanna et al. «Whole slide imaging equivalency and efficiency study». In: *Journal of Pathology Informatics* 11 (2020), pp. 1–10 (cit. on p. 19).

[60] Cyllene V Hedvat. «Digital microscopy: past, present, and future». In: *Archives of Pathology & Laboratory Medicine* 134.11 (2010), pp. 1666–1670 (cit. on p. 19).

[61] Sanjay Mukhopadhyay et al. «Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: a multicenter blinded randomized noninferiority study of 1992 cases (pivotal study)». In: *The Lancet Oncology* 19.4 (2018), pp. 329–336 (cit. on p. 19).

[62] Nehal Atallah, Michael Toss, Clare Verrill, Manuel Salto–Tellez, David Snead, and Emad Rakha. «Potential quality pitfalls of digitalized whole slide image of breast pathology in routine practice». In: *Modern Pathology* 35 (Dec. 2021). DOI: 10.1038/s41379-021-01000-8 (cit. on p. 20).

[63] M. Kasthuri. «Performance analysis of gradient-based image edge detection». In: *International Journal of Health Sciences* 6.S5 (2022), p. 9134. DOI: 10.53730/ijhs.v6ns5.9134 (cit. on p. 28).

[64] Zhu Wei, Bai Junqi, Zhai Shangli, Du Hanyu, Miao Feng, Wang Shoufeng, Liu Wen, Liu Yu, and Wang Xingpeng. «Neighborhood gradient-based infrared weak and small target detection method under complex background». Patent. 2020 (cit. on p. 28).

[65] Fancy Joy and V. Vijayakumar. «A Novel User-Friendly Application for Foreground Detection with Post-Processing in Surveillance Video Analytics». In: *International Journal of Electrical & Electronics Research* (2022). DOI: 10.37391/ijeer.100477 (cit. on p. 28).

[66] Radhamadhab Dalai and Kishore Kumar Senapati. «An Innovative Tree Gradient Boosting Method Based Image Object Detection from a Uniform Background Scene». In: *Pattern Recognition and Image Analysis* 28.2 (2018), pp. 287–295. DOI: 10.1134/S1054661818020165 (cit. on p. 28).

[67] Umut Avci. «Handling Imbalanced Data in Predictive Maintenance: A Resampling-Based Approach». In: *Proceedings of the IEEE HORA Conference*. 2023. DOI: 10.1109/hora58378.2023.10156799 (cit. on p. 29).

[68] Carlos T. Calafate I. de Zarzà J. de Curtò. «Optimizing Neural Networks for Imbalanced Data». In: *Electronics* (2023). DOI: 10.3390/electronics12122674 (cit. on p. 29).

[69] Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. Springer, 2013 (cit. on p. 30).

[70] Lisa M. McShane, Douglas G. Altman, Willi Sauerbrei, Sheila E. Taube, Massimo Gion, and Gary M. Clark. «Assessment of Immunohistochemistry for Biomarker Research in Oncology». In: *Journal of Clinical Oncology* 23.4 (2000), pp. 725–734 (cit. on p. 31).

[71] Donald B. Rubin. «Inference and Missing Data». In: *Biometrika* 63.3 (1976), pp. 581–592 (cit. on p. 31).

[72] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009 (cit. on p. 31).

[73] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. «SMOTE: Synthetic Minority Over-sampling Technique». In: *Journal of Artificial Intelligence Research*. 2002, pp. 321–357 (cit. on p. 32).

[74] Ivan Tomek. «Tomek Links for Pattern Classification». In: *IEEE Transactions on Systems, Man, and Cybernetics* 6.6 (1976), pp. 769–772 (cit. on p. 32).

[75] Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard. «A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data». In: *SIGKDD Explorations* 6.1 (2004), pp. 20–29 (cit. on p. 32).

[76] Karen Simonyan and Andrew Zisserman. «Very deep convolutional networks for large-scale image recognition». In: *3rd International Conference on Learning Representations, ICLR 2015*. 2015. URL: https://arxiv.org/abs/1409.1556 (cit. on p. 33).

[77] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. «Deep residual learning for image recognition». In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778 (cit. on p. 33).

[78] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. «Densely connected convolutional networks». In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708 (cit. on p. 33).

[79] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. «Going deeper with convolutions». In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9 (cit. on p. 33).

[80] François Chollet. «Xception: Deep learning with depthwise separable convolutions». In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1251–1258 (cit. on p. 33).

[81] Zhuang Liu et al. «A ConvNet for the 2020s». In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2022) (cit. on p. 33).

[82] Yarin Gal and Zoubin Ghahramani. «Dropout as a Bayesian approximation: Representing model uncertainty in deep learning». In: *Proceedings of the 33rd International Conference on Machine Learning*. 2016, pp. 1050–1059. URL: https://arxiv.org/abs/1506.02142 (cit. on pp. 34, 39).

[83] Daily Milanés Hermosilla, Rafael Trujillo Codorniu, René López-Baracaldo, Roberto Sagaro Zamora, Denis Delisle Rodriguez, John Villarejo Mayor, and José Nuñez Alvarez. «Monte Carlo Dropout for Uncertainty Estimation and Motor Imagery Classification». In: *Sensors* 21 (Oct. 2021), p. 7241. DOI: 10.3390/s21217241 (cit. on pp. 40, 47).

[84] Anil Bhattacharyya. «On a Measure of Divergence between Two Statistical Populations Defined by their Probability Distributions». In: *Bulletin of the Calcutta Mathematical Society* 35 (1943), pp. 99–109 (cit. on p. 40).

[85] Leo Breiman. «Random Forests». In: *Machine learning* 45.1 (2001), pp. 5–32 (cit. on p. 42).

[86] Jerome H. Friedman. «Greedy Function Approximation: A Gradient Boosting Machine». In: *Annals of Statistics* 29.5 (2001), pp. 1189–1232 (cit. on p. 42).

[87] Pierre Geurts, Damien Ernst, and Louis Wehenkel. «Extremely Randomized Trees». In: *Machine Learning* 63.1 (2006), pp. 3–42 (cit. on p. 42).

[88] Tianqi Chen and Carlos Guestrin. «XGBoost: A Scalable Tree Boosting System». In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2016, pp. 785–794 (cit. on pp. 42, 43).

[89] Mostafa Karami. *Machine Learning Algorithms for Radiogenomics: Application to Prediction of the MGMT Promoter Methylation Status in mpMRI Scans*. Oct. 2022. DOI: 10.13140/RG.2.2.34778.31689 (cit. on p. 44).