



UNIVERSITY OF TRENTO

PHD PROGRAM IN BIOMOLECULAR SCIENCES

DEPARTMENT OF CELLULAR, COMPUTATIONAL AND
INTEGRATIVE BIOLOGY – CIBIO

37th Cycle

**Exploiting Interpretability in
AI-Driven Medical Imaging:
Strategies for Transparency and
Clinical Integration in
Computational Pathology**

Tutor
Giuseppe Jurman

Advisor
Enrico Domenici

Ph.D. Thesis of
Matteo Pozzi

Academic Year: 2023/2024

Declaration

I Matteo Pozzi confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Trento, 12/08/2024

Abstract

The integration of artificial intelligence (AI) into medical imaging has the potential to revolutionize diagnostic processes and patient care. However, the "black-box" nature of some deep learning models poses significant challenges for clinical adoption, as medical professionals require transparency and interpretability to trust and effectively utilize these tools. This thesis aims to enhance the interpretability and explainability of deep learning models in medical imaging by incorporating both post-hoc explainability techniques and interpretable-by-design methods, with a particular focus on computational pathology.

The research begins with a comprehensive review of the automation in pathology, highlighting how automated systems can improve the robustness and consistency of pathology practices. This review underscores the importance of automation as a foundation for integrating AI applications into clinical workflows, enabling the handling of large data volumes with high precision and laying the groundwork for reliable AI model development.

Addressing the challenge of data scarcity—a significant barrier in training effective AI models—the thesis investigates the use of synthetic data generation in computational pathology. Synthetic histopathological images are created not only to augment existing datasets but also to evaluate their potential as substitutes for original datasets, thereby easing data sharing constraints. XAI techniques are applied to assess the usability of these synthetic images in training AI models. The findings reveal that, when properly validated, synthetic data can mitigate limitations due to data scarcity and enhance model performance in detecting and classifying pathological features.

The research further delves into the quantification of tumor-infiltrating lymphocytes (TILs) by leveraging interpretable-by-design definition of the research. Rather than developing new segmentation models, the study utilizes existing segmentation outputs to calculate TIL scores in accordance with international guidelines. The design ensures that the segmentation masks used are accurate and interpretable, aligning AI-generated outputs with standardized clinical scoring systems. This method enhances the reliability and clinical relevance of the findings, providing valuable insights into the tumor microenvironment crucial for prognosis and personalized treatment strategies.

Collectively, the studies presented in this thesis contribute to a coherent strategy aimed at advancing computational pathology through the integration of interpretability and explainability in deep learning models. By systematically addressing key challenges—automation, explainability, data scarcity, and standardization—the research works toward the overarching goal of enhancing diagnostic accuracy and efficiency in medical imaging.

In conclusion, the thesis demonstrates that integrating interpretability and ex-

plainable techniques into deep learning models is both feasible and beneficial for clinical practice. The use of post-hoc explainability methods and interpretable-by-design approaches not only improves model transparency but also fosters trust among medical professionals. This work lays a foundation for future AI applications in medicine that prioritize transparency, reliability, and clinical alignment, ultimately contributing to improved patient outcomes and the advancement of personalized healthcare.

Contents

1	Introduction	2
1.1	Artificial intelligence in medicine	2
1.2	XAI: Approaches to uncover the black-box	3
1.3	Computational pathology	5
1.3.1	AI in Computational Pathology	6
1.3.2	Biomarkers and Computational Pathology	8
1.3.3	Latest trends and future direction	9
1.4	Thesis structure	11
1.5	Thesis aim	13
2	Automation in digital pathology	15
2.1	Summary	15
2.2	Abstract	16
2.3	Introduction	16
2.4	Pre-analytical processes	17
2.4.1	Tracking samples	17
2.4.2	Tissue processing	19
2.4.3	Automation in tissue embedding	19
2.4.4	Automatic microtome	20
2.4.5	Automation in slide staining and coverslipping	21
2.4.6	Collaborative robots	22
2.5	Analytical processes	23
2.5.1	Digital and computational pathology	23
2.5.2	Structured synoptic reporting	26
2.6	Post-analytical processes	27
2.6.1	Storage and biobanking of tissue and specimens	27
2.6.2	Digital imaging archiving	28
2.7	Conclusion and Future Directions	29
2.8	Disclaimer	29
2.9	Declarations	30
3	Explainability in endoscopic imaging	31
3.1	Summary	31
3.2	Abstract	32
3.3	Introduction	32
3.4	Methods	33
3.4.1	Outcome definition	34
3.4.2	Dataset description	34

3.4.3	Data preprocessing	35
3.4.4	Deep learning architecture	36
3.4.5	Differential learning rates	36
3.4.6	Ensemble learning	37
3.4.7	Performance evaluation metrics	37
3.4.8	Model interpretation	38
3.4.9	Implementation	38
3.5	Results	38
3.5.1	Classification	38
3.5.2	Interpretability	40
3.6	Discussion	42
3.7	Conclusions	43
3.8	List of abbreviations	43
3.9	Declarations	43
3.10	Ethics approval and consent to participate	43
3.11	Consent to publish	43
3.12	Availability of data and materials	44
3.13	Competing interests	44
3.14	Funding	44
3.15	Authors' contributions	44
3.16	Acknowledgements	44
3.17	Figures	44
3.18	Tables	45
4	Synthetic data in digital pathology	47
4.1	Summary	47
4.2	Abstract	48
4.3	Introduction	49
4.4	Materials and Methods	52
4.4.1	Dataset	52
4.4.2	Preprocessing pipeline	52
4.4.3	Denoising Diffusion Probabilistic Model	54
4.4.4	Evaluations	55
4.4.5	Implementation	59
4.5	Results	59
4.5.1	Measuring the Similarity Between Generated and Real Images Reveals Generated Data Resembles Real Images	60
4.5.2	The Generated Data Exhibit Performance Comparable to Real Data in A Deep Learning-based Task	62
4.5.3	Explainability-mediated analysis further reveals insights about the features learned from both datasets	63
4.5.4	The Generated Data is Biologically Realistic in the Eyes of Experts	65
4.6	Discussion	67
4.7	Data and Code Availability	69
4.8	Additional Informations	69
4.8.1	Fundings	69

5	Investigating TILs in ductal carcinoma in situ	70
5.1	Summary	70
5.2	abstract	71
5.3	Introduction	71
5.4	Materials and Methods	73
5.4.1	Data collection	73
5.4.2	DCIS segmentation	73
5.4.3	TILs detection	74
5.4.4	Pathologist TILs score	75
5.4.5	Automatic TILs score computation	75
5.4.6	Statistical Analysis	75
5.5	Results and Discussion	76
5.6	Acknowledgments	78
6	Discussion	79
6.1	Conclusion and Future Directions	80

Chapter 1

Introduction

1.1 Artificial intelligence in medicine

The evolution of medical practice from the early 20th century to the present has been marked by significant advancements and milestones that have transformed patient care. As we navigate through this journey, it becomes evident that each era has contributed uniquely to the way healthcare is delivered, ultimately enhancing the quality of life for countless individuals.

In the early 20th century, medical practice was largely characterized by rudimentary diagnostic tools and limited treatment options. Physicians relied on basic observational skills and a limited understanding of human biology to diagnose and treat illnesses. The 1920s marked a pivotal moment in this evolution with the introduction of antibiotics, which revolutionized the treatment of bacterial infections. This innovation drastically reduced mortality rates and changed the landscape of infectious disease management. Furthermore, the development of vaccines significantly enhanced public health, leading to the near-eradication of diseases such as polio and smallpox. These breakthroughs not only saved lives but also laid the groundwork for future advancements in preventive medicine [1, 2].

As we moved into the second half of the 20th century, advances in physics and biomedical engineering allowed physicians to non-invasively examine the body's structures and functions through technologies such as ultrasound, computerized axial tomography (CAT), nuclear magnetic resonance (NMR), and later, magnetic resonance imaging (MRI). These imaging modalities provided clinicians with unprecedented insights into human anatomy and pathology, enabling more accurate diagnoses and tailored treatment plans. The ability to visualize internal structures without invasive procedures transformed surgical practices and improved patient outcomes significantly.

Another significant wave of advancement came with breakthroughs in understanding the genetic code, building on Watson and Crick's iconic 1953 paper on DNA's molecular structure. This work paved the way for the field of genomics, which has revolutionized our understanding of diseases at molecular level. The Human Genome Project, started in 1990, aimed to map all human genes and has since provided invaluable insights into genetic disorders, genetics and genomics of cancer and potential therapies [3]. Despite these advancements, the scientific community has yet to fully exploit these technological innovations, leading to increasing efforts in decoding complex patterns hidden within genomic data [4].

The digital era has guided healthcare into a new age of transformation, characterized by rapid advancements in technology that have fundamentally altered how medical professionals interact with patients and manage health information [5]. The integration of digital technologies into healthcare systems has enhanced accessibility, efficiency, and patient engagement. Electronic health records (EHRs) have replaced paper-based documentation, allowing for seamless sharing of patient information among healthcare providers. This shift not only improves coordination of care but also reduces the risk of errors associated with manual record-keeping.

Telemedicine has emerged as a significant innovation within this digital landscape, enabling patients to consult with healthcare providers remotely through video calls or mobile applications. This accessibility is particularly beneficial for individuals in rural or underserved areas who may have limited access to specialized care. Additionally, wearable devices and mobile health applications empower patients to monitor their health metrics in real-time, fostering a proactive approach to personal health management.

Moreover, big data analytics plays a crucial role in transforming healthcare delivery by allowing providers to analyze vast amounts of patient data for trends and insights that can inform clinical decisions. Predictive analytics can identify at-risk populations and facilitate early interventions, ultimately improving health outcomes while reducing costs [6].

Artificial intelligence (AI) represents one of the most groundbreaking advancements in healthcare technology today. By harnessing machine learning algorithms and vast datasets created by the previous technological waves, AI is revolutionizing various aspects of patient care, from diagnosis to treatment planning and beyond. AI-powered diagnostic tools can analyze medical images with remarkable accuracy, often matching or exceeding human performance in identifying conditions such as tumors metastasis or bone fractures [7, 8].

In addition to imaging analysis, AI is being utilized to develop personalized treatment plans based on individual patient data. Machine learning models can predict how patients will respond to specific treatments by analyzing genetic information alongside clinical history, thereby enabling tailored therapies that enhance efficacy while minimizing side effects. [9, 10]

As we continue to explore the potential of AI in healthcare, it is essential to address ethical considerations surrounding data privacy and algorithmic bias. Ensuring equitable access to AI-driven innovations will be crucial as we strive toward a future where technology enhances rather than hinders patient care.

1.2 XAI: Approaches to uncover the black-box

The integration of AI in healthcare care is not merely a technological advancement but a paradigm shift that demands rigorous ethical and interpretative frameworks. The rare implementation of AI systems in daily clinical practice might seem in conflict with their good performance and the potential benefits offered by their exploitation in the healthcare domain. Obstacles to the clinical translation of AI systems include the need of large annotated datasets to train the algorithms, the ability to generalize to data from the clinical practice (which are often characterized by high variability) and, importantly, the need of understanding and interpreting the

system predictions [11, 12] Explainable AI (XAI) has emerged as a critical approach to address the "black box" nature of traditional machine learning models.

In healthcare, the consequences of AI-driven decisions are profound and directly impact human lives. Unlike other domains, medical AI systems must provide not just accurate predictions but also comprehensible reasoning behind those predictions. This necessitates developing AI models that are transparent, interpretable, and aligned with clinical reasoning. Several key motivations underscore the importance of explainable AI in healthcare:

1. **Clinical Trust and Adoption:** Clinicians require a clear understanding of how AI makes specific recommendations to integrate these technologies into their decision-making processes.
2. **Ethical Accountability:** Transparent AI models enable proper attribution of responsibility and facilitate comprehensive audit trails for medical decisions.
3. **Regulatory Compliance:** Emerging healthcare regulations increasingly mandate interpretability and transparency in AI-driven medical technologies.
4. **Bias Mitigation:** Explainable models allow for more effective identification and mitigation of potential algorithmic biases that could perpetuate healthcare disparities.
5. **Patient Engagement:** Providing patients with comprehensible explanations of AI-assisted diagnoses can improve their understanding and participation in their healthcare journey.

The wide range of XAI techniques varies significantly in many aspects, reflecting the diversity of possible explanations. These techniques can be categorized according to several criteria, which will be discussed in the following sections and are summarized in Figure 1.1.

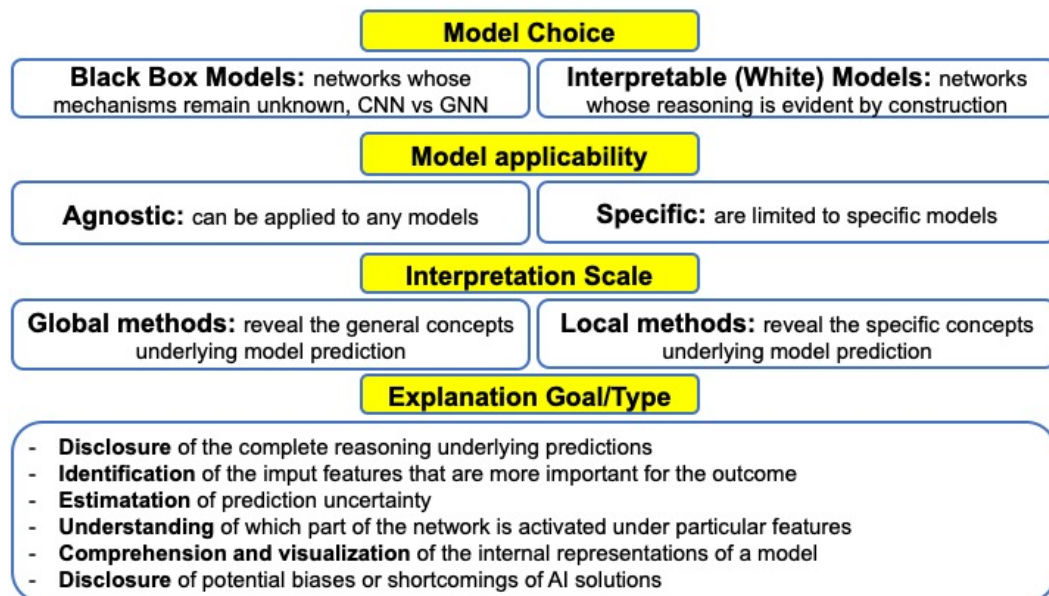


Figure 1.1: XAI classification. XAI techniques are classified according to multiple criteria.

A key distinction in the explainability of ML and DL solutions lies in the choice of the model. On one hand, there are black-box models, characterized by their complexity and deep structures, whose internal mechanisms are not transparent. Interpretability for black-box models can be achieved through post hoc methods, applied after training, which act as an additional layer of explainability. Examples of such methods include feature visualization techniques like SHapley Additive ExPlanations (SHAP) [13], Layer-wise Relevance Propagation (LRP) [14], Class Activation Maps (CAM) [15], and their gradient-based variants such as Grad-CAM [16] and Grad-CAM++ [17].

On the other hand, an alternative approach involves interpretable models (also known as white models, white box, or glass box). These models are inherently transparent, with reasoning that is clear by design. They do not require supplementary tools for explanation, and their outputs are easily interpretable. Decision trees, ProtoPNet [18], and additive MIL [19] are examples of this category.

Another criterion for classifying XAI methods is the scale of the interpretation they provide. Global XAI methods offer explanations about the entire model, revealing the general logic underlying its predictions. These methods aim to uncover the overarching concepts driving a model's outputs. In contrast, local XAI methods focus on specific predictions, identifying which input features are relevant for a given outcome without generalizing to other instances. Such methods typically highlight the features influencing a specific prediction but do not explain how these features are used or apply their findings to other cases. Recently, the concept of glocal explainability was introduced for methods that bridge the gap, demonstrating both the concepts learned by a model and how they contribute to specific outcomes [20].

Alternatively, XAI techniques can be classified as either model-specific - applicable to particular types of models (e.g., CNNs) - or model-agnostic - which can be used with any algorithm. With the growing efficiency of DL solutions, model-agnostic techniques are gaining prominence as they provide a standardized approach to comparing different models.

Finally, XAI methods can be differentiated by the objectives and types of explanation they offer. These include, among others, elucidating the complete reasoning behind predictions, identifying the most influential input features, estimating prediction uncertainties, understanding which parts of a network are activated by specific features, visualizing internal representations, and revealing potential biases or shortcomings in AI solutions. Each of these explanation goals addresses different aspects of the reasoning of a model, contributing to greater transparency.

Given these diverse goals, it is crucial to identify the target explanation to select the most appropriate XAI technique. For example, physicians may prioritize understanding why a specific output was generated to support or challenge diagnoses and prognoses. Developers, meanwhile, may focus on understanding internal mechanisms to enhance model performance or uncover potential biases.

1.3 Computational pathology

Traditional pathology involves a series of meticulous steps in diagnosing diseases, primarily through the examination of tissue samples. The process typically begins with the collection of tissue specimens, which are then processed in a laboratory.

These specimens are embedded in paraffin, sliced into thin sections, and stained to highlight specific features. Pathologists then examine these stained slides under a microscope to identify any abnormalities or diseases. This manual process is labor-intensive and requires highly trained specialists to interpret the slides accurately. These steps can easily lead to errors, making it crucial to standardize every stage of the process.

Healthcare workforce shortages, particularly in pathology, represent a significant challenge in Europe. Among European countries, Germany faces an especially acute shortage, with one of the lowest densities of pathologists—one pathologist per 47,989 inhabitants—compared to the European average of one per 32,018 inhabitants [21]. This shortage is part of a broader trend of healthcare professional shortages across Europe, impacting various medical disciplines and regions.

In this context, digitization has emerged as a promising solution to alleviate some of these challenges by enhancing efficiency, productivity, and accuracy [22, 23]. Automation and digitization in clinical pathology are transformative, cutting across multiple stages of the traditional workflow. These advancements include optimizing the tracking and management of tissue specimens, automating processes such as tissue embedding, cutting, and staining, as well as digitizing and securely storing digital slides.

Digital slides, also known as whole slide images (WSI), are high-resolution digital representations of traditional glass slides, offering a new paradigm for clinical pathology. This digitization greatly enhances information accessibility by making slides available for viewing, sharing, and analysis on digital platforms, enabling seamless collaboration among experts across geographical boundaries. Such accessibility is particularly transformative for healthcare systems in underdeveloped countries or smaller, rural areas, where resources and expertise are often limited [24]. By bridging these gaps, digital slides facilitate consultations and second opinions, thereby improving diagnostic accuracy and patient outcomes.

Beyond improving accessibility, WSI also unlock unprecedented opportunities for technological innovation, particularly through the integration of artificial intelligence (AI). The ability to digitize pathology data creates a vast repository of structured and analyzable information, laying the groundwork for AI applications. Advanced algorithms, whether based on traditional methods or modern neural networks, are increasingly being developed to analyze digital slides with remarkable speed and precision. These AI tools can assist pathologists in various tasks, from identifying specific cellular features to predicting disease prognosis, thus enhancing efficiency and reducing diagnostic variability [23].

1.3.1 AI in Computational Pathology

The integration of artificial intelligence (AI) in digital pathology is transforming diagnostic practices, enhancing accuracy and efficiency [25]. Several AI-based tools have received regulatory approval for clinical use, particularly in the diagnosis of various cancers. Some notable and not exhaustive examples are:

1. PathAI: PathAI has developed machine learning algorithms that assist pathologists in diagnosing diseases from histopathological images. Their tools have been validated for use in breast cancer diagnosis and are designed to improve diagnostic accuracy by reducing variability among pathologists.

2. Visiopharm: A leading provider of AI-driven precision pathology software for research and diagnostics. It offers tools that enable accurate tissue-based research and clinical decision support, with nine IVDR-cleared diagnostic algorithms for EU customers¹.
3. Ibex Medical Analytics: Known for its deep learning-based cancer diagnostic support software, Ibex Medical Analytics is currently the most widely deployed artificial intelligence platform in pathology¹.
4. Paige Prostate Detect: This tool has received the first-ever FDA clearance for a clinical AI application in pathology. It is designed to assist in the detection and diagnosis of prostate cancer, enhancing the capabilities of pathologists².
5. Aignostics: Focuses on pioneering computational pathology for pharmaceutical research, clinical trials, and CDx development. Their AI tools are designed to provide personalized insights for clinicians and strengthen healthcare delivery¹.
6. Deep Lens: Acquired by Paradigm, Deep Lens uses AI to match cancer patients with clinical trials. Their VIPER platform analyzes clinical data and pathology reports for a deeper understanding of a patient's cancer³.
7. Deep 6 AI: Utilizes natural language processing (NLP) to streamline and enhance patient recruitment for clinical trials. It analyzes unstructured medical data, including electronic health records (EHRs) and pathology reports, to identify potential candidates for trials³.

Research in computational pathology has been deeply intertwined with advancements in the general field of deep learning. However, unlike traditional computer vision tasks where fully annotated datasets are often feasible, whole slide images (WSIs) present unique challenges due to their sheer size, complexity, and the labor-intensive nature of detailed annotation. Annotating WSIs at the pixel or region level for large datasets is not only time-consuming but often impractical, given the level of expertise required and the variability in pathological assessments.

This realization has shifted the focus of the field towards more data- and labor-efficient frameworks, such as weakly supervised learning (WSL), multiple instance learning (MIL), and self-supervised learning (SSL). These methodologies leverage limited or indirect labels, enabling the effective use of large datasets without exhaustive annotations.

Weakly Supervised Learning (WSL)

Weakly supervised learning involves training models using labels that are coarse, incomplete, or noisy. In the context of computational pathology, WSL is often employed when only slide-level labels (e.g., the presence or absence of cancer) are available, rather than detailed pixel- or patch-level annotations [26]. Models learn to infer patterns that correlate with these high-level labels while implicitly localizing relevant features within the image. Techniques such as attention mechanisms or class activation maps (CAMs) are frequently used to enhance interpretability and guide the model toward regions of interest.

Multiple Instance Learning (MIL)

Multiple instance learning is a specific framework within weakly supervised learning that addresses situations where labels are associated with a set of instances (a "bag") rather than individual instances. In computational pathology, a WSI is often divided into smaller patches, with the entire slide labeled as positive (e.g., containing cancer) or negative (e.g., benign). The model processes the patches and learns to identify which ones contribute to the positive label. MIL has been successfully used for tasks such as tumor detection and grading, where the presence of a few critical patches determines the overall diagnosis. Among the earliest works to successfully apply MIL to computational pathology is the study by Campanella et al. [27], which demonstrated the efficacy of MIL for cancer detection across various tissue types using weakly labeled WSIs. This landmark paper highlighted the scalability of MIL and its potential to handle large-scale datasets.

Building on this foundation, the Mahmood lab introduced the CLAM [28] (Clustered Attention MIL) framework, which extended traditional MIL by incorporating attention mechanisms. CLAM improved the interpretability and performance of MIL-based models by clustering similar instances and dynamically weighting their contributions to the final prediction. This approach enabled more accurate identification of relevant regions within WSIs, setting a new standard in the field.

Furthermore, innovations in MIL architecture, such as transformer-based MIL models and hybrid frameworks that combine CNNs with graph-based methods, have significantly improved the performance and generalization of MIL approaches [29, 30, 31]. These methods leverage both spatial and contextual information, offering a deeper understanding of the intricate relationships within WSIs.

Self-Supervised Learning (SSL)

Self-supervised learning is a pretraining technique that leverages unlabeled data by creating surrogate tasks, often called pretext tasks, to learn meaningful feature representations. For example, in computational pathology, models can be trained to predict the spatial arrangement of patches, reconstruct masked regions of an image, or discriminate between transformations applied to images. Once pretrained, these models can be fine-tuned on smaller labeled datasets for specific downstream tasks, such as segmentation or classification.

Paired with state-of-the-art architectures like vision transformers (ViTs) or convolutional neural networks (CNNs), SSL has demonstrated powerful results in computational pathology [32, 31, 30]. It significantly reduces the reliance on large annotated datasets and improves generalization by capturing rich, domain-specific features during pretraining. SSL has also enabled transfer learning in pathology, where models trained on one dataset (e.g., for one cancer type) can be adapted to new datasets or tasks with minimal additional training.

1.3.2 Biomarkers and Computational Pathology

Biomarkers are quantifiable indicators of physiological or pathological states, and span genetic, proteomic, and morphological domains, enabling personalized healthcare strategies and improving patient outcomes [33]. The integration of biomarkers

into medical practice has revolutionized disease prevention, diagnosis, and treatment, serving as critical tools for translating biological insights into clinical action. Their utility arises from their ability to bridge molecular mechanisms with clinical phenotypes, offering insights into disease risk, progression, and treatment efficacy. For example, protein biomarkers like prostate-specific antigen (*PSA*) have long been used for cancer screening, while genetic biomarkers such as *BRCA1* and *BRCA2* mutations inform hereditary cancer risk assessments. Another example, in cardiovascular disease, are biomarkers like natriuretic peptides and cardiac troponins, which proved to be pivotal in diagnosing and prognosticating heart failure and myocardial infarction, respectively [34]. Biomarkers can be categorized according to their temporal or diagnostic applications [35]. Prognostic biomarkers are used to predict the likelihood of a clinical event, while predictive biomarkers are used to identify individuals who are more likely than others to experience a favorable or unfavorable effect from exposure to a medical treatment. Other relevant category of biomarkers are the diagnostic, used to detect or confirm the presence of a disease or medical condition, and the risks biomarkers, identified and approved for their effectiveness in predicting an individual's likelihood of developing a particular disease or medical condition in the future. In recent years, the field of computational pathology has emerged as a transformative force in biomarker discovery and application. Computational pathology leverages artificial intelligence and deep learning technologies to extract clinically relevant information from routine histology images, which can enhance the accuracy and efficiency of clinical decision-making, by transforming subjective morphological assessments into reproducible biomarkers. Examples are represented by AI tools which assist in the precise evaluation of immune biomarkers like PD-L1 and tumor-infiltrating lymphocytes (TILs), improving consistency and reducing inter observer variation in pathology clinics, ultimately helping clinicians in patient's selection for immunotherapy. AI algorithm are usually deployed with the aim of quantifying clinically relevant biomarkers, in a variety of ways: from counting to density estimation, from shape analysis to network and topological modelling. Using a variety of different algorithmic approaches AI is playing a crucial role in standardizing the evaluation of already known biomarkers but also in discovering new candidates.

1.3.3 Latest trends and future direction

Foundation models (FMs) have emerged as a transformative force in computational pathology, significantly altering the diagnostic workflow of pathologists. These models, trained on vast and diverse datasets, have shown remarkable capabilities in generalizing to various tasks without the need for labeled data. For instance, models like Virchow, trained on over 100,000 diagnostic histopathological slides, have demonstrated superior performance across multiple clinically relevant tasks [36]. The integration of these models has introduced a new paradigm in the field, enabling the analysis of large, unlabeled whole-slide images and providing new insights into cancer diagnosis and treatment [36, 37, 38, 39, 40, 41].

The convergence of data from multiple modalities—such as histopathology, molecular profiling (e.g., genomics, transcriptomics), and clinical metadata—has revolutionized the diagnostic process. AI models that can integrate these diverse data streams provide a more holistic view of a patient's condition, allowing for more

precise and personalized diagnostic insights. For instance, combining genomic mutation data with histological patterns can refine cancer subtype classification and guide therapeutic strategies. Multi-modal models represent a significant leap toward comprehensive pathology workflows that unify molecular and morphological insights [42, 43].

Large Language Models (LLMs) have the potential to revolutionize the way pathologists approach disease diagnosis and treatment. These models, trained on extensive datasets encompassing medical literature, pathology reports, and digital imagery, have demonstrated remarkable capabilities in integrating and analyzing multimodal data [42]. The integration of LLMs with computational pathology allows for real-time interaction with pathologists during slide inspection and report creation. This interaction enhances the diagnostic process by providing context-relevant outputs and assisting in the generation of detailed and accurate pathology reports.

All these advancements and approaches rely heavily on vast amounts of diverse and representative data. The creation of large, publicly accessible datasets that aggregate information from multiple institutions across countries is essential for ensuring equitable representation of all patient populations. Such datasets are not only critical for developing AI models that perform well across diverse demographics but also provide a transparent, public alternative to proprietary, closed-source models. Publicly accessible datasets foster collaboration, reproducibility, and trust in AI-driven systems.

A notable European initiative in this direction is the Big picture project [44], which aims to collect and harmonize data from multiple partners across Europe. This initiative represents a significant step toward fostering cross-border collaboration, ensuring that datasets encompass a wide variety of patient populations, imaging techniques, and clinical scenarios. Providing robust support to develop, scale, and sustain such initiatives is crucial. Investments in infrastructure, policy frameworks, and funding are needed to enable these efforts to thrive. By prioritizing public datasets and collaborative platforms, the field can ensure the development of fair, high-performing models without compromising on transparency or equity.

1.4 Thesis structure

This thesis comprises four manuscripts, each addressing distinct yet interconnected challenges in the realm of interpretable AI and its applications to medical imaging. Together, they construct a comprehensive narrative that aligns with the overarching aim of enhancing the transparency, reliability, and clinical integration of AI-driven solutions in computational pathology.

The research spans automation in digital pathology, explainability in medical imaging, the utility of synthetic data in overcoming dataset limitations, and the automating the quantification of tumor-infiltrating lymphocytes (TILs) to compute a standard clinical score. Below, I outline the contributions of each manuscript and provide context for their connections within the broader framework of interpretable AI.

Manuscript 1: Automation in Digital Pathology

The first manuscript explores the state-of-the-art advancements in automation for pathology laboratories, emphasizing pre-analytical, analytical, and post-analytical processes. It highlights how automated systems reduce variability, enhance diagnostic reproducibility, and enable more efficient workflows. My primary contribution involved authoring the chapter on computational pathology, where I detailed how digitalization has paved the way for leveraging AI and computer-aided tools to enhance diagnostics and workflows within the pathology domain.

Manuscript 2: Explainability in Endoscopic Imaging

The second manuscript investigates the application of explainable AI (XAI) techniques in endoscopic imaging, specifically addressing the "black-box" nature of deep learning models. By employing both post-hoc techniques, the manuscript demonstrates how AI systems can be investigated using gradient-based techniques. My main contributions were the implementation of training in a cross-validation setting and the implementation and interpretation of the explainability techniques and results, ensuring their relevance to clinical workflows and medical practitioners.

Manuscript 3: Synthetic Data in Computational Pathology

The third manuscript tackles the challenge of data scarcity in computational pathology by exploring synthetic data generation. It evaluates the effectiveness of synthetic datasets in training deep learning models and the potential of XAI techniques to validate these datasets. My role involved designing the study and the experiments to validate the results. This included implementing rigorous evaluation metrics and developing a comprehensive framework for synthetic data generation and validation using explainable artificial intelligence. Additionally, I contributed to the interpretation of results, ensuring their alignment with the study's objectives.

Manuscript 4: Investigating Tumor-Infiltrating Lymphocytes (TILs)

The fourth manuscript focuses on the quantification of TILs, a critical biomarker in cancer prognosis and treatment planning. This study leverages interpretable-by-design techniques to align AI-generated outputs with standardized clinical scoring systems. My contributions included training the segmentation network, executing and integrating the object detection results, and subsequently performing the statistical analyses.

While each manuscript addresses a unique aspect of the integration of AI system in the medical domain, they share common goals of enhancing clinical trust, improving diagnostic accuracy, and fostering the integration of AI in medical practice. The

first manuscript, a review, is meant to give a general overview of the digital pathology (DP) field, establishing a foundational understanding. The second manuscript introduces the concept of XAI. The latter two manuscripts represent my primary contributions, where I expand on the concepts and competencies introduced in the initial works, applying them to specific challenges and demonstrating their practical impact. The first manuscript establishes the foundational need for automation, which underpins the methodological advancements explored in the subsequent studies. Manuscripts two and three delve into specific challenges—explainability and data scarcity—offering innovative solutions that build on the automation framework. Finally, the fourth manuscript implements an "interpretable-by-design" strategy, utilizing segmentation and detection outputs to systematically compute a scoring mechanism. This approach replaces the arbitrary and subjective methods currently employed by pathologists, offering a more objective and standardized framework for clinical application.

1.5 Thesis aim

The primary aim of this thesis is to enhance the interpretability and explainability of deep learning models applied to medical imaging, with a focus on computational pathology. As deep learning models become increasingly prevalent in medical diagnostics, their "black-box" nature poses significant challenges for clinical adoption. Medical professionals require transparent and understandable AI systems to trust and effectively utilize them in patient care. This thesis addresses this need by integrating interpretability and explainable techniques into deep learning models, ensuring that the models not only perform accurately but also provide insights into their decision-making processes.

To achieve this aim, the thesis explores both post-hoc explainability techniques and interpretable-by-design methods:

1. **Post-Hoc Explainability Techniques:** These methods are applied after the model has been trained, providing explanations for the model's predictions without altering its internal architecture. The thesis investigates various post-hoc techniques, such as saliency maps, attention mechanisms, and feature importance analyses, to interpret the outputs of deep learning models in medical imaging. By applying these techniques, the research seeks to uncover how models arrive at specific conclusions, enabling clinicians to validate and trust the AI's decisions, as well as validate synthetic data.
2. **Interpretable-by-Design Techniques:** In contrast to post-hoc methods, interpretable-by-design approaches involve building models with inherent transparency. This means designing model architectures that are intrinsically understandable, often by incorporating domain knowledge and constraints from the outset. The thesis develops models that align closely with clinical workflows and standards, such as using segmentation outputs that adhere to international guidelines for tumor-infiltrating lymphocytes (TILs) assessment. By integrating interpretability into the model design, these methods aim to produce AI systems whose operations are transparent and whose outputs are directly meaningful to medical practitioners.

By combining these two approaches, the thesis aims to create deep learning models that are both powerful and accessible to clinicians. The research seeks to demonstrate that enhancing interpretability does not necessarily come at the expense of model performance. Instead, it can lead to more robust and clinically relevant AI tools. The ultimate goal is to facilitate the integration of AI into medical practice by overcoming one of the key barriers to adoption: the lack of transparency in AI decision-making processes.

Through a series of studies and applications, the thesis examines the effectiveness of these techniques in real-world scenarios, such as disease classification, synthetic data generation, and quantification of pathological features. By addressing challenges like data scarcity and the need for standardization, the research contributes to the development of AI models that are not only accurate but also trustworthy and aligned with clinical needs.

In summary, the thesis aims to bridge the gap between advanced AI technologies and practical medical applications by focusing on interpretability and explainability.

By doing so, it aspires to enhance the collaboration between AI systems and medical professionals, ultimately improving patient outcomes through more informed and reliable diagnostics.

Chapter 2

Automation in digital pathology

2.1 Summary

DOI: 10.1007/s00428-023-03637-z

Introduction The paper discusses the pressing need for automation in pathology laboratories to standardize practices, reduce manual errors, and enhance diagnostic precision. Traditional manual processes, such as tissue preparation and slide analysis, are labor-intensive and prone to variability. The integration of advanced technologies and automation aims to address these challenges while improving overall efficiency.

Methods The study categorizes automation in pathology into three primary phases:

1. Pre-analytical Phase: Focuses on specimen tracking and preparation, leveraging barcode systems, robotic arms, and automated tissue processors to streamline workflows.
2. Analytical Phase: Explores digital pathology solutions, including whole-slide imaging (WSI) and computational algorithms for image analysis, classification, and feature extraction.
3. Post-analytical Phase: Investigates digital storage systems for tissue samples and digital slides, ensuring long-term accessibility and data integrity through automated biobanking and imaging archives.

Results Automation significantly enhances diagnostic accuracy and reproducibility. Examples include automated tissue processors reducing variability, while robotic systems improve sample handling. Whole-slide imaging combined with AI algorithms facilitates efficient and accurate pathology workflows. Additionally, automated biobanking systems improve specimen traceability and retrieval efficiency.

Discussion The study highlights that while automation provides transformative benefits, challenges such as high costs, privacy concerns, and the need for regulatory frameworks persist. Future directions emphasize integrating AI-driven insights, machine learning models, and improved interoperability between digital systems to advance pathology practices further.

Cutting-edge technology and automation in the pathology laboratory

Enrico Munari¹, Aldo Scarpa², Luca Cima³, Matteo Pozzi^{4,5}, Fabio Pagni⁶,
Francesco Vasuri⁷, Stefano Marletta^{2,8}, Angelo Paolo Dei Tos⁹, Albino Eccher¹⁰

¹ Pathology Unit, Department of Molecular and Translational Medicine, University of
Brescia, Piazza Del Mercato, 15, 25121 Brescia, BS, Italy.

² Pathology Unit, Department of Diagnostics and Public Health, University of Verona,
Verona, Italy.

³ Pathology Unit, Department of Laboratory Medicine, Santa Chiara Hospital, APSS,
Trento, Italy.

⁴ Bruno Kessler Foundation, Trento, Italy.

⁵ University of Trento, CIBIO Department, Trento, Italy.

⁶ Pathology Unit, Department of Medicine and Surgery, University of Milano-Bicocca,
IRCCS Fondazione San Gerardo Dei Tintori, Monza, Italy.

⁷ Pathology Unit, IRCCS, Azienda Ospedaliero-Universitaria Di Bologna, Bologna, Italy.

⁸ Department of Pathology, Pederzoli Hospital, Peschiera del Garda, Italy.

⁹ Surgical Pathology and Cytopathology Unit, Department of Medicine-DIMED,
University of Padua School of Medicine, Padua, Italy.

¹⁰ Section of Pathology, Department of Medical and Surgical Sciences for Children and
Adults, University of Modena and Reggio Emilia, University Hospital of Modena,
Modena, Italy.

2.2 Abstract

One of the goals of pathology is to standardize laboratory practices to increase the precision and effectiveness of diagnostic testing, which will ultimately enhance patient care and results. Standardization is crucial in the domains of tissue processing, analysis, and reporting. To enhance diagnostic testing, innovative technologies are also being created and put into use. Furthermore, although problems like algorithm training and data privacy issues still need to be resolved, digital pathology and artificial intelligence are emerging in a structured manner. Overall, for the field of pathology to advance and for patient care to be improved, standard laboratory practices and innovative technologies must be adopted. In this paper, we describe the state-of-the-art of automation in pathology laboratories in order to lead technological progress and evolution.

2.3 Introduction

Efforts to standardize surgical pathology laboratory processes and reduce manual work have increased over the past decades, aiming to enhance diagnostic accuracy and patient care outcomes. The handling of anatomic pathology samples is critical, as loss or incorrect storage can have serious diagnostic, legal, and ethical implications. Recommended conditions for storage include controlled temperature and humidity for paraffin-embedded blocks and secure, traceable systems for glass slides

[45, 46]. The Italian Ministry of Health’s Superior Health Council has highlighted these issues in their guidelines [47].

From collection to storage, it is crucial to maintain a secure and controlled chain of custody for biological samples, ensuring quality, traceability, and proper conservation. Improving compliance and process efficiency requires solutions that automate and simplify labelling, archiving, and search processes. Automation, defined as the use of devices to replace or supplement human effort in a process [48], is key.

Standardization in tissue processing, analysis, and reporting [49] is a major focus in surgical pathology, ensuring precision and repeatability of diagnostic findings, as well as clarity in diagnostic reporting. New technologies, including digital pathology systems and artificial intelligence techniques, are being developed and applied to enhance diagnostic accuracy, though adoption has been gradual due to concerns about data privacy, cost, and compatibility [50, 51]. Also, molecular pathology provides results that need to be precise and require standardized analytical procedures before implementation [52]. In this setting, quality assurance and control systems play a crucial role, serving as adjuvants to ensure the accuracy and reliability of results [53, 54, 55].

Proper tracking, storage, and conservation of specimens are critical, impacting diagnostic accuracy, patient care, and research. The evolution of surgical pathology and patient care relies heavily on the adoption of advanced technologies and standardized practices. This paper describes the state-of-the-art in pathology laboratory automation, aiming to inspire innovation tools and processes to support operators, organizations, and, most importantly, patients.

2.4 Pre-analytical processes

The pre-analytical phase of tissue processing comprises all the steps, starting from receiving tissue specimens to the submission of histopathology slides for interpretation [45].

The application fields of automation in the pre-analytical phase of pathology include specimen collection and tracking, processing, embedding, cutting, and staining (Fig. 1 and Table 1).

2.4.1 Tracking samples

It has become vital to create automated systems that can assist with tracking and managing the workflow of the specimens due to the rising volume of specimens received. In order to track the location, status, and stage of processing for each specimen, barcode scanning equipment and laboratory information management software (LIMS) are utilized [56, 57].

Labs can decrease the possibility of human mistakes and accelerate the turnaround time for diagnostic tests by automating the workflow process for specimens. Automated systems, for instance, can generate barcode labels that can be applied to the specimen container for tracking and notify laboratory employees when a specimen is

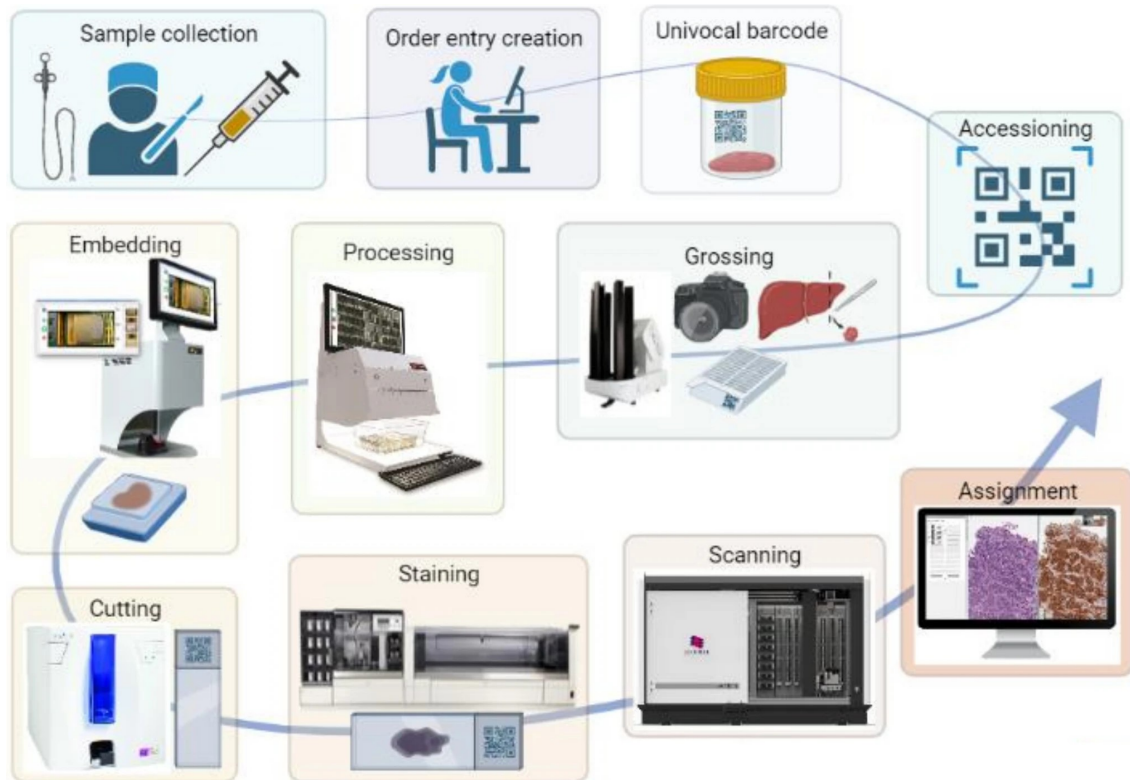


Figure 2.1: Automated workflow in the pre-analytical processes.

received. The specimen's location and status in the LIMS can be updated by scanning the barcode as it passes through each stage of processing. This enables the laboratory staff to keep track of each specimen's progress, see any delays or problems that require attention, and know who did a certain action with the specimen and when.

PathTracker™ (SPOT Imaging, Sterling Heights, MI, USA) is a laboratory solution for bulk barcode scanning that incorporates technology to acquire, process, analyze, and log all the barcodes in the field of view, with a reported scanning time of 30 s for a 150 cassette processing basket. Any damaged or poorly printed barcodes are flagged, and PathTracker™ then provides a set of correction tools to correct damaged barcodes automatically or manually, thus ensuring continuous workflow.

FinderFLEX (LOGIBIOTECH, Alseno, Italy) is a robotic unit for handling and scanning cytohistological samples. Thanks to a multi-articulated mechanical arm, FinderFLEX can handle and insert slides, macrosection slides, biopsy cassettes, super mega cassettes, and vials into the appropriate racks in a fully automated and secure way. The operator simply has to turn on the device and log in to add any new samples for storage. Using a latest-generation barcode scanner, FinderFLEX also rapidly scans any barcodes, QR codes, and Data Matrix 2D codes, communicating directly with the LIS to ensure a systematic and traceable sample management and handling process. FinderFLEX identifies and manages the samples directly from their standard racks and containers, significantly reducing handling times, also thanks to the automatic transmission of the gripper fingers. The device is also equipped with simple and intuitive software and a touchscreen panel, which the

operator can access in total safety in case of an emergency.

2.4.2 Tissue processing

The importance of standardization in tissue processing within anatomic pathology cannot be overstated. It ensures uniformity by reducing discrepancies and confirming that any variations are due to the samples themselves. This uniform approach simplifies quality control, making it easier to detect and rectify any issues like the presence of contaminants [58].

In the past, pathologists and technicians would spend countless hours manually preparing tissue samples for diagnosis. Tissue fixation and processing may now be carried out rapidly, precisely, and with a minimum of human involvement because of the development of automated methods and tools.

Tissue-Tek Xpress[®] x 120 tissue processor (Sakura Finetek, Tokyo, Japan) allows continuous streamlining of the histology workflow using vacuum infiltration to offer consistent results in rapid time, distributing cases uniformly and decreasing workloads, processing large tissue specimens in 2.5 h.

Running numerous protocols simultaneously on a single instrument is made possible by the HistoCore PEGASUS Plus tissue processor (Leica Biosystems, Wetzlar, Germany), providing a completely integrated system with the capacity to record each cassette individually, including cassette ID, amount, and color, as well as basket ID, user ID, and reagent information.

Compared to manual processing, automated tissue fixation and processing have a number of benefits. First, since all processes are meticulously regulated by the computer, automation lessens the possibility of mistakes and unpredictability in tissue processing. This may result in more precise and reliable diagnostic findings, enhancing patient care and outcomes. The second benefit of automation is quicker processing times since the computer can regulate the timing of each step to maximize effectiveness. Finally, automation frees up laboratory staff to concentrate on important duties like quality control.

2.4.3 Automation in tissue embedding

One of the most critical steps in the histology procedure is embedding; after the tissue processing stage, this laborious operation is done manually and requires proper training and experience. The correct orientation of the tissue within the paraffin is of paramount importance since a badly oriented specimen will result in an uninformative section and can lead to tissue loss at cutting, with detrimental consequences for the patient. The technician embeds surgical specimens and biopsies one at a time, making sure they are positioned correctly, which is frequently a laborious and time-consuming process. For this operation to produce the best circumstances for the cutting phase, trained specialists with good manual dexterity are needed.

Compared to manual embedding, automated embedding systems have a number of benefits, such as improved productivity, standardized processing, and less manual

labor.

By automating the process of embedding tissues as part of the processing protocols, the Synergy system (Milestone Medical, Sorisole, Italy) eliminates the need to manually reopen the cassettes and reposition the tissues. A carefully created rack, specialized molds, and pads make up the Synergy technology system. Through the use of a single tissue processing and embedding methodology, the sponges used for the pads guarantee the specimens' correct orientation and facilitate cutting at the microtome stage.

The Tissue-Tek AutoTEC[®] a120, in conjunction with Tissue-Tek[®] Paraform[®] cassettes and Tissue-Tek[®] Paraform[®] Tissue Orientation Gels (Sakura Finetek, Tokyo, Japan), is a component of Sakura's SMART automation concept to automate the manual work and produce a continuous flow in the lab. Such gels are made to securely hold and keep tiny tissue samples oriented. A complete system for automating cassette embedding with a throughput of up to 120 cassettes per hour is provided by the Tissue-Tek[®] Paraform[®] Sectionable Cassette System once the tissue is correctly oriented at grossing. This system locks the specimen during processing and embedding, minimizing tissue loss and eliminating the need for specimen reorientation.

These types of automated embedding systems appear to be superior compared to manual embedding, especially in terms of productivity and uniformity. Automated embedding can increase the accuracy and reliability of diagnostic testing by lowering the possibility of human error and variability. It must be kept in mind, however, that tissues can vary greatly in size, shape, and consistency, and not all may be suitable for automatic embedding. Some delicate or irregularly shaped samples may require manual embedding to ensure proper orientation and preservation.

2.4.4 Automatic microtome

Microtomes, the cornerstone of pathology labs since the nineteenth century, have radically transformed tissue analysis by producing ultra-thin sections for detailed cellular structure examination and disease pathology investigation. Despite their indispensable role, microtome operation remains an artisanal task, demanding skillful handling and precise adjustments. The critical challenges of section thickness variation and tissue distortion call for innovative approaches and advanced automation to ensure reliable, reproducible results. The automated microtome operates by slicing the tissue sample into thin slices with the help of a motorized cutting blade. The instrument's control panel allows for the generation of tissue sections with various thicknesses by adjusting the section's thickness. The instrument's automation also makes sure that tissue segment thickness is constant, lowering the possibility of mistakes and inaccurate diagnostic findings.

In this regard, an automatic microtome AS-410M has been developed by Dainippon Seiki (Nagaokakyo, Japan), which automatically performs high-precision and quality histological cuts according to the pre-established requirements for each case or tissue. The cut is then transferred to a slide where it is deposited and stretched; subsequently, the slide is stored in a drying chamber from where it can be collected. The cuts obtained are very homogeneous and of high quality. In addition,

the equipment may include roughing modules, cut quality control, slide printing, and connection to the Laboratory Information System (LIS) for full traceability of the samples. The approximate production is 250 blocks in a 7-h work shift, with the possibility to run 24 h a day.

Sakura's Tissue-Tek AutoSection[®] Automated Microtome offers one-touch trimming and customizable sectioning, coupled with numerous integrated safety measures. It aligns the block with the blade edge, ensuring precise XYZ positioning. This system enables consistent block orientation, regardless of prior trimming or sectioning on other microtomes, thus conserving both tissue and technician efforts.

Some limitations in applying such technology could be related to the fact that extremely hard or brittle materials might be challenging to cut consistently; moreover, very little biopsy might still require human hands and expertise to avoid the loss of precious tissue.

2.4.5 Automation in slide staining and coverslipping

The adoption of automated staining technology has accelerated the processing of huge sample numbers while also minimizing human error, enhancing consistency, and improving staining procedures' efficiency and dependability.

Hematoxylin and eosin (H&E)-stained slides represent the cornerstone of morphological diagnosis, and their importance cannot be overestimated [59]. Every step of the process can be automated, resulting in greater reproducibility, precision, and reliability; notably, it has been shown that automated individual staining protocol(s) as opposed to batch-stained slides might be preferable for digital pathology [60]. An example of an individual slide staining system is the Ventana HE 600 (Roche Diagnostics, Basel, Switzerland).

Using labeled antibodies, immunohistochemistry (IHC) is a potent diagnostic method in pathology that enables the identification of particular antigens in tissue slices. IHC staining has become more efficient thanks to automation, which has optimized incubation periods, temperature ranges, and reagent concentrations—elements crucial for precise antigen–antibody reactions. The automated systems also reduce background noise and non-specific staining, which raises the signal-to-noise ratio and the overall caliber of the stained slides.

Modern automated staining systems have been created by different companies to meet the various needs of pathology labs.

One example is the VENTANA BenchMark line of automated slide stainers (Roche Diagnostics, Basel, Switzerland), which provides complete IHC and in situ hybridization (ISH) staining solutions.

The BOND-PRIME automatic staining platform from Leica Biosystems (Wetzlar, Germany) can adapt to different workflow demands like batch, continuous, single slide, or STAT cases, or a combination of these for both IHC and ISH. Another example is the Tissue-Tek Genie[®] system from Sakura Finetek (Tokyo, Japan), a fully automated, random access stainer for IHC and ISH, with independent staining stations for handling slides with different antibodies and probes simultaneously and at any time.

A key step in the preparation of a high-quality histological glass slide is coverslipping. The quality of the coverslipping is important since the presence of air bubbles, excess or lack of mounting medium, and dried mounted slides can impair

the diagnosis. There are three types of coverslipping methods, namely, the classic glass coverslip, the liquid method, and the film method. The film method is the only automatic and has been demonstrated to be the fastest, with significantly less air bubbles and staining alterations compared with the other two methods [61], thus resulting in the best method for the production of glass slides for digital scanners.

2.4.6 Collaborative robots

In numerous situations, it is challenging to automate manual processes. Devices, even those manufactured by the same company, often lack sufficient coordination to transfer materials. A common daily laboratory task involves moving sections between rack systems, such as transferring samples from a staining platform to a coverslipping device [62]. The process can be time-consuming and may result in material loss due to the risk of components falling or breaking.

There is ample opportunity for enhancing production flow and intelligently integrating various steps, in addition to the need for further process development. The increasing adoption of robotic systems for material transfer across processes is a product of collaborative robotics. Collaborative robots, or “cobots,” feature sensors that facilitate safe human–robot interaction without necessitating protective barriers. Flexible, camera-assisted gripping devices also contribute to the functionality of these systems, allowing them to operate effectively. The Tissue-Tek SmartConnect® from Sakura (Tokyo, Japan) represents a cutting-edge technological advancement in laboratory automation, bridging the gap between human expertise and efficient, reliable processes. This collaborative robot has been designed to work seamlessly alongside laboratory technicians, assisting in various tasks while promoting accuracy and productivity. Once the Tissue-Tek Xpress®×120 is loaded through SmartConnect, automated tissue processing begins. SmartConnect then independently transfers the magazines to the Tissue-Tek AutoTEC® a120 embedder. Ultimately, SmartConnect delivers standardized, high-quality, embedded blocks prepared for microtomy. Laboratories can therefore improve workflow, lower human error, and boost overall effectiveness by putting such a system in place.

Additionally, the incorporation of cutting-edge technology into these systems, such as machine learning and artificial intelligence, might result in even more precise and accurate treatment of samples, enhancing the overall outcomes.

Such robots can benefit the lab by:

- eliminating manual and accidental errors as well as contamination risks;
- simplifying routine activities and improving processes and workflows;
- increasing productivity and efficiency;
- ensuring complete sample tracking and traceability, guaranteeing their quality
- reducing repetitive manual processes performed by health care staff, freeing up more time for strategic activities with high added value;
- helping improve patient satisfaction and, most importantly, patient safety.

2.5 Analytical processes

The application fields of automation in the analytical phase of pathology include digital pathology and the analytical process performed by computational pathology algorithms (Fig. 2 and Table 2), as well as synoptic reporting and data entry templates.

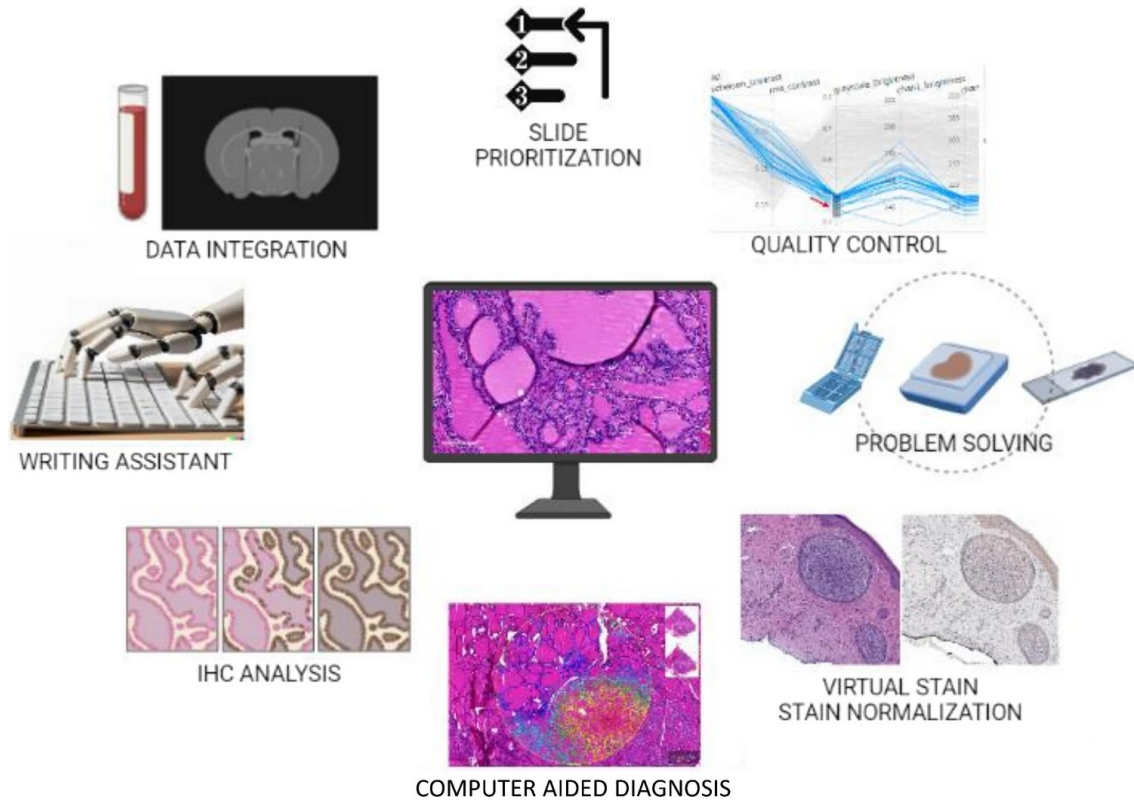


Figure 2.2: Automated workflow in the pre-analytical processes.

2.5.1 Digital and computational pathology

Modern diagnostic medicine’s crucial domains of digital pathology and computational pathology are redefining how doctors approach the analysis, diagnosis, and treatment of diseases [51].

Pathology glass slide scanners have revolutionized digital pathology by enabling the conversion of histological samples on glass slides into high-resolution digital images. This has enhanced information accessibility, storage, and sharing, fostering global collaboration among health care professionals. Over time, these scanners have significantly improved in terms of speed, resolution, and capabilities.

Notable products in the market include the NanoZoomer series (Hamamatsu, Hamamatsu City, Japan), Aperio (Leica Biosystems, Wetzlar, Germany), IntelliSite (Philips, Eindhoven, the Netherlands), Pannoramic series (3DHISTECH, Budapest, Hungary), and Axioscan (Zeiss, Oberkochen, Germany). Each of these devices offers impressive image quality, processing speed, and capacity, catering to the diverse needs of diagnostic laboratories and research facilities.

Pathologists may see, examine, and exchange high-resolution digital images of histological and cytological samples thanks to the digitalization of glass slides using whole slide imaging, or “digital pathology.”

Through the use of this technology, pathologists may now collaborate remotely, consult with specialists around the world, improve diagnostic precision, and speed up patient care [63].

The development and implementation of machine learning algorithms and artificial intelligence (AI) to evaluate digital slides is the subject of computational pathology. Using this method, quantifiable data may be extracted from the digitized slides and utilized to discover unique patterns and biomarkers, increasing diagnostic accuracy and enabling tailored medication [50].

Numerous developments in whole slide imaging technologies, rising processing power, and the accessibility of enormous annotated datasets have all contributed to the growth of digital and computational pathology. Investment and research in these areas have also increased as a result of the growing need for diagnostic solutions that are more effective, accurate, and cost-efficient.

The diagnosis of cancer is one of the main uses of digital and computational pathology. When it comes to the automated diagnosis and categorization of different tumor forms, including breast, lung, and prostate cancer, machine learning algorithms have achieved extraordinary results. These algorithms can examine digital histopathology images to find neoplastic cells, differentiate benign tumors from malignant ones, and even identify the subtypes and grades of the malignancies [64, 65, 66, 67].

In this regard, such technology may reduce the workload for pathologists, reduce interobserver variability, and enable more reliable and precise diagnoses by automating these procedures [68].

In order to pinpoint potential beneficial uses of AI in pathology, Heinz et al. conducted an anonymous online survey involving 75 domain experts in computational pathology from both academic and industrial backgrounds [69]. The survey results suggested that the most promising future application is seen as predicting treatment response directly from standard pathology slides.

As a matter of fact, among different applications in translational medicine, digital pathology is being actively investigated to predict response and identify patients most likely to respond to treatment.

In the era of immune-oncology, the selection of patients who may benefit the most from immune checkpoint inhibitor-based therapies (ICI) like PD-1/PD-L1 blockade is a major and still unsolved issue [70]. Notably, besides PD-L1 expression on tumor and immune cells, the immune contexture represented by tumor-infiltrating lymphocytes (TILs) has been demonstrated to have strong predictive potential [71].

Interestingly, Park et al. have developed an AI-based algorithm for the analysis and quantification of TILs in the tumor microenvironment, capable of defining three immune phenotypes (IP): inflamed, immune-excluded, and immune-desert [72]. These authors demonstrated that patients with inflamed tumors have a better prognosis both in terms of OS and PFS and, in particular, that patients with inflamed neoplasms and high expression of PD-L1 show a significant improvement in survival compared to patients with high expression of PD-L1 but non-inflamed tumors.

Such results underscore the fact that the application of image analysis offers

increased accuracy and efficiency by automatically measuring multiple parameters that are impossible to achieve by eye.

Besides tumor pathology, computational pathology is also being investigated and applied in critical but often neglected fields like transplantation pathology, a highly specialized area of pathology that examines both post-transplant graft biopsy results for rejection or graft damage as well as organ donor biopsy for organ allocation, as well as in many different fields of functional and non-neoplastic pathology [73, 74, 75].

Overall, it appears clear that digital and computational pathology provide very useful methods for managing and interpreting massive datasets from various sources, such as genomics, proteomics, and clinical data, in the age of big data. Through the use of machine learning algorithms, integrative data analysis, a greater understanding of disease causes, and the discovery of novel possibilities for diagnosis, prognosis, and treatment are all made possible.

One of the primary obstacles to the integration of digital pathology into clinical practice, as perceived by administrators, is associated with its cost. In this regard, Ho and colleagues elaborated a financial projection for digital pathology implementation at a large health care organization in order to estimate potential operational cost savings [76]. The projected savings were based on two main benefits associated with the use of digital pathology: (1) potential improvements in workflow/productivity and lab consolidation; and (2) avoided treatment costs due to reduced rates of interpretive errors by general, non-subspecialist pathologists. The authors projected that the total cost savings over 5 years could reach approximately \$18 million. This suggests that if the costs of acquiring and implementing digital pathology do not exceed this value, the return on investment becomes attractive to hospital administrators.

Currently, different integrated digital pathology systems are being implemented around the world, providing clear examples of the feasibility of the implementation of digital pathology workflows both in small and large pathology departments supporting large and distributed health care organizations with complex patient demographic profiles [77, 78, 79]; also, official guidelines have been published [80].

Finally, the value of digital pathology in the education of anatomic pathology is beyond doubt, with a growing number of resources like digital pathology atlases. Essential skills like identifying features, providing differential diagnoses, annotating, taking photographs, describing, and presenting are all improved through the use of such resources. The way these resources are used seems to play a crucial role in overcoming the reluctance to use digital tools among certain learners. Regularly integrating these resources into unidentified case discussions, educational collections, and tutorials has the potential to dramatically improve and speed up the learning process [81].

Despite significant advancements and prospective applications, digital and computational pathology still has a number of issues that need to be resolved. Because medical data is sensitive and professionals must share photos and information, privacy and security issues with the data are raised. To ensure consistency and interoperability across various systems and organizations, it is also crucial to standardize digital imaging processes, data formats, and annotation strategies. Machine learning algorithms must also undergo thorough validation and testing before being integrated into clinical processes in order to guarantee their dependability and

clinical utility [82]; strategies for preventing model accuracy losses in the contest of artifacts must also be developed [83].

We believe that the new generation of pathologists, besides having solid and comprehensive anatomic pathology training, will also need to expand their cultural background to include at least the basic principles of computational pathology and image analysis in order to bridge the cultural gap between medicine, computer science, and data analytics. This will not mean that pathologists will have to be a sort of hybrid professional, but surely they will need to have the ability to collaborate with computer scientists to understand and overcome the possible limitations of new technological approaches and, importantly, to be the main actors in this paradigm shift.

The future of digital and computational pathology is still bright, despite these difficulties, as it will make it possible to significantly increase diagnosis accuracy and give patients a more thorough grasp of disease processes by combining new imaging modalities with machine learning algorithms.

2.5.2 Structured synoptic reporting

The practice of structured reporting in pathology is of utmost significance as it fosters uniformity and comprehensiveness in recording vital cancer data. This standardization not only amplifies the clarity and usefulness of reports for immediate patient care but also guarantees that invaluable data is systematically captured for secondary purposes such as research, quality assurance, and public health management.

The International Collaboration on Cancer Reporting (ICCR) has played a pivotal role in propelling this global standardization, thereby contributing to enhanced patient outcomes and breakthroughs in cancer research [84, 85].

The ICCR envisions improving patient outcomes through internationally standardized pathology reporting. The formulation of evidence-based datasets, encompassing all significant and current reporting information for any specific cancer, results in more exhaustive pathology cancer reports, refined cancer staging, and the optimization of treatment protocols for cancer patients.

Beyond the development of datasets, the ICCR has pinpointed two additional key areas of focus for the future. The first is the translation of datasets into multiple languages to expedite the adoption of reporting standards in both developed and low-to-middle-income countries (LMICs). The second is the conversion of dataset standards into machine-readable formats to facilitate their electronic implementation and global data interoperability.

The creation of evidence-based datasets, which include all essential and contemporary reporting details for each specific cancer, not only leads to more thorough pathology reports on cancer but also improves cancer staging and fine-tune treatment approaches for cancer patients [86]. Furthermore, these datasets represent the basis for the creation of nationwide networks between pathology laboratories as is the case with the Pathological Anatomy National Automated Archive (PALGA) that has been operating in the Netherlands since 1971. The aim of such organization is to promote communication and information exchange between participating laboratories and to provide potentially useful data for health care professionals in the interest of patient care and research [87].

2.6 Post-analytical processes

The application fields of automation in the analytical phase of pathology include storage and biobanking of tissues and specimens and digital imaging archiving (Fig. 3 and Table 3).

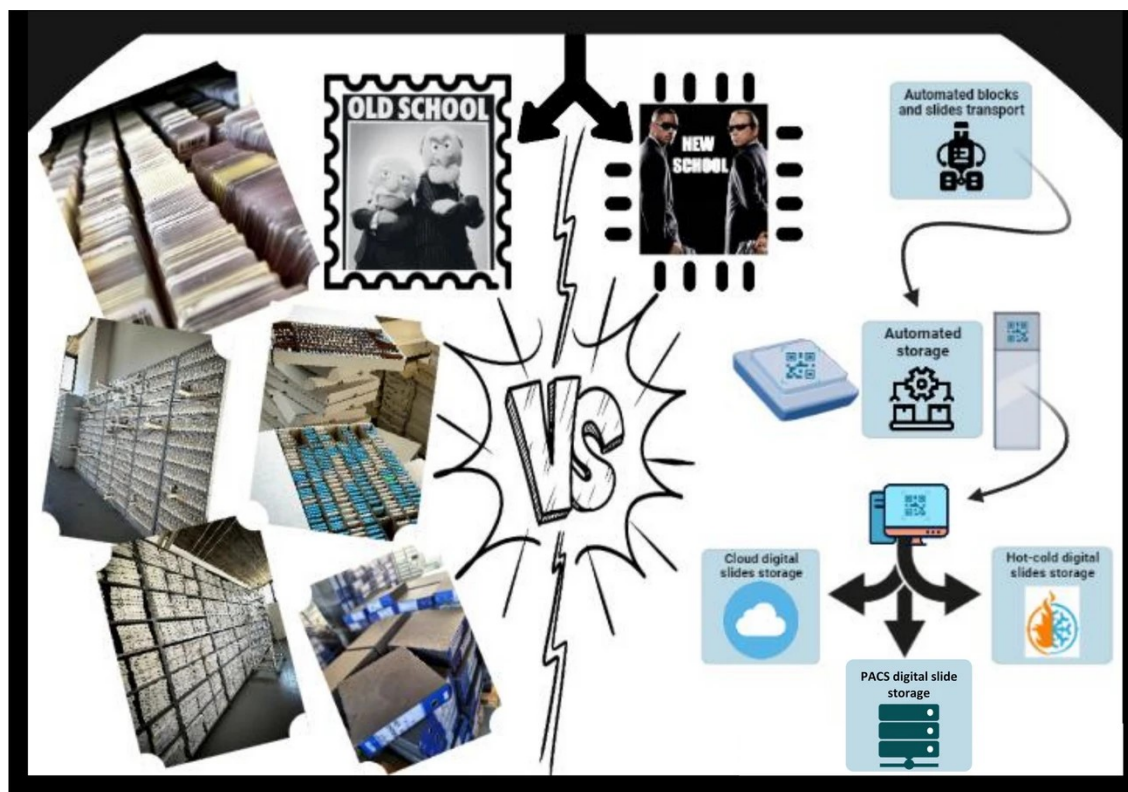


Figure 2.3: Shifting paradigms in storage and archiving.

2.6.1 Storage and biobanking of tissue and specimens

Tissue and surgical specimens should be kept in a way that protects their integrity and averts deterioration or contamination. Depending on the specimen type and the intended purpose, this can entail freezing, refrigeration, or formalin fixation [88].

Professional associations such as the Joint Commission on Accreditation of Health Care Organizations (USA) and the College of American Pathologists (USA) advise that tissue blocks and slides be kept for a long enough time to ensure that the patient is treated properly. The UK's Royal College of Pathologists advises keeping blocks for life and histology slides and smears for 10 years [89].

Different companies have created automated histology cassette storage and management solutions to improve traceability, speed up archiving and retrieval, protect sensitive patient tissue and biopsy blocks, and cut down on sorting and storage time [90].

To preserve lab security, increase productivity, and optimize procedures, built-in automation must be integrated. Reliable laboratory information system (LIS) integration and user-friendly software on board boost performance and lower errors. Complete traceability and straightforward cassette retrieval are also necessary for

maintaining an efficient operation. Additionally, a safe, secure storage facility that is constantly watched over is needed to ensure the preservation, integrity, and quality of samples [91].

To make sure that each task is managed and monitored in the right and timely manner, many solutions have been put forth by vendors, such as Arkive BC™ (Menarini Diagnostics, Florence, Italy), which may be easily interfaced with middleware or included into the LIS as needed. Authorized technicians may comprehensively oversee all areas of the operations thanks to the user-friendly interface of the onboard software.

For histology/cytology slides, the same manufacturer has produced a product named Arkive SL™, a device for loading slides to be archived and the retrieval of archived slides. The SLTrack tool automatically records a full audit trail of each slide and SL item in the lab, enabling users to trace, identify, and retrieve samples quickly and efficiently. Additional units are added to build a modular system, which increases overall capacity without limiting the number of slides that can be retained for short-, medium-, and long-term storage requirements.

SmartCABINET and ClientCABINET (LOGIBIOTECH, Alseno, Italy) are smart units for the automated and traceable storage of cytohistological samples. They are an intelligent automated system which can help operators perform all phases of the storage and retrieval of cytohistological samples. Thanks to the software installed, “pick-to-light” technology, Wi-Fi connection, and the use of samAPP, each storage and retrieval operation is carried out in a secure, traceable, and recorded manner. They are modular and flexible, as numerous identical units can be added to each SmartCABINET, and each of these can in turn command an infinite number of ClientCABINETs to adapt to any storage volume requirement and space available.

2.6.2 Digital imaging archiving

In pathology, digital image archiving describes the procedure of keeping WSIs and related diagnostic data for analysis and future use. The benefit of digital imaging archiving is that the data is always accessible and independent of traditional archives [92].

A digital archive makes information easier to find and faster to retrieve, resistant to deterioration, able to view earlier case comments, and simple to share with coworkers [93].

The methodology, high-volume scanning, and particularly the enormous storage capacity required, as well as the associated costs, provide unique problems when developing a fully digitized slide library [94].

Medical management systems that store and handle digital images and data in a vendor-neutral format are called vendor-neutral archives (VNAs). Regardless of vendor or manufacturer, a VNA may save WSIs and associated data from digital pathology systems and link with other health care IT systems like EHR and LIS. VNAs can store images on dedicated hardware or link to third-party systems’ images. Central VNAs enable image backup, disaster recovery, business continuity, and interoperability with external organizations and health information exchanges. When new imaging technologies like radiology PACS are introduced, they can reduce picture data migration. Federated VNA may be cheaper and faster to install due to their lower hardware infrastructure requirements, but they may have more

trouble connecting to multiple image sources [95].

Digital imaging archiving enables pathologists to analyze enormous amounts of data and find patterns and trends for research and therapeutic decision-making. Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) is crucial. With over 350,000 ideas and a million relationships, SNOMED CT is the most comprehensive, multilingual clinical health care vocabulary in the world [96].

SNOMED topography (SMOMED T) and SNOMED morphology (M) codes could be used to choose cases from a digital archive, which could help minimize the archive's size and cost while preserving the benefits of quick and easy retrieval of WSIs [93].

2.7 Conclusion and Future Directions

Automation in surgical pathology has demonstrated immense potential for enhancing the accuracy, efficiency, and overall quality of patient care. Through the integration of advanced technologies such as robotics, AI, and machine learning, pathology laboratories can reduce human error, streamline workflows, and expedite the diagnostic process. The future of surgical pathology is undeniably intertwined with the advancements in automation, paving the way for more accurate diagnoses, improved patient outcomes, and a more profound understanding of diseases.

2.8 Disclaimer

The cited platforms and vendors are meant to serve as examples stemming from the authors' knowledge and experience. Such examples are not intended as endorsements and might not accurately represent the latest technological advancements.

Acknowledgements The authors are thankful to Davide Seminati, MD and Vincenzo L'Imperio, MD (Department of Medicine and Surgery, Pathology, University of Milan-Bicocca, Milano, Italy) for their precious help in preparing the images.

Author contribution All authors contributed to the study conception and design. Material preparation and data collection were performed by Enrico Munari, Albino Eccher, and Luca Cima. The first draft of the manuscript was written by Enrico Munari; Aldo Scarpa, Luca Cima, Matteo Pozzi, Fabio Pagni, Francesco Vasuri, Stefano Marletta, Angelo Paolo Dei Tos, and Albino Eccher commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open access funding provided by Università degli Studi di Verona within the CRUI-CARE Agreement.

Data Availability The authors confirm that the data supporting the findings of this study are available within the article.

2.9 Declarations

Conflict of interest Albino Eccher, Angelo Paolo Dei Tos, and Aldo Scarpa are consultants for Logibiotech. Enrico Munari, Luca Cima, Matteo Pozzi, Fabio Pagni, Francesco Vasuri, and Stefano Marletta have no relevant financial or non-financial interests to disclose

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Chapter 3

Explainability in endoscopic imaging

3.1 Summary

DOI: 10.1186/s12911-022-02043-w

Introduction The study focuses on the challenges of diagnosing inflammatory bowel diseases (IBD), such as Crohn’s Disease (CD) and Ulcerative Colitis (UC), using endoscopic imaging. These conditions often present with overlapping features, making accurate diagnosis difficult. Leveraging explainable AI (XAI) models, the study aims to provide a reliable and interpretable decision-support tool for clinicians.

Methods The study develops a deep learning model to classify endoscopic images into three categories:

1. Healthy vs. pathological (P vs. N).
2. Crohn’s Disease vs. Ulcerative Colitis (UC vs. CD).
3. Ulcerative Colitis vs. healthy (UC vs. N).

The approach involves data preprocessing, feature extraction, and ensemble learning for model optimization. Interpretability techniques like Grad-CAM are used to identify key regions contributing to the model’s decisions.

Results The deep learning model achieves:

- Matthews Correlation Coefficient (MCC) >0.9 for P vs. N and UC vs. N classifications.
- MCC >0.6 for UC vs. CD classification.

These results indicate robust performance in differentiating healthy samples from pathological ones and between IBD subtypes.

Discussion The model demonstrates the potential of AI in supporting clinical endoscopy by improving diagnostic accuracy and reducing subjectivity. However, challenges include dataset variability, generalizability to unseen data, and ethical considerations regarding model interpretability.

Automatically detecting Crohn’s disease and Ulcerative Colitis from endoscopic imaging

Marco Chierici¹, Nicolae Puica², Matteo Pozzi^{1,3}, Antonello Capistrano⁴, Marcello Dorian Donzella⁵, Antonio Colangelo⁴, Venet Osmani¹, Giuseppe Jurman¹

¹ Fondazione Bruno Kessler, via Sommarive, 18, 38123 Trento, Italy.

² PagoPA S.p.A., Rome, Italy.

³ Università degli studi di Trento, via Calepina, 14, 38122 Trento, Italy.

⁴ GPI S.p.A., via Ragazzi del ’99, 13, 38123 Trento, Italy.

3.2 Abstract

Background

The SI-CURA project (*Soluzioni Innovative per la gestione del paziente e il follow up terapeutico della Colite UlceRosA*) is an Italian initiative aimed at the development of Artificial Intelligence (AI) solutions to discriminate pathologies of different nature, including Inflammatory Bowel Disease (IBD), namely Ulcerative Colitis (UC) and Crohn’s Disease (CD), based on endoscopic imaging of patients (P) and healthy controls (N).

Methods

In this study we develop a deep learning (DL) prototype to identify disease patterns through three binary classification tasks, namely i) discriminating positive (pathological) samples from negative (healthy) samples (P vs N); ii) discrimination between Ulcerative Colitis and Crohn’s Disease samples (UC vs CD) and, iii) discrimination between Ulcerative Colitis and negative (healthy) samples (UC vs N).

Results

The model derived from our approach achieves a high performance of Matthews correlation coefficient (MCC) >0.9 on the test set for *P versus N* and *UC versus N*, and $MCC > 0.6$ on the test set for *UC versus CD*.

Conclusion

Our DL model effectively discriminates between pathological and negative samples, as well as between IBD subgroups, providing further evidence of its potential as a decision support tool for endoscopy-based diagnosis.

3.3 Introduction

Inflammatory bowel diseases (IBD), including Crohn’s disease (CD) and Ulcerative colitis (UC), are chronic and recurrent diseases. CD patients have healthy parts of

the intestine mixed in between inflamed areas, while UC induces a continuous inflammation of the colon. Further, CD may occur in all the layers of the bowel walls, while UC only affects the innermost lining of the colon. Although they both have an undetermined etiology, research advances have outlined some of the pathways leading to their insurgence: 1) genetic predisposition associated with the environment induces disruption of the intestinal microbial flora; 2) the structure of the epithelial cells and of the immune system of the intestine determine the risk of developing the disease. IBD is diagnosed using a combination of endoscopy (for CD) or colonoscopy (for UC) and imaging studies, such as contrast radiography, magnetic resonance imaging, or computed tomography. Physicians may also check stool samples to make sure symptoms are not being caused by an infection or run blood tests to help confirm the diagnosis. Still, a definite diagnosis of IBD remains a challenging task [97], often affected by subjective judgement [98]. Several automated approaches have been published in the recent literature [98, 99] attempting to provide computational support to improve the diagnostic task, with machine learning (ML) models playing a major role, however several challenges remain [100, 101]. In this line Stidham et al. [102] found deep learning model performance to be similar to an experienced clinician in evaluating severity of UC, while Takenaka et al. [103] derived a high performance model in identifying UC patients with endoscopic remission. Klein et al. [104] focused on analysis of colonic biopsies of CD patients to predict post-biopsy clinical phenotypes and outcomes, similar to Waljee et al. [105, 106] developing a model to predict IBD-related hospital admission and disease flare.

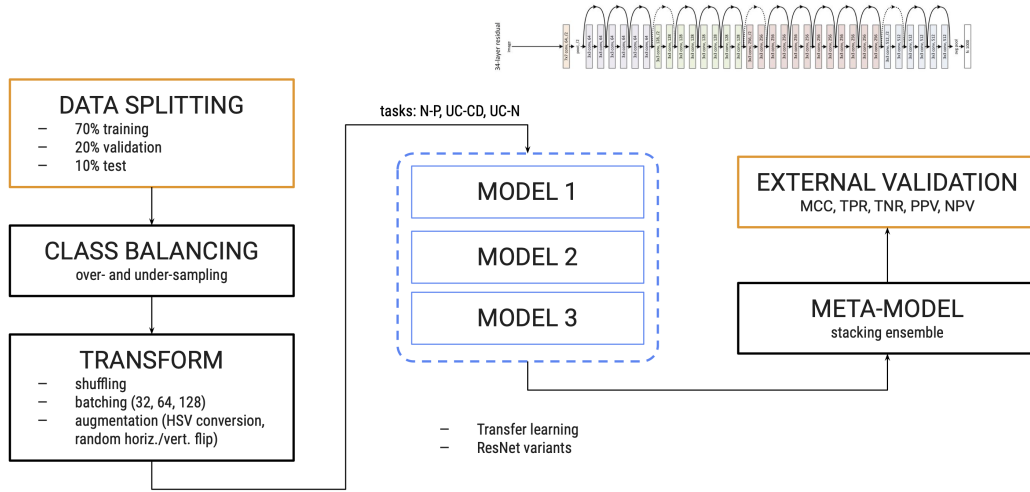
As a consequence, the emergence of Artificial Intelligence (AI) solutions based on Deep Learning (DL) models comes as no surprise. The models exploit the potential of artificial neural networks to automatically process different types of data [107, 108], including endoscopic imaging [109]. Our work naturally embeds in this research line, combining recent DL architectures with more classical ML strategies such as ensemble learning to further enhance the predictive performance. The promising results obtained can support the clinicians in providing a more objective and reliable diagnosis, thus reducing the risk of misidentification of CD and UC, an important aspect considering different treatment options and follow-ups of the two conditions [109].

3.4 Methods

We tackle the problem of IBD detection from endoscopic imaging with an ensemble learning approach based on a fine-tuned ResNet architecture. An overview of the general workflow is presented in Fig. 3.1.

Briefly, the input images first undergo a random partitioning into train and test sets: the train set is used for model development in a k-fold cross-validation (CV) schema, while the test set is kept apart to obtain predictions from the trained model. Then, after a preprocessing phase aimed at preparing the images for the predictive modeling, three pre-trained Residual Network models [110] of increasing complexity are fine-tuned on the preprocessed images in a transfer learning setting. Finally, the models are combined in a single meta-model with a stacking ensemble method.

Figure 3.1: *Experimental workflow*. After an upstream processing of the input images (blocks “Data splitting”, “Class balancing”, “Transform”), for each classification task three pre-trained ResNet variants of increasing complexity (“Model i ”, $i = 1, 2, 3$) are used as weak learners and fine-tuned in 5-fold cross-validation on the input data (see Methods). A meta-model is then built by stacking ensemble of the three weak learners, evaluating the performance on the external test set in terms of different classification metrics (see Methods). N-P, negative versus positive; UC-CD, Ulcerative Colitis versus Crohn’s Disease; UC-N, Ulcerative Colitis versus negative; MCC, Matthews Correlation Coefficient (MCC); TNR, true negative rate; TPR, true positive rate; NPV, negative predictive value; PPV, positive predictive value



3.4.1 Outcome definition

In this work, we investigate three major outcomes, namely: i) discrimination between negative (healthy) samples and positive (pathological) samples (N-P), ii) discrimination between Ulcerative Colitis and Crohn’s Disease (UC-CD) samples and, iii) discrimination between negative samples and Ulcerative Colitis (UC-N).

3.4.2 Dataset description

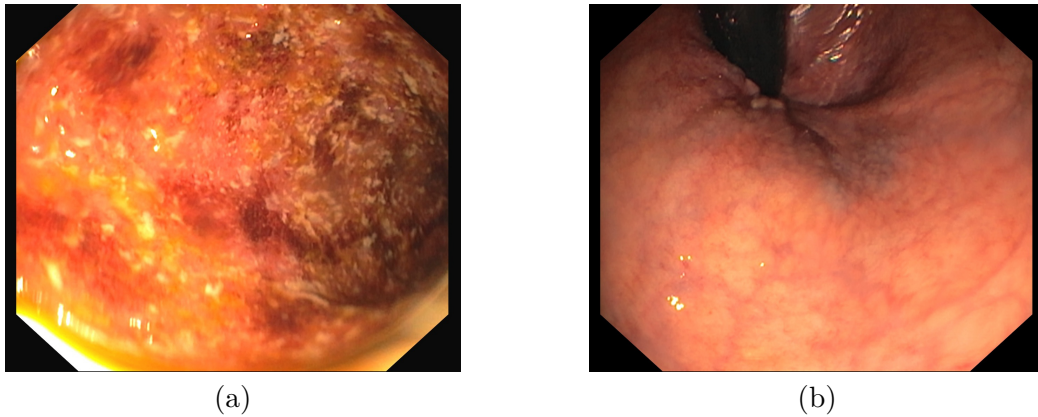
The SI-CURA dataset includes 14,226 three-channel RGB (red, green, blue) endoscopic images of different sizes split between “Positive” patients (P, $n=11,404$) and “Negative” healthy controls (N, $n=2,822$). In this work, we consider a single image to be a sample for the classification task. Positive (pathological) samples are further labeled as Ulcerative Colitis (UC), Crohn’s disease (CD), and Inflammatory bowel disease (IBD). Table 3.4 illustrates the sample stratification for each class and Figure 3.2 shows an example of positive and negative images.

In developing and evaluating the model, the overall dataset was split into two main subsets: training (90% of total sample size) and test (10%). The training set was further split for model development according to a 5-fold cross-validation schema. Since the P data is $3\times$ the N data, under- and over-sampling were used to balance the class distributions using the Imbalanced-dataset-sampler Python li-

Table 3.1: Number of elements (images) in each class, in the format “raw data (preprocessed data)”. P, positive; N, negative; UC, Ulcerative Colitis; CD, Crohn’s disease; IBD, inflammatory bowel disease.

P			N
UC	CD	IBD	
4,388 (3,594)	5,949 (4,098)	1,067 (823)	2,822 (2,815)

Figure 3.2: An example of positive (a) and negative (b) images within the SI-CURA dataset.



brary¹, which automatically estimates the sampling weights and mitigates overfitting when used in conjunction with data augmentation techniques.

3.4.3 Data preprocessing

An image editor was used to remove undesired artifacts, such as signs, writings, medical instruments, black, white, and corrupted images. Further, all images were converted to JPEG with lossy compression to have a consistent format across the dataset. Table 3.4 details the number of elements in each class after preprocessing.

Data augmentation is a technique commonly used to increase the amount of relevant data in the original dataset, thus providing additional examples to the neural network. Augmentation exploits the fact that convolutional neural networks are invariant to transformations such as translation, size, viewpoint, rotation. The following techniques were used: RGB to HSV (hue, saturation, value) colour model conversion, random horizontal flip, and random vertical flip, all of which had an equal probability of being applied.

Several experiments were performed with different batch and image sizes. The best results were obtained with a batch size of 32, 64, and 128, depending on the model’s depth. To speed up the training time, we rescaled all images to 700x700 pixels, equal to the average image dimension.

¹<https://github.com/ufoym/imbalanced-dataset-sampler>

3.4.4 Deep learning architecture

We tackled the tasks through a transfer learning approach using the following variants of Residual Networks (ResNet) [110]: ResNet18, ResNet34, ResNet50, ResNet101, and ResNet152, all of which were imported pre-trained on ImageNet from the PyTorch Hub². The choice of using pre-trained models leverages the information already learned by the network after an extensive training on the large ImageNet dataset. The fully-connected head of the ResNet networks was swapped with one performing binary classification: an untrained sequential module ending with a linear transformation and two outputs. Moreover, the rectified linear unit function (ReLU) [111] and dropout technique (with $p = 0.5$ probability) were used in this module. ReLU activation has been widely used in deep neural networks due to its advantages over other functions (e.g. sigmoid): most importantly, it overcomes the vanishing gradient problem [112] that has been afflicting neural networks for several years. We adopted dropout to increase regularisation, preventing the coadaptation of the neurons. The weights of the layers were frozen except for the last two, which were updated with a small variation of the learning rate. In a typical transfer learning approach, the layers of the network are frozen to the trained weights except for the last (“higher”, or “top”) layers (farther to the input). This is because the first-layer features are more general and the last-layer features are more specific [113]. There are different strategies for fine-tuning the higher layers: typically, the final fully-connected classification part of the network is trained, leaving the convolutional layers frozen. However, it is also possible to unfreeze one or more convolutional layers as well to fine-tune the performance. We chose to unfreeze the last two convolutional layers as a compromise between achieving a good performance and avoiding overfitting, following recommendations in the literature [114]. The optimiser used for the training was Adaptive Movement Estimation (Adam) [115], a replacement for Stochastic Gradient Descent that combines the advantages of the Adaptive Gradient (AdaGrad) and Root Mean Square Propagation (RMSProp) algorithms, making it a popular choice in recent studies [116]: AdaGrad maintains a per-parameter learning rate that improves performance on problems with sparse gradients, and RMSProp also maintains per-parameter learning rates using a decaying average of recent partial gradients. This means that Adam improves parameter optimisation, particularly in online and non-stationary (e.g., noisy) tasks. The actual number of epochs was determined by early stopping based on the validation loss, with a patience of 30 epochs. The maximum number of epochs was set to 200, a value high enough to allow the net to learn as much as possible until the early stopping technique takes place.

3.4.5 Differential learning rates

The learning rate is one of the most important hyperparameters that can be tuned to improve optimisation convergence. It is usually chosen by time-consuming techniques such as grid search that exhaustively experiments with different values, picking the one that works best in terms of performance on the validation set. A more convenient approach, which was adopted in this study, is the learning rate finder [117]. During each epoch, the algorithm makes the learning rate cyclically

²<https://pytorch.org/hub/>

vary between a lower and an upper bound. The loss corresponding to each learning rate is computed, and the learning rate yielding the steepest drop in the loss is chosen.

The learning rate finder method was complemented by the use of different learning rates during the training, an approach commonly referred to as “differential learning rates”, which is particularly useful in a transfer learning setting. In fact, only the layers immediately preceding the fully-connected classifier head were unfrozen and trained. Since the ResNet model is pre-trained on the large ImageNet dataset, these layers only need a small amount of fine-tuning to achieve good prediction performance. In particular, the learning rate values for the ResNet’s `layer4` (just before the classifier) and `layer3` were reduced by a third and by a ninth, respectively.

3.4.6 Ensemble learning

Ensemble learning is a machine learning paradigm where multiple models (“weak learners”) are trained to solve the same problem and finally combined to achieve better results. Two ensemble methods were used in this study: averaging prediction and stacking. While the former keeps the models independent from each other and averages their predictions, the latter combines the predictions of the weak learners through a final classifier, resulting in a meta-model. We combined three weak learners in a stacking ensemble meta-model: ResNet34, ResNet50, and ResNet101 were used as weak learners in the N-P and UC-N cases, and ResNet34, ResNet50, and ResNet152 in the UC-CD case. Other combinations were tried, but they performed worse. The ensemble meta-model has six input features (two for each weak learner) and two outputs features, and it was further trained for several epochs by using the early stopping criterion.

3.4.7 Performance evaluation metrics

The metrics used for assessing the model performance in both training and evaluation stages are Matthews Correlation Coefficient (MCC), true positive rate (TPR), true negative rate (TNR), positive predictive value (PPV), negative predictive value (NPV), and the confusion matrices. The MCC was used as the main metric because it is particularly suitable for binary and multiclass classification [118], and it is generally regarded as a balanced performance measure that can be used even if the classes are of very different sizes. MCC is computed from the values of the confusion matrix, true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), as in the following:

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

The values of MCC range between -1 and 1, where 1 means perfect classification, -1 perfect misclassification (inverse prediction), and 0 random guess or, generally speaking, absence of correlation between the predictions and the ground truth.

We also computed the training and validation losses and the 95% studentized bootstrap confidence intervals for the training MCC.

3.4.8 Model interpretation

In an attempt to better understand which input features are deemed important by the trained deep learning models, we used two explainable AI (XAI) algorithms implemented in the Captum library³, namely Saliency and Guided backpropagation. Saliency [119] is one of the most straightforward approaches for estimating input attribution: it computes the gradient of the model output with respect to the input pixels of a given image. While a simple approach to deep model interpretation, Saliency maps cannot capture the input feature interactions and tend to be noisy. Guided backpropagation [120] (GBP) is an extension of the Saliency algorithm aiming to alleviate these issues: GBP also computes the gradient of the output with respect to the input; in addition, it overrides ReLU backpropagation, resulting in backpropagation of non-negative gradients only. Each method computes an attribution value (the higher, the more important) for each input pixel: the resulting attribution maps are then qualitatively assessed by overlaying them on the original image.

3.4.9 Implementation

The following frameworks and libraries were used: PyTorch v1.5.0 with torchvision v0.6.0a0, Captum v0.4.1, CUDA 10.2, and Jupyter v6.1.4. Color model conversion was performed by the PIL library. The packages were installed in a Anaconda virtual environment for improved reproducibility. The analyses were run on a ppc64le server, provided by GPI S.p.A., with 128 CPUs (max 4023 MHz, min 2394 MHz) and 4 NVIDIA Tesla P100 SXM2 16GB GPUs. The training of neural networks was performed on multiple GPUs using PyTorch's `DataParallel`.

3.5 Results

3.5.1 Classification

Figures 3.3-3.5 show the cross-validation and test set performance of the weak learners and the ensemble meta-models for all classification tasks.

The test set confusion matrices are reported in Fig. 3.6 for the best-performing weak learner of each task.

Table 3.5 reports the results in terms of average cross-validation MCC with 95% confidence intervals and the test set MCC, TPR, TNR, PPV, and NPV. The rows of Table 3.5 are ranked task-wise by decreasing test set MCC. The average-prediction ensemble models performed consistently worse for all tasks and are not shown.

According to Table 3.5 and Fig. 3.3, the top three nets (or weak learners) for N-P are ResNet50, ResNet34, and ResNet101. While the first one is trained with an unbalanced training data loader, the latter two nets are trained using a balanced data loader.

As for the UC-CD task, the best results were obtained by ResNet34, ResNet50, and ResNet152 (Table 3.5, Fig. 3.4). The first two nets were trained with a balanced data loader and the last one with an unbalanced data loader. In this task,

³<https://captum.ai/>

Figure 3.3: *N-P classification results*. Metrics values for the ensemble model (resnet34-50-101) and the weak learners resnet34, resnet50, and resnet101. Light blue: cross-validation; dark blue: test set. Black bars represent 95% confidence intervals.

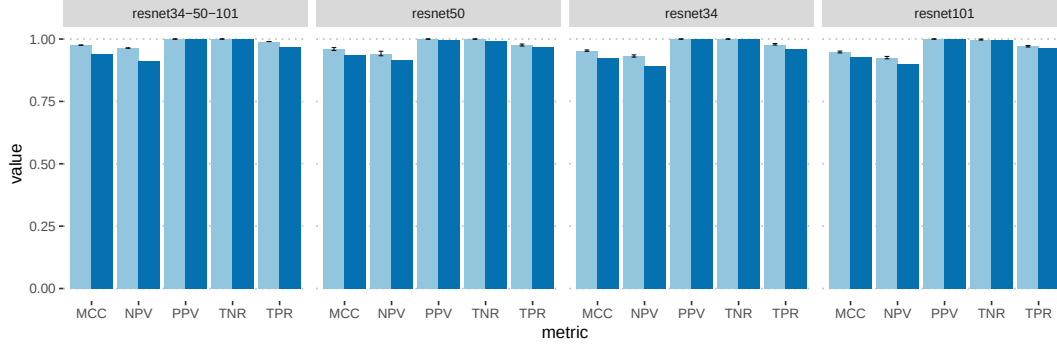


Figure 3.4: *UC-CD classification results*. Metrics values for the ensemble model (resnet34-50-152) and the weak learners resnet34, resnet50, and resnet152. Light blue: cross-validation; dark blue: test set. Black bars represent 95% confidence intervals.

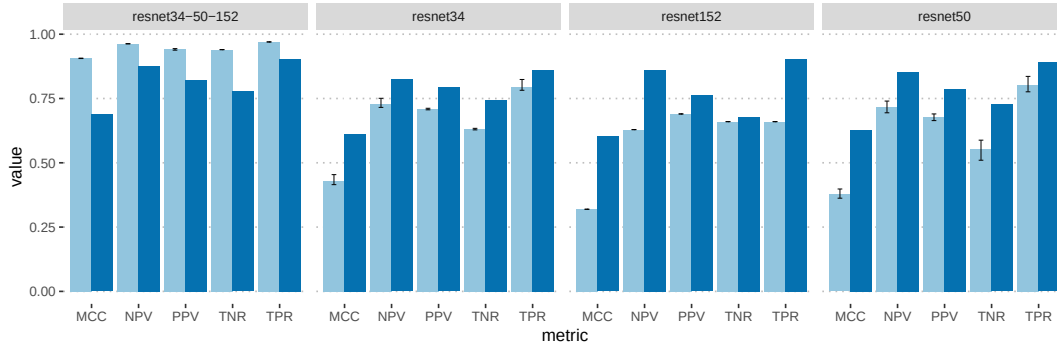
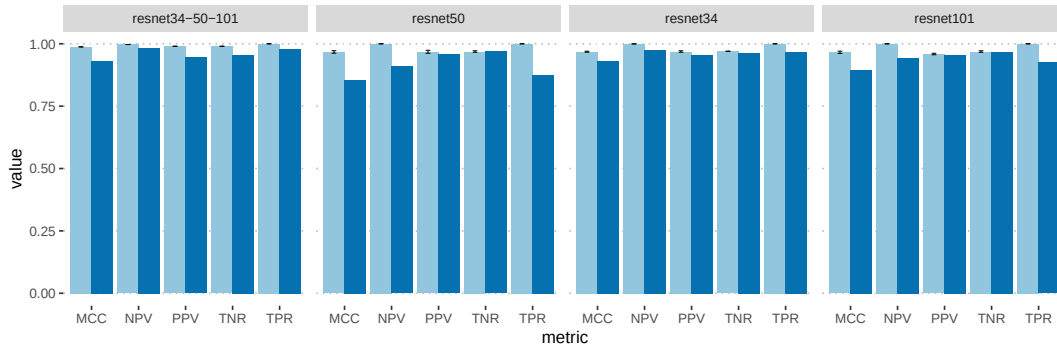


Figure 3.5: *UC-N classification results*. Metrics values for the ensemble model (resnet34-50-101) and the weak learners resnet34, resnet50, and resnet101. Light blue: cross-validation; dark blue: test set. Black bars represent 95% confidence intervals.



the classes are not as imbalanced as in N-P (Table 3.4): however, the dataset balancing technique helped maintain balanced the confusion matrices in the training and evaluation phase. As expected, the meta-model slightly improves over the weak

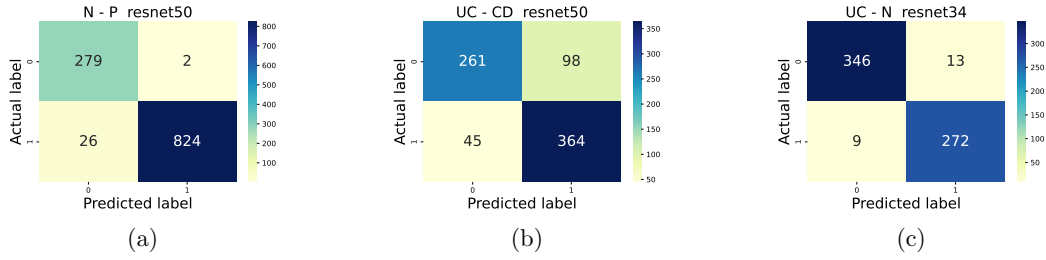
Figure 3.6: *Confusion matrices*. Test set confusion matrices for the best-performing weak learners of each classification task (see Table 3.5).

Table 3.2: Metrics values for the best trained nets, ranked by decreasing Matthews Correlation Coefficient (MCC) on the test set; cv, ts: cross-validation and test set; TNR, true negative rate; TPR: true positive rate; NPV, negative predictive value; PPV, positive predictive value; resnetX-Y-Z: meta-model obtained by stacking ensemble of the three weak learners resnetX, resnetY, and resnetZ.

Task	Net	MCC_{cv}	MCC_{ts}	TNR_{ts}	TPR_{ts}	NPV_{ts}	PPV_{ts}
N-P	resnet34-50-101	0.976 (0.976,0.976)	0.940	1.000	0.968	0.912	1.000
	resnet50	0.960 (0.954,0.966)	0.937	0.993	0.969	0.915	0.998
	resnet101	0.948 (0.944,0.952)	0.929	0.996	0.964	0.900	0.999
	resnet34	0.954 (0.951,0.956)	0.925	1.000	0.960	0.892	1.000
UC-CD	resnet34-50-152	0.906 (0.905,0.907)	0.688	0.777	0.902	0.875	0.822
	resnet50	0.377 (0.362,0.398)	0.629	0.727	0.890	0.853	0.788
	resnet34	0.428 (0.415,0.454)	0.611	0.744	0.861	0.824	0.793
	resnet152	0.319 (0.319,0.319)	0.602	0.677	0.905	0.862	0.761
UC-N	resnet34-50-101	0.988 (0.988,0.988)	0.931	0.955	0.979	0.983	0.945
	resnet34	0.968 (0.966,0.970)	0.930	0.964	0.968	0.975	0.954
	resnet101	0.964 (0.961,0.970)	0.895	0.967	0.925	0.943	0.956
	resnet50	0.966 (0.962,0.972)	0.855	0.969	0.875	0.909	0.957

learners on the test set (Table 3.5).

The best-performing networks for the UC-N task are a ResNet34 model with a balanced data loader and an ensemble model using ResNet50, ResNet34, and ResNet101 as weak learners trained with unbalanced, balanced, and unbalanced data loaders, respectively. The best metrics are shown in Table 3.5. In this case, the performance improvement of the ensemble model over the single best-performing weak learner ResNet34 is limited.

The performance on the test set is reasonably comparable to that in cross-validation for all tasks and models, except for the UC-CD ensemble model which exhibits higher CV performance (MCC=0.906 vs. 0.688, Fig. 3.4, Table 3.5).

3.5.2 Interpretability

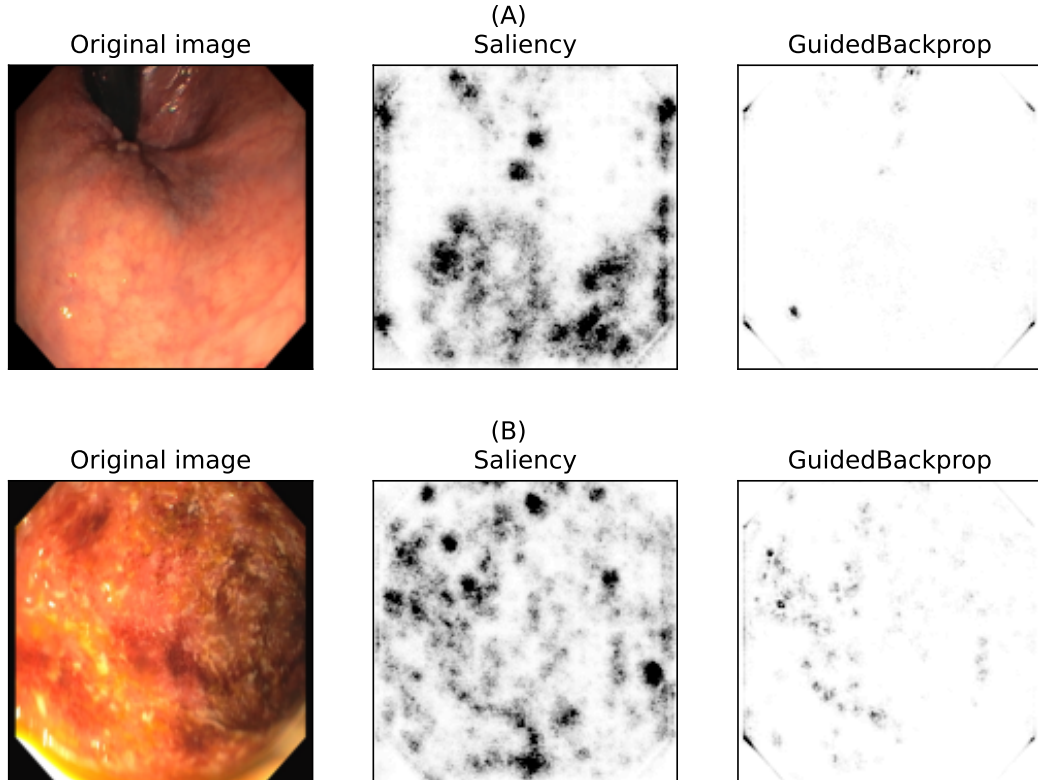
We applied the chosen model interpretation methods to a random subsample of test set N-P images using a ResNet50 model, which was the best performing weak learner on the task N-P (Table 3.5). For two representative images (1 Positive, 1 Negative), Fig. 3.7 visualizes the Saliency and Guided Backpropagation (GuidedBackProp) attribution maps overlaid on the original image, which is also shown separately for

Table 3.3: Metrics values for the best trained nets, ranked by decreasing Matthews Correlation Coefficient (MCC) on the test set; cv, ts: cross-validation and test set; TNR, true negative rate; TPR: true positive rate; NPV, negative predictive value; PPV, positive predictive value; resnetX-Y-Z: meta-model obtained by stacking ensemble of the three weak learners resnetX, resnetY, and resnetZ.

Task	Net	MCC_{cv}	MCC_{ts}	TNR_{ts}	TPR_{ts}	NPV_{ts}	PPV_{ts}
N-P	resnet34-50-101	0.976 (0.976,0.976)	0.940	1.000	0.968	0.912	1.000
	resnet50	0.960 (0.954,0.966)	0.937	0.993	0.969	0.915	0.998
	resnet101	0.948 (0.944,0.952)	0.929	0.996	0.964	0.900	0.999
	resnet34	0.954 (0.951,0.956)	0.925	1.000	0.960	0.892	1.000
UC-CD	resnet34-50-152	0.906 (0.905,0.907)	0.688	0.777	0.902	0.875	0.822
	resnet50	0.377 (0.362,0.398)	0.629	0.727	0.890	0.853	0.788
	resnet34	0.428 (0.415,0.454)	0.611	0.744	0.861	0.824	0.793
	resnet152	0.319 (0.319,0.319)	0.602	0.677	0.905	0.862	0.761
UC-N	resnet34-50-101	0.988 (0.988,0.988)	0.931	0.955	0.979	0.983	0.945
	resnet34	0.968 (0.966,0.970)	0.930	0.964	0.968	0.975	0.954
	resnet101	0.964 (0.961,0.970)	0.895	0.967	0.925	0.943	0.956
	resnet50	0.966 (0.962,0.972)	0.855	0.969	0.875	0.909	0.957

comparison.

Figure 3.7: *Model interpretability*. Qualitative visualization of attribution maps by Saliency and Guided backpropagation (GuidedBackProp) algorithms for a Negative (top row) and Positive (bottom row) input image. The original image is also shown for comparison. The attribution maps are visualized as heatmaps, with darker shades representing more important features according to the algorithm.



We observe that GuidedBackProp attributions are less noisy than Saliency ones, as expected. Moreover, typical endoscopic features of IBD such as mucosal erythemas appear to have higher attribution values according to GuidedBackProp (Fig. 3.7, bottom). Using a gradient-based XAI method such as Saliency and Guided backpropagation is particularly appealing because it allows a straightforward visual interpretation of the input features deemed important by the underlying model. However, the identified features should be the object of further analyses since they may be affected by different kinds of bias, such as illumination (Fig. 3.7).

3.6 Discussion

In this work, we developed and evaluated a prototype DL framework based on ResNet architectures merged by ensemble learning, able to identify disease patterns from endoscopic images of different IBDs, namely Ulcerative Colitis and Crohn's Disease, and distinguish negative (healthy) samples. The DL models achieve a test set performance $MCC=0.940$ for the classification of healthy controls versus IBD patients, $MCC=0.688$ for Ulcerative Colitis versus Crohn's Disease, and $MCC=0.931$ for Ulcerative Colitis versus healthy individuals. On the UC-CD task, we observed a lower, yet relatively good, predictive performance ($MCC=0.688$). This result might be due to the intrinsic difficulty in distinguishing between these two clinical subtypes of IBD, using solely colonoscopy-based diagnosis, whereas additional methods such as endoscopic ultrasonography may be more effective in differentiating UC from CD [121]. It should be noted that the results of the ensemble model demonstrate that there is only a marginal increase in predictive performance compared to using a single ResNet, possibly because it is difficult to further improve over already high-performance results. Overall, the obtained results in all the three classification tasks indicate a very good to excellent predictive performance, highlighting the potential of the framework to evolve into a valuable tool for the clinicians in their diagnostic tasks. Given the several subtleties of endoscopic imaging, linked to both the intrinsic disease features and to potential artifacts (for instance due to light effects), disagreement on the diagnosis is not uncommon among clinicians dealing with this task: the automated system proposed here can provide an additional view on the problem, clarifying the issues hampering the correct assessment of the pathology. Nonetheless, despite the encouraging results, the current study should be considered a proof of concept rather than a consolidated pipeline, which we have already planned to improve in future development. In this regard, additional model architectures other than ResNet could be evaluated, and different loss functions can be used to overcome the data limitation involving unbalanced classes. Moreover, the neural networks could be trained for more epochs, with additional combinations of hyperparameters, including automated approaches to data preprocessing and artefact removal, such as traditional image segmentation [122]. Additionally, alternative ensemble models can be tested, trained with images in different colour spaces. The model outcome can also be improved by enhancing the training data, even through augmenting techniques creating synthetic data: for example, generative adversarial networks. Finally, strengthening the resampling strategy will further improve the overall reproducibility of the study, while the analysis of the data trajectories across the DL layers can provide a valuable direction regarding the model interpretability.

3.7 Conclusions

The results presented in this study demonstrate the vast potential of deep neural networks in discriminating between pathological and negative samples as well as discriminating between IBD subgroups, namely Ulcerative Colitis and Crohn's Disease. Further development of this work and other studies in this area in general have the potential to become powerful tools in the hands of clinicians, aiding diagnosis and the clinical decision process. However, additional studies would be required to evaluate the utility of machine-learning aided diagnosis of IBD in clinical practice.

3.8 List of abbreviations

IBD: Inflammatory bowel diseases
CD: Crohn's disease
UC: Ulcerative colitis
AI: artificial intelligence
XAI: explainable AI
DL: deep learning
ML: machine learning
RGB: red, green, blue
HSV: hue, saturation, value
ReLU: rectified linear unit
ResNet: residual neural network
MCC: Matthew's correlation coefficient
TPR: true positive rate
TNR: true negative rate
PPV: positive predictive value
NPV: negative predictive value

3.9 Declarations

3.10 Ethics approval and consent to participate

The study that produced the data used for these analyses was conducted in accordance with the Declaration of Helsinki and approved by a local ethics committee in Italy, namely the Istituto Tumori Bari "Giovanni Paolo II" - IRCCS, with the protocol number n. 333 / C.E. of 31/07/2019. Informed consent to participate was obtained from the study participants. The team acquired the permission to access and analyze the clinical patient data by GPI S.p.A. through the Contract 210202009864 between GPI S.p.A. and FBK.

3.11 Consent to publish

Not applicable.

3.12 Availability of data and materials

The data generated and analysed during the current study are not publicly available due to privacy restrictions and commercial confidentiality, however they can be made available in a de-identified manner upon reasonable request and an appropriate agreement, contacting the corresponding author.

3.13 Competing interests

GPI S.p.A. took part in the writing of the manuscript.

3.14 Funding

This work was supported by GPI S.p.A. through the Contract 210202009864 between GPI S.p.A. and FBK within the framework project “*SI-CURA – Soluzioni Innovative per la gestione del paziente e il follow up terapeutico della Colite UlceRosA*” funded by Regione Puglia with the POR Puglia FESR-FSE 2014-2020 Innonetwork. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

3.15 Authors’ contributions

MC performed the XAI analysis, prepared figures and tables, and coordinated the revision of the paper after reviewing. NP preprocessed the data, wrote the original scripts, ran the predictive modeling, and contributed to an earlier version of the manuscript. MC, VO, GJ coordinated the project, interpreted the results, and wrote the manuscript. MP implemented the k-fold cross-validation, ran the predictive modeling, and contributed to XAI analysis. AC1, MDD, AC2 provided access to the endoscopic imaging data and interpreted the results. All authors read and approved the final version of the manuscript.

3.16 Acknowledgements

A shorter version of this work has been previously presented online during the CIBB 2021 conference (<https://davidechicco.github.io/cibb2021/papers.html>).

3.17 Figures

Figure 3.1 *Experimental workflow.* After an upstream processing of the input images (blocks “Data splitting”, “Class balancing”, “Transform”), for each classification task three pre-trained ResNet variants of increasing complexity (“Model i ”, $i = 1, 2, 3$) are used as weak learners and fine-tuned in 5-fold cross-validation on the input data (see Methods). A meta-model is then built by stacking ensemble of the three weak learners, evaluating the performance on the external test set in terms of different classification metrics (see Methods).

N-P, negative versus positive; UC-CD, Ulcerative Colitis versus Crohn’s Disease; UC-N, Ulcerative Colitis versus negative; MCC, Matthews Correlation Coefficient (MCC); TNR, true negative rate; TPR, true positive rate; NPV, negative predictive value; PPV, positive predictive value

Figure 3.2 An example of positive (a) and negative (b) images within the SICURA dataset.

Figure 3.3 *N-P classification results.* Metrics values for the ensemble model (resnet34-50-101) and the weak learners resnet34, resnet50, and resnet101. Light blue: cross-validation; dark blue: test set. Black bars represent 95% confidence intervals.

Figure 3.4 *UC-CD classification results.* Metrics values for the ensemble model (resnet34-50-152) and the weak learners resnet34, resnet50, and resnet152. Light blue: cross-validation; dark blue: test set. Black bars represent 95% confidence intervals.

Figure 3.5 *UC-N classification results.* Metrics values for the ensemble model (resnet34-50-101) and the weak learners resnet34, resnet50, and resnet101. Light blue: cross-validation; dark blue: test set. Black bars represent 95% confidence intervals.

Figure 3.6 *Confusion matrices.* Test set confusion matrices for the best-performing weak learners of each classification task (see Table 3.5).

Figure 3.7 *Model interpretability.* Qualitative visualization of attribution maps by Saliency and Guided backpropagation (GuidedBackProp) algorithms for a Negative (A, top row) and Positive (B, bottom row) input image. The original image is also shown for comparison. The attribution maps are visualized as heatmaps, with darker shades representing more important features according to the algorithm.

3.18 Tables

Table 3.4: Number of elements (images) in each class, in the format “raw data (preprocessed data)”. P, positive; N, negative; UC, Ulcerative Colitis; CD, Crohn’s disease; IBD, inflammatory bowel disease.

P			N
UC	CD	IBD	
4,388 (3,594)	5,949 (4,098)	1,067 (823)	2,822 (2,815)

Table 3.5: Metrics values for the best trained nets, ranked by decreasing Matthews Correlation Coefficient (MCC) on the test set; cv, ts: cross-validation and test set; TNR, true negative rate; TPR: true positive rate; NPV, negative predictive value; PPV, positive predictive value; resnetX-Y-Z: meta-model obtained by stacking ensemble of the three weak learners resnetX, resnetY, and resnetZ.

Task	Net	MCC_{cv}	MCC_{ts}	TNR_{ts}	TPR_{ts}	NPV_{ts}	PPV_{ts}
N-P	resnet34-50-101	0.976 (0.976,0.976)	0.940	1.000	0.968	0.912	1.000
	resnet50	0.960 (0.954,0.966)	0.937	0.993	0.969	0.915	0.998
	resnet101	0.948 (0.944,0.952)	0.929	0.996	0.964	0.900	0.999
	resnet34	0.954 (0.951,0.956)	0.925	1.000	0.960	0.892	1.000
UC-CD	resnet34-50-152	0.906 (0.905,0.907)	0.688	0.777	0.902	0.875	0.822
	resnet50	0.377 (0.362,0.398)	0.629	0.727	0.890	0.853	0.788
	resnet34	0.428 (0.415,0.454)	0.611	0.744	0.861	0.824	0.793
	resnet152	0.319 (0.319,0.319)	0.602	0.677	0.905	0.862	0.761
UC-N	resnet34-50-101	0.988 (0.988,0.988)	0.931	0.955	0.979	0.983	0.945
	resnet34	0.968 (0.966,0.970)	0.930	0.964	0.968	0.975	0.954
	resnet101	0.964 (0.961,0.970)	0.895	0.967	0.925	0.943	0.956
	resnet50	0.966 (0.962,0.972)	0.855	0.969	0.875	0.909	0.957

Chapter 4

Synthetic data in digital pathology

4.1 Summary

DOI: 10.1038/s41598-024-79602-w

Introduction This paper addresses the shortage of annotated whole-slide images (WSIs) for training AI models in pathology by generating synthetic data using Denoising Diffusion Probabilistic Models (DDPMs). Synthetic data offers a scalable alternative for overcoming the challenges of data scarcity and high annotation costs.

Methods

1. Data Generation: Synthetic WSIs are created by training DDPMs on existing datasets of histopathological images.
2. Validation: The study employs multiple strategies to evaluate synthetic data:
 - Visual quality assessment by domain experts.
 - Quantitative similarity metrics to compare synthetic and real WSIs.
 - AI performance comparison using synthetic vs. real WSIs in training tasks.

Results

- Synthetic WSIs closely resemble real ones in both visual and quantitative metrics.
- Training AI models with synthetic data yields comparable performance to real data in downstream tasks.
- Domain experts verify that the synthetic data maintains biological plausibility and utility for diagnostic tasks.
- Including heterogeneous evaluation strategies reveal complementary insight on the quality of the generated dataset.

Discussion Synthetic data presents a promising avenue for reducing reliance on extensive manual annotations and broadening training datasets. However, ensuring the quality and representativeness of synthetic images is crucial for practical applications. Further development in generating domain-specific synthetic data is necessary.

Generating and evaluating synthetic data in digital pathology through diffusion models

Matteo Pozzi^{1,2†}, Shahryar Noei^{1†}, Erich Robbi^{1,3}, Luca Cima⁵, Monica Moroni¹, Enrico Munari⁵, Evelin Torresani⁴, Giuseppe Jurman^{1*}

¹ Fondazione Bruno Kessler, Data Science for Health Unit, Via Sommarive 18, 38123 Povo, Trento, Italy.

² Università degli Studi di Trento, Department for Computational and Integrative Biology, Via Sommarive 9, 38123 Povo, Trento, Italy.

³ Università degli Studi di Trento, Department of Information Engineering and Computer Science, Via Sommarive 9, 38123 Povo, Trento, Italy.

⁴ Pathology Unit, Department of Laboratory Medicine, Santa Chiara Hospital, APSS, Trento, Italy.

⁵ Department of Pathology and Diagnostic, University and Hospital Trust of Verona, Verona, Italy.

† These authors contributed equally to this work.

* Corresponding author: Giuseppe Jurman, Data Science for Health Unit, Fondazione Bruno Kessler, Via Sommarive 18, 38123 Povo, Trento, Italy, giuseppe.jurman@fbk.eu (+39) 0461-314523.

4.2 Abstract

Synthetic data is becoming a valuable tool for computational pathologists, aiding in tasks like data augmentation and addressing data scarcity and privacy. However, its use necessitates careful planning and evaluation to prevent the creation of clinically irrelevant artifacts.

This manuscript introduces a comprehensive pipeline for generating and evaluating synthetic pathology data using a diffusion model. The pipeline features a multifaceted evaluation strategy with an integrated explainability procedure, addressing two key aspects of synthetic data use in the medical domain.

The evaluation of the generated data employs an ensemble-like approach. The first step includes assessing the similarity between real and synthetic data using established metrics. The second step involves evaluating the usability of the generated images in deep learning models accompanied with explainable AI methods. The final step entails verifying their histopathological realism through questionnaires answered by professional pathologists. We show that each of these evaluation steps are necessary as they provide complementary information on the generated data’s quality.

The pipeline is demonstrated on the public GTEx dataset of 650 Whole Slide Images (WSIs), including five different tissues. An equal number of tiles from each tissue are generated and their reliability is assessed using the proposed evaluation pipeline, yielding promising results.

In summary, the proposed workflow offers a comprehensive solution for generative AI in digital pathology, potentially aiding the community in their transition towards digitalization and data-driven modeling.

4.3 Introduction

In recent years, rapid advancements in scanning technologies have revolutionized traditional pathology, giving rise to the field of digital pathology [50, 123, 124],,. Instead of traditional glass slides, the focus now shifts to whole slide images (WSIs)—high-resolution digital representations of complete tissue sections. To create WSIs, tissue specimens undergo meticulous scanning using specialized slide scanners. These scanners capture every microscopic detail, including cell structures, tissue patterns, and staining variations. The resulting WSIs serve as the digital counterparts of physical glass slides [125]. Pathologists can now review WSIs remotely, breaking geographical constraints. Collaborative efforts among pathologists occur virtually, with insights shared and challenging cases discussed. Deep learning algorithms have demonstrated impressive capabilities in automating critical tasks such as cancer detection or grading [126, 127, 128, 129] as well as cell segmentation [130, 131, 132]. However, the success of these deep learning algorithms heavily relies on the availability and quality of large-scale, annotated datasets [133], which can be challenging in the medical domain where data scarcity and privacy concerns are prominent [134]. To address this issue, synthetic data generation has gained momentum, by producing data that maintains similar statistical properties as real data, making it a promising avenue to tackle privacy concerns while working with sensitive data such as medical images [135, 136]. We explore this approach by using synthetic data as a standalone solution by simulating a scenario in which we anonymize proprietary medical data, enabling institutions to share and utilize data without compromising patient privacy.

At the heart of synthetic data generation lie generative models, such as Generative Adversarial Networks (GANs) [137]. GANs have found widespread application in generating synthetic images across various medical domains [138, 139, 140], including digital histopathology images [141, 142, 143, 144]. However, recent advancements have shown that diffusion models [145] have surpassed GANs in generating both natural and medical images [146, 147]. Diffusion models are a class of generative models inspired by non-equilibrium thermodynamics. These models employ a two-step process to learn the data distribution to generate new data afterwards. In the first step, a Markov chain of diffusion steps is defined, where noise is progressively added to the input. Then, the model learns to reverse this process by generating novel data starting from the noisy inputs [145]. The usage of diffusion models in medical image generation includes (but is not limited to) brain magnetic resonance images [148, 149], microscopic blood cells images [150], positron emission tomography heart images [151], and chest x-ray [152]. A comprehensive review on the subject has been recently published [153]. The first application of

diffusion models in generating histopathology images dates to [154], where the authors successfully employed denoising diffusion models to generate Hematoxylin and Eosin (H&E) tiles of three different subtypes of gliomas. Their results demonstrated that diffusion models outperform GANs for this task based on common quantitative evaluation metrics used in image generation quality assessments. Following their success, the authors in [155] also demonstrated the ability and superiority of diffusion models over GANs in generating colon histopathology images containing six different cell nucleus types and their corresponding masks. Additionally, authors in [156] used vision transformer-based diffusion models to generate colon histopathology images with higher quality than those generated using GANs. Before exploiting generated data it is crucial to evaluate them. However, the evaluation of generated data remains a topic of debate, with various studies adopting different approaches. The lack of a standardized evaluation pipeline to comprehensively assess generated data poses a major caveat. In this study, we aim to tackle this issue by introducing a general framework for generating and, more importantly, thoroughly evaluating synthetic pathology images from various viewpoints. We will illustrate how these multifaceted evaluations yield complementary insights that cannot be obtained through any single method alone. The proposed synthetic pathology data generation and evaluation pipeline, as demonstrated in Figure 4.1, begins with the preprocessing stage, where tiles are extracted from WSIs. Subsequently, a diffusion model is trained to generate data conditioned on different classes. After the training process, a sampling phase is performed to generate a specific number of synthetic data for each class. Lastly, the evaluation step process encompasses three distinct sets of assessments. The first set involves quantitative metrics that measure the similarity between the generated and real images. For our quantitative evaluation, we employed well-established distribution based metrics commonly used in the field of synthetic images, including Inception Score (IS) [157], Fréchet Inception Distance (FID) [158] improved precision-recall [159], and density-coverage [160] as well as IL-NIQE no-reference image quality assessment method [161]. The second set evaluates the practical usability of the generated images in deep learning models. To complement this set we investigated the convolutional filters learned by the classifiers, trained on both real and generated data, using the Concept Relevance Propagation (CRP) algorithm [162], an explainable artificial intelligence approach. Lastly, a qualitative assessment is conducted to validate the biological realism of the generated data, a crucial aspect not targeted in previous assessments. This comprehensive approach ensures the reliability and usefulness of the generated dataset for our selected task. We applied the pipeline on 650 WSIs of five different tissues derived from the Genotype-Tissue Expression (GTEx) [163]. We specifically chose to include multiple tissues, since most previous studies have focused on only one tissue type [154, 155, 156]: such conservative approach fails to demonstrate whether similar success can be expected for other tissue types with varying complexities of morphology.

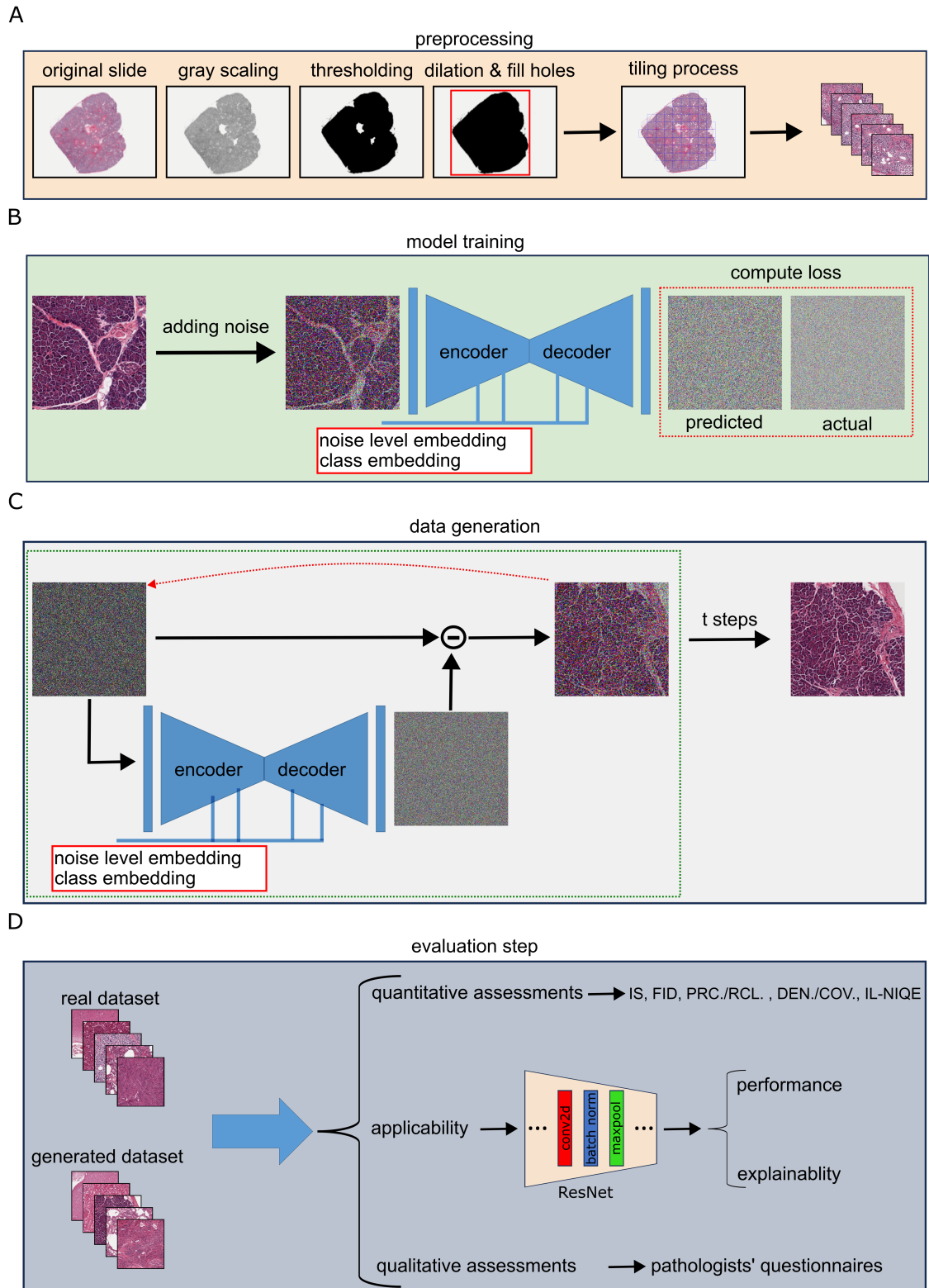


Figure 4.1: The proposed pipeline. **(A)** Demonstrates the preprocessing step, where tissue is detected in a slide, followed by tiling to extract batches of tiles. **(B)** Illustrates the model training step, where extracted tiles are input into a denoising diffusion model, with noise initially added to images. The U-Net model predicts the added noise using embeddings of noise level and tissue type to compute the loss and drive the training process. **(C)** Depicts the data generation step, where the trained model predicts noise, which is subtracted from the original noise to create a less noisy image. This process is repeated for t steps to generate the final image. **(D)** The generated dataset undergoes evaluations using quantitative metrics such as Inception Score (IS), Frechet Inception Distance (FID), precision-recall (PRC, RCL), and density-coverage (DEN, COV) to compare real and generated datasets. IL-NIQE assesses image quality, and ResNet classifiers evaluate biological realism via expert pathologists’ assessments.

To the best of our knowledge, this study represents a novel and reproducible approach to histopathology image generation and evaluation. What sets our work apart is the thorough evaluation of the generated data from diverse angles, encompassing different tissue types and utilizing a range of evaluation metrics. Through our emphasis on multifaceted evaluation, including the pioneering application of explainability methods, our objective is to offer a comprehensive understanding of the capabilities and limitations of the proposed generative model within the digital pathology domain.

4.4 Materials and Methods

4.4.1 Dataset

Images were obtained from the GTEx study [164]. GTEx dataset contains WSIs collected using an Aperio glass slides scanner using H&E staining [163].

Among the 53 available human tissues, we retained data associated exclusively to samples obtained from our designated tissues of interest, namely the Brain, Kidney, Lung, Pancreas, and Uterus. This filtering process resulted in approximately 5000 samples’ metadata. Subsequently, we randomly selected a sample set of 650 specimens (subjects) from the filtered metadata to obtain their corresponding WSIs, resulting in one slide per subject. In cases where a WSI was not available for downloading, the specimen was excluded from the final dataset (the list of the downloaded WSIs is available in Supplementary Table S1).

Out of the 648 downloaded WSIs, a total of 589 have been used as training set for the deep learning models, while the remaining ones have been exclusively used for testing purposes.

4.4.2 Preprocessing pipeline

WSIs demonstrate a considerably high spatial resolution: in our case, a spatial base resolution of 0.494 μm per pixel on a 20x magnification factor. Such granularity makes them impractical as direct inputs for neural networks.

It is commonly recommended to divide a WSI into uniform-sized tiles which can then be utilized as input for deep learning models [165, 166]. Therefore, a preprocessing stage was deemed necessary to adequately prepare the collected data for model training.

In our study, we employed the Histolab library [167] to facilitate the tiling process of the slides. This library offers tools to generate tissue masks and perform the tiling process [167].

The WSIs do not necessarily contain only tissue information. Therefore, the creation of a tissue mask given a slide is a crucial step, as it ensures that the generated tiles primarily contain tissue-related information while minimizing extraneous background areas. We observed that the following ordered steps (depicted in Figure 4.1A) define an adequate tissue mask:

1. **RgbToGrayscale filter:** This process involves the conversion of a 3-channel image into a single channel grayscale image.
2. **OtsuThreshold filter:** Given a grayscale image, returns a binary thresholded image with separated foreground and background [168].
3. **BinaryDilation:** Performs a dilation on a binary image, i.e., the output image's pixel value is determined by assigning it the maximum value among its neighboring pixels.
4. **BinaryFillHoles filter:** This process involves the task of filling all regions within a binary image.

We have carefully selected parameters to define the attributes of the tissue mask extractor. For the OtsuThreshold image filter, we have chosen the operator.lt parameter, which performs the “less than” operation between the pixels of the image and the estimated Otsu's threshold. Additionally, we employed a disk size of 5 and performed a single iteration for the BinaryDilation operation. Furthermore, we utilised the pre-set structural element (i.e. an element with a square connectivity equal to one) for the BinaryFillHoles operation.

In addition to the mask generation, other parameters are required in order to tile a WSI:

- **Tile Size:** Specifies the final resolution of the extracted tiles in pixels.
- **Level:** Specifies the level of tile extraction in a slide, the lower it is (minimum value 0), the higher the magnification factor will be.
- **Pixel Overlap:** Specifies in pixels how much the tiles will overlap between each other.

We decided to extract non-overlapping tiles with a resolution of 512x512 pixels from slides at the magnification factor of 5x (therefore with resolution of 1.976 μm per pixel). These parameters were chosen upon consultation with pathologists aiming to optimize the tissue classification task, providing sufficient contextual information for tissue differentiation, and avoiding excessive zooming that could render the tiles unidentifiable.

This resulted in 63247 tiles that were divided in two sets: from the 589 training WSIs, 56990 tiles have been extracted and used for training the generative model and the classifier, while the remaining 6257 tiles have been extracted from the 59 test WSIs and used for testing purposes. Nevertheless, even with a precisely defined mask, not all generated tiles may consistently capture informative attributes of the original slide. As illustrated in (the left panel of) Supplementary Figure S1, some tiles lack discernible morphological characteristics of the tissue, indicating that they may not contribute valuable information for the models. To filter such tiles, we exploited the Complexity metric C as inspired by [169] and it is defined as equation 4.1:

$$C_t = \frac{1}{NM} \sum_{(x,y)} |G_x(x,y) + G_y(x,y)| \quad (4.1)$$

where the pixel-wise gradient (G) in the horizontal (x) and vertical (y) directions is computed for each tile t . The total sum of gradients is then multiplied by a term that is dependent on the width N and height M of the tile. A tile t with a non-uniform tissue will exhibit elevated pixel gradient values, thereby resulting in a correspondingly high value for C_t .

We applied a chosen threshold (i.e., 0.4) based on visual inspection to retain only the tiles that exhibited sufficient distinguishable features. This step resulted in the removal of 948 uninformative tiles, allowing us to retain a total set of 62299 tiles, ranging from 10891 for Brain to 14173 for Pancreas. Additional details are summarized in Supplementary Table S2.

4.4.3 Denoising Diffusion Probabilistic Model

For our generative model, we utilized the Denoising Diffusion Probabilistic Models (DDPM) proposed by [145]. DDPM is a specific variant of diffusion models that focuses on denoising the input data by learning a probabilistic model capable of generating high-quality outputs. To create a dataset covering five different tissue types, we followed the classifier-free-guidance approach outlined in [147]. In short, the pipeline works as follows.

First, Gaussian noise is introduced into the input images. The magnitude of this noise is determined randomly by a scheduler, resulting in varying levels of noise being applied to the input images. This approach allows the model to explore a spectrum of image noises, ranging from subtle distortions to an isotropic Gaussian distribution. The noisy image along with an embedding vector representing the level of noise and a trainable embedding vector representing the class (in our case the tissue type) are fed to a U-Net [170] based neural network, which aims to learn to output the noise. This prediction is then compared to the actual added noise, forming the basis for calculating the loss and guiding the model's training process. This process is illustrated in Figure 4.1B.

Once the training phase is complete, the image generation phase begins. Starting with a baseline of pure isotropic noise, the trained network predicts the corresponding noise level. The difference between this predicted noise and the initial noise results in the creation of a less noisy image, which then becomes the input for the network once again. This iterative procedure is repeated for t steps, resulting in the generation of a final image, as illustrated in Figure 4.1C.

The model implementation can be found online. The images were resized to 256 pixels, as experimenting with different sizes revealed that this produced the most favorable outcomes. For model training, we employed a batch size of 26. The learning rate was set at 1e-5, and the training procedure encompassed 215000 steps. We empirically found the learning rate by monitoring the sampled data after every 1000 steps. The number of steps for data generation (t) was set to 250. Visual inspection of image generation and model checkpointing was done every 1000 steps. Training was done till a convergence in the training loss was noticed as well as sampling during the training provided convincing results, since it has been observed that the quality of the generated images can be improved even when the training loss was stable. Class conditioning was done using a trainable `torch.nn.Embeddings` variable. Lastly, the generated images were further enhanced using a sharpening filter with a factor of 3.

4.4.4 Evaluations

To comprehensively evaluate the quality of the generated images from various perspectives, we devised multiple sets of experiments ranging from widely recognized metrics for assessing generated images quality as well as expert human evaluation. In the following, we will give an overview of the details of each of these experiments.

Canonical quantitative evaluation metrics

We assessed the quality of the generated images using well-established metrics in the field of synthetic images, including Inception Score (IS) [157], Fréchet Inception Distance (FID) [158], improved precision-recall [159], and density-coverage [160] as well as IL-NIQE no-reference image quality assessment method [161]. Here, we will briefly explain each metric.

The IS is calculated using equation 4.2.

$$IS = \exp \left\{ \mathbb{E}_{x \sim P_g} \left[D_{KL}(p(y|x) || p(y)) \right] \right\} \quad (4.2)$$

where y is a set of classes (in our case tissues), x is the image, $p(x|y)$ is the conditional probability, $p(y)$ is the marginal class probability, D_{KL} stands for the Kullback Leibler divergence and $\mathbb{E}_{x \sim P_g}$ denotes average over all the generated images.

where y is a set of classes (in our case tissues), x is the image, $p(x|y)$ is the conditional probability, $p(y)$ is the marginal class probability, D_{KL} stands for the Kullback-Leibler divergence, and $\mathbb{E}_{x \sim P_g}$ denotes average over all the generated images.

High values of IS indicate that the generated images have a high diversity across different classes while maintaining good within-class similarities [?]. We utilized the `torchmetrics` implementation to calculate IS in our study, splitting the images 5 times and using the Inception v3 network [171] unbiased logits as the features to calculate the conditional and marginal probabilities.

Fréchet Inception Distance (FID) is a measure to quantify the similarity between the distributions of the generated and real images [158]. It achieves so by passing the generated and real images through the original Inception v3 network, extracting the outputs from a specified layer, and calculating equation 4.3:

$$FID = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}}) \quad (4.3)$$

Here, μ_r and Σ_r represent the estimated average and covariance matrices of real images, while μ_g and Σ_g denote the same for generated images in the Inception_v3 feature space. The trace operation is denoted by Tr . To calculate FID, we utilized the `torchmetric` implementation, setting the `feature` parameter to 2048, which corresponds to the last pooling layer. Lower values of FID indicate that the distributions of the real and generated images are more similar [?].

Both FID and IS rely on the imagenet [172] pre-trained Inception v3 network which might not be ideal for those images whose distribution significantly differs from natural images [173, 174]. To account for this, we additionally incorporated methods that assess the quality of generated images by examining the similarity of the manifolds between the real and generated data. The authors in [159] have proposed improved precision and recall measuring the overlap of the manifolds of the generated and real data. Briefly, they are calculated by projecting the data to a feature space and estimating the manifolds for real and generated data by surrounding each datapoint with a hyper-sphere that reaches its k th nearest neighbor [159]. Precision is defined as the percentage of generated samples within the real data manifold, while recall is defined as the percentage of real images that fall within the generated data manifold [159]. Higher values for both indicate greater overlap between the real and generated data. Precision and recall have been demonstrated to suffer in the presence of outliers [160]. To overcome this problem, the authors in [160] have proposed the density and coverage metrics, which offer improved manifold estimation (for a more detailed explanation, please refer to [160]).

In this study, we calculated the precision, recall, density and coverage using the code provided by [160]. We set the number of neighbors to 10. To prevent any potential bias that may rise due to networks being trained on specific datasets [160], the feature space for real and generated data was created by projecting the data onto a 100-dimension space through a ResNet-50 network [110] with random weights (for result comparability, we also performed analyses using an ImageNet pretrained network). We modified the last fully connected layer of the network to output a 100-dimensional vector. Since the imbalance between the number of real and generated images can bias the estimation [160], we addressed this issue by randomly sampling an equal number of generated images per tissue from the real data. To obtain reliable estimates, we repeated the entire process five times and reported the mean and standard error of the mean.

IL-NIQE [161] is a blind image quality assessment (BIQA) method which aims to evaluate the quality of images and videos without the need for a reference, non distorted image. The quality score is computed by a modified Bhattacharyya distance, which measures the deviation of the statistics of the input image from a reference set of statistics learned from a high quality set of natural images. This process is not done for the whole image but is done after dividing the image in patches and then averaging each quality score over the entire image. To compute the IL-NIQE score we used the PyTorch Toolbox for Image Quality Assessment python package. We computed IL-NIQE score for each image in both datasets and then reported the average score for each dataset. Moreover, we also computed the IL-NIQE score for each class of both synthetic and real dataset.

Deep learning-based practical assessments

While the aforementioned evaluation measures provide valuable insights into the quality of generated images and their similarity to real ones, they may not provide sufficient information about more complex morphological features (such as histopathological plausibility) or the suitability of the generated images to be used as a replacement to the real images in machine learning studies. These aspects are of utmost importance and are one of the primary focuses of our current study. Therefore, additional evaluation criteria are required. To do so, we designed a set of experiments using deep convolutional neural networks (CNNs) [175].

We used the ResNet-50 architecture [110] with randomly initialized weights to train a tissue classifier in two different scenarios. In the first scenario, we trained the network from scratch using the real tiles employed to train the diffusion model. Based on results demonstrated in [176], we put aside 30% of slides specifically for internal validation, resulting in 39294 tiles for training. This approach reduces the potential of data leakage related biases in the evaluation of models. To ensure that the network learns relevant morphological features associated with tissue types, we applied a set of following augmentations during training: spatial transformations, including random rotation and random horizontal-vertical flipping, chromatic transformations, such as jittering and random channel swapping, to mitigate the effects of staining [177], and robustness enhancing augmentations like blurring using a Gaussian filter and random erasing of parts of the image. The performance was evaluated on both the real tiles reserved for testing (check Materials and Methods, Dataset section) and the generated tiles. Additionally, we retrained the network on a subset of the real tiles, equivalent in number to the generated tiles, to compare the results and eliminate the possibility of performance differences due to varying dataset sizes.

The network was trained for 100 epochs with a batch size of 32. We employed the AdamW optimizer [178] with an initial learning rate of 0.0002. The learning rate was decreased using a cosine annealing scheduler with a t_{max} parameter of 100 [179]. The model corresponding to the highest accuracy on the validation was saved for further steps. The model and training were implemented using pytorch and pytorchlightning.

In the second scenario, adopting the exact same training setting as the first scenario, we trained a network from scratch using the generated tiles. The only distinction was that, since the generated images were on a tile level and not slide level, we reserved 30% of the tiles for internal validation, resulting in 10500 tiles for training. Subsequently, the trained network was evaluated on two distinct test sets. The first test set comprised a randomly selected 15% subset of the slides utilized in training of the diffusion model (referred to as the internal test-set). In contrast, the second test set consisted of real slides that were not seen by the diffusion model (the same test set used to assess the performance of the network trained on real data).

We evaluated the performance of the models on the test sets using multi-class accuracy (ACC) and Matthews’s correlation coefficient (MCC) [180, 181, 182].

The explainability analysis

We employed Concept Relevant Propagation (CRP) [162] as a means to assess whether the ResNet-50 utilized similar concepts, represented by convolution filters in CNNs, to classify images belonging to the same class, regardless of the dataset

it was trained on. This analysis was crucial to ensure that the classifier employed consistent strategies in identifying the tissue of the tile, regardless of its origin (real or generated). The insights gained from this analysis provided valuable information for evaluating both the classifier itself and the quality of the synthetic tiles generated by the diffusion model.

CRP combines local and global approaches to provide robust and comprehensive explanations of a model’s predictions. It is based on Layer-wise Relevance Propagation (LRP) [183] but extended to disentangle explanations based on concepts learned by the network. CRP computes concept-conditional heatmaps to locate activation of concepts in the input space and provides dataset-level statistics to investigate how different classes or images activate specific channels in the model.

We extracted important concepts for each class from the network trained on the generated dataset by randomly selecting one generated image per class and choosing the convolutional layer 4 which, in the ResNet-50 architecture, is the last convolutional before the output. This choice is justified as we aimed to identify higher-level concepts, which are typically encoded in deeper layers closer to the output of the network [184, 185]. The same process was done for the network trained on the real data.

Afterwards we constrained the relevance propagation by conditioning it on both the class and the layer. This allowed us to determine the significance of each channel in the selected layer for accurately classifying the given tile as belonging to its corresponding class.

In our analysis, we extracted a list of the top 5 channels per class per model, sorted by relevance. The most relevant concept per class were further utilized for feature visualization, where we obtained the most representative samples for each concept from both synthetic and real datasets, along with attribution heatmaps. Additionally, we assessed the prevalent class associated with these concepts, evaluating the integrity and efficacy of the concept extraction procedure. This analysis also helped in assessing the generalizability of the findings by examining whether the same associations hold true when employing the identical model on distinct datasets. This analysis was done using the CRP python package.

Qualitative evaluations

It is acknowledged that human expert evaluation remains the ultimate benchmark, particularly in the case of medical images. For this reason, we asked three professional pathologists (LC, EM, ET) to assist us in evaluating the quality of our generated images.

We did so by designing two different questionnaires. The first half of the first questionnaire included 100 questions. Pathologists were presented with three tiles from a single tissue and tasked with identifying the corresponding tissue. Half of the questions featured tiles derived from real data, while the remaining half consisted of generated images. The tiles “type” (real vs generated) was unknown to pathologists, the questions were randomized, and an equal number of tissues were included. The aim was to assess the pathologists’ ability to detect tissues using a limited number of tiles and to examine potential variations in performance between real and generated data. The second half of the first questionnaire consisted of an additional 100 questions. This time, pathologists were required to determine

whether a given tile was real or generated, based on a single tile. The number of questions was equal for each tissue and data type (real or generated). Similar to the previous section, the questions were randomized across tissues and the distinction between real and generated data. All the tiles were randomly selected, but they were manually reviewed to ensure diversity. The examples of the questions from the first questionnaire are depicted in Supplementary Figure S2.

The follow up questionnaire was sent one week after receiving responses to the first questionnaire. Pathologists were asked to rate from 0 (not important) to 5 (very important) how much they had employed some specific image features in distinguishing between real and generated tiles. The aim of this phase was to figure out whether the pathologists' approach predominantly relied on "visual attributes", such as blurriness, chromatic features, or the presence of artifacts, or if they also found biological or morphological irregularities in their discrimination between real and generated images. Out of the seven questions, four corresponded to visual attributes and three corresponded to biological attributes. The full list of questions is available in Supplementary Table S3.

4.4.5 Implementation

All data processing and model training codes were implemented in Python (version 3.7.16 or higher). Any specific packages used were specified in the corresponding Materials and Methods section. For the evaluation questionnaires, we employed JavaScript and the Google Apps Script APIs.

The pre-processing, classifier training and evaluations were performed on a single machine with 2x Nvidia GTX-1080Ti GPUs, an Intel i7-6700 CPU, and 32GB RAM. The diffusion model training was conducted on a cluster with 2 x NVIDIA V-100 GPUs with 32GB VRAM.

All the codes are available in the Git-Hub repository.

4.5 Results

We used the set of 56990 extracted real tiles as input for the diffusion model and performed training over 215000 steps. Confirming a converged loss and preliminary visual assessments of image quality, we produced 3001 synthetic images per tissue. This choice was motivated by our available resources and time constraints. A sample pair of real and generated images for each tissue are depicted in Figure 4.2.

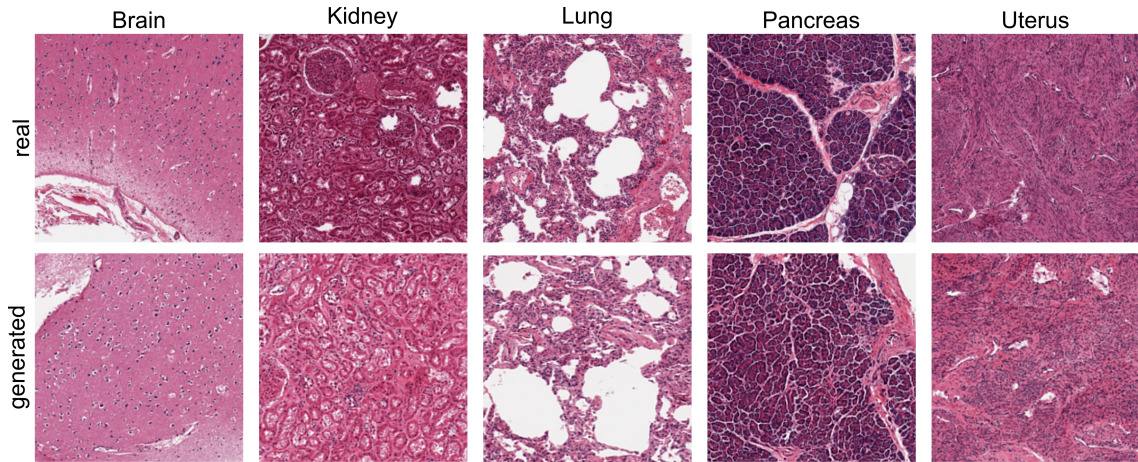


Figure 4.2: Example of real and generated image pair for each tissue. The top row displays a representative image from the real dataset for each tissue, while the bottom row showcases a generated image from the same tissue.

After our initial visual examinations, which confirmed a strong resemblance between the generated images and actual data, we moved forward with our evaluation pipeline.

4.5.1 Measuring the Similarity Between Generated and Real Images Reveals Generated Data Resembles Real Images

After generating the synthetic dataset, we conducted a quantitative assessment of its quality using various synthetic image evaluation metrics. These results are summarized in Table 4.1.

Table 4.1: Results of quantitative evaluation metrics. The table presents the values for quantitative metrics, including the Inception Score (IS), precision, recall, density, coverage, Fréchet Inception Distance (FID), and IL-NIQE. The values for IS, precision, recall, density, and coverage are reported as the mean $\pm$ standard error of the mean based on repeated calculations. FID is calculated using the entire dataset. The IL-NIQE score is calculated as the average of IL-NIQE scores for all generated data. The arrows indicate whether higher ($\uparrow$) or lower ($\downarrow$) values are more desirable. Further details regarding the calculation and interpretation of these metrics can be found in the Materials and Methods section.

Metric	IS ($\uparrow$)	FID ($\downarrow$)	Precision ($\uparrow$)	Recall ($\uparrow$)
Value	2.9115 ± 0.0339	35.1140	0.9620 ± 0.0033	0.9385 ± 0.0091

Metric	Density ($\uparrow$)	Coverage ($\uparrow$)	IL-NIQE ($\downarrow$)
Value	0.7877 ± 0.0186	0.8516 ± 0.0101	59.66

The overall results indicate that the generated images closely resemble the real images. Our generated images achieved a high Inception Score (2.9115 ± 0.0339 , typical reported values are between 2 to 3 where higher values are better), surpassing previous attempts at generating synthetic digital pathology images [154, 155, 156]. However, the Inception Score achieved with our generated images is still lower to the score that would have been achieved with our real dataset (3.2892 ± 0.0149). The FID value, although slightly higher, is still comparable to values reported in other studies (FID values in synthetic pathology typically range from 10 to 40 where lower values are more desired) [154, 155, 156, 186]. It is important to note that these studies utilized a smaller field of view, which can result in a lower FID. For instance, in [155], increasing the field of view by a factor of two (from x20 to x10, the only experiment equivalent to our field of view) more than doubled the FID (reaching a value of 38.1), which exceeds our reported FID. Additionally, based on results reported in [156], changing the dataset results in a significant change in the achieved FID. The precision and recall values approach the maximum value of 1 and outperform previous work [154, 156]. However, it is important to consider that these studies calculated precision and recall using ImageNet pretrained networks, which can impact the results [160]. To ensure a fairer comparison, we recalculated these values using an ImageNet pretrained ResNet50. The results demonstrate that our precision outperforms the other studies, while our recall remains comparable to most studies. Nevertheless, there is a significant decrease in recall compared to [34.], although this decrease is accompanied by a substantial increase in precision. The summary of these comparisons can be found in Supplementary Table S4. To the best of our knowledge, no previous study has reported density and coverage metrics, but our values are also close to the maximum achievable value of 1, indicating a high similarity between the generated images and real ones. Similarly, to our knowledge, IL-NIQE has not been previously utilized to evaluate the quality of generated pathology images (However, typical IL-NIQE values for real images range from around 30 to 60, where lower is better). The average score obtained on the generated data (59.66) is higher than the average score obtained from the real data (33.25), suggesting that the generated data may appear less realistic than actual data, as anticipated. Regardless, since to our knowledge there are no other studies reporting IL-NIQE scores in the evaluation of diffusion models, direct assessment of the level of realism remains challenging. However, it is important to note that direct comparisons between studies may not be completely fair and feasible due to differences in datasets, tile sizes, zoom levels, and tissues. Our main focus in this paper is not solely on improving the quality of the generated images. Nevertheless, we present these quantitative metrics to demonstrate that our generated images exhibit similar quality to those of other previously published studies that have utilized diffusion models.

The calculated metrics provide an assessment of the overall quality of the generated dataset. However, for a more detailed analysis, we also calculated the FID and IL-NIQE individually for each class. The results are presented in Table 4.2. While an overall increase in FID was expected when calculating it separately for each class, we observed variations in this increase across different tissues. Specifically, the brain and pancreas tissues demonstrated a more substantial increase in FID compared to the uterus and lung tissues, which maintained their FID values close to the overall FID score. Interestingly, the class-wise IL-NIQE scores do not consistently corre-

late with class-wise FID scores. Specifically, the discrepancy between the average IL-NIQE scores of real and synthetic data suggests that the generated data for lung, pancreas, and uterus exhibit lower quality compared to those for brain and lung. These results highlight the importance of examining the performance of the generative model at a class level using multiple methods, as it provides valuable insights into the variability of generated data quality across different tissue types.

Table 4.2: Class-wise FID and IL-NIQE. To better understand the quality of the generated data for each tissue, we calculated the Fréchet Inception Distance (FID) and IL-NIQE scores for synthetic (synth.) and real data separately for each tissue type, as described in the Materials and Methods. The FID scores indicate the dissimilarity between the generated and real images, while the IL-NIQE score serves as a measure of image quality, capturing both structural and perceptual aspects.

Metric	Brain	Kidney	Lung	Pancreas	Uterus
FID	58.92	45.89	39.19	64.98	33.97
IL-NIQE (synth. data)	56.03	57.96	61.36	59.97	63.00
IL-NIQE (real data)	32.73	33.04	34.46	30.84	35.23

4.5.2 The Generated Data Exhibit Performance Comparable to Real Data in A Deep Learning-based Task

The performance of the trained ResNet-50 for tissue classification under various scenarios is summarized in Table 4.3. The table provides an overview of the test results obtained from different experimental settings. Furthermore, to gain a deeper understanding of the classification outcomes, the corresponding confusion matrices are visualized in Supplementary Figure S3.

Table 4.3: Performance of the trained neural networks for the tissue classification task. The accuracy (ACC) and Matthews correlation coefficient (MCC) values are reported for various training and testing scenarios as described in the Materials and Methods section.

Train set – Test set	ACC	MCC
Real data – External real data	98%	0.9705
Real data – Generated data	98%	0.9796
Generated data – Internal real data	95%	0.9331
Generated data – External real data	93%	0.9133
Reduced real data – External real data	97%	0.9606
Reduced real data – Generated data	98%	0.9765

The results indicate that there is no significant difference in the performance of the network trained on real data when tested on either unseen real data or generated data. This finding suggests that the generated images have successfully captured the diverse morphological patterns present in the original tissues, as the network’s

ability to classify them remains consistent. In contrast, when evaluating the network trained on generated data, a decrease in performance is observed for both seen and unseen real data. It is noteworthy that earlier studies have also documented a reduction in performance when relying solely on generated data as opposed to real data [156]. This reduction in performance can be attributed to several factors. Firstly, the smaller number of training samples available for the network trained on generated data compared to real data may contribute to lower accuracy. In fact, when we conducted an experiment by retraining the network using an equal number of real image tiles as generated ones, we observed a small decrease in accuracy when tested on external real data as reported in Table 4.3. Additionally, as evident from the confusion matrices, the decline in performance is notably pronounced for certain tissues such as pancreas and lung. This observation suggests that these tissues present more intricate and complex characteristics, potentially posing challenges for the diffusion model in generating diverse and accurate tiles. Interestingly, this phenomenon contrasts with some of our quantitative metrics, such as the class-wise FID, which might have predicted greater difficulties in tissues like the brain. However, the outcome of the deep learning assessment did not align with this expectation. It appears to correlate more closely with the results of the IL-NIQE, as that measure predicted lower quality of lung, pancreas, and uterus images, underscoring the significance of multifaceted evaluations.

Despite these drawbacks, the overall results are promising, as they demonstrate that the generated images exhibit similar patterns to those found in real tiles. Furthermore, by utilizing a small number of generated images, we can achieve performance levels comparable to those attained using real data.

4.5.3 Explainability-mediated analysis further reveals insights about the features learned from both datasets

We employed CRP to identify the top channel or concept activated for each class in both ResNets, one trained on real data and the other on generated data, analyzing 5 concepts for each network in total. These lists are presented in Supplementary Table S5. Once we determined the class-specific concepts, we used them to visualize which input pixels activate these concepts.

For instance, in Figure 4.3A, channel 428, associated with kidney-related concept in the ResNet trained on real data, is activated by kidney samples from both generated and real data. This is somewhat expected from the real data, given how we extracted this channel. However, it is not taken for granted for the generated data. Notably, the input pixels activating channel 428 are part of the same anatomical structure, the kidney’s glomerulus. In simpler terms, channel 428 in the ResNet trained on real data is activated by the glomeruli in kidney samples both in real and generated data.

What’s crucial is that when we examine the concept that is most activated by kidney samples in the network trained on generated data, shown in Figure 4.3B, a similar pattern emerges. In both real and generated data, the input pixels activating this channel correspond to the glomeruli of kidney images.

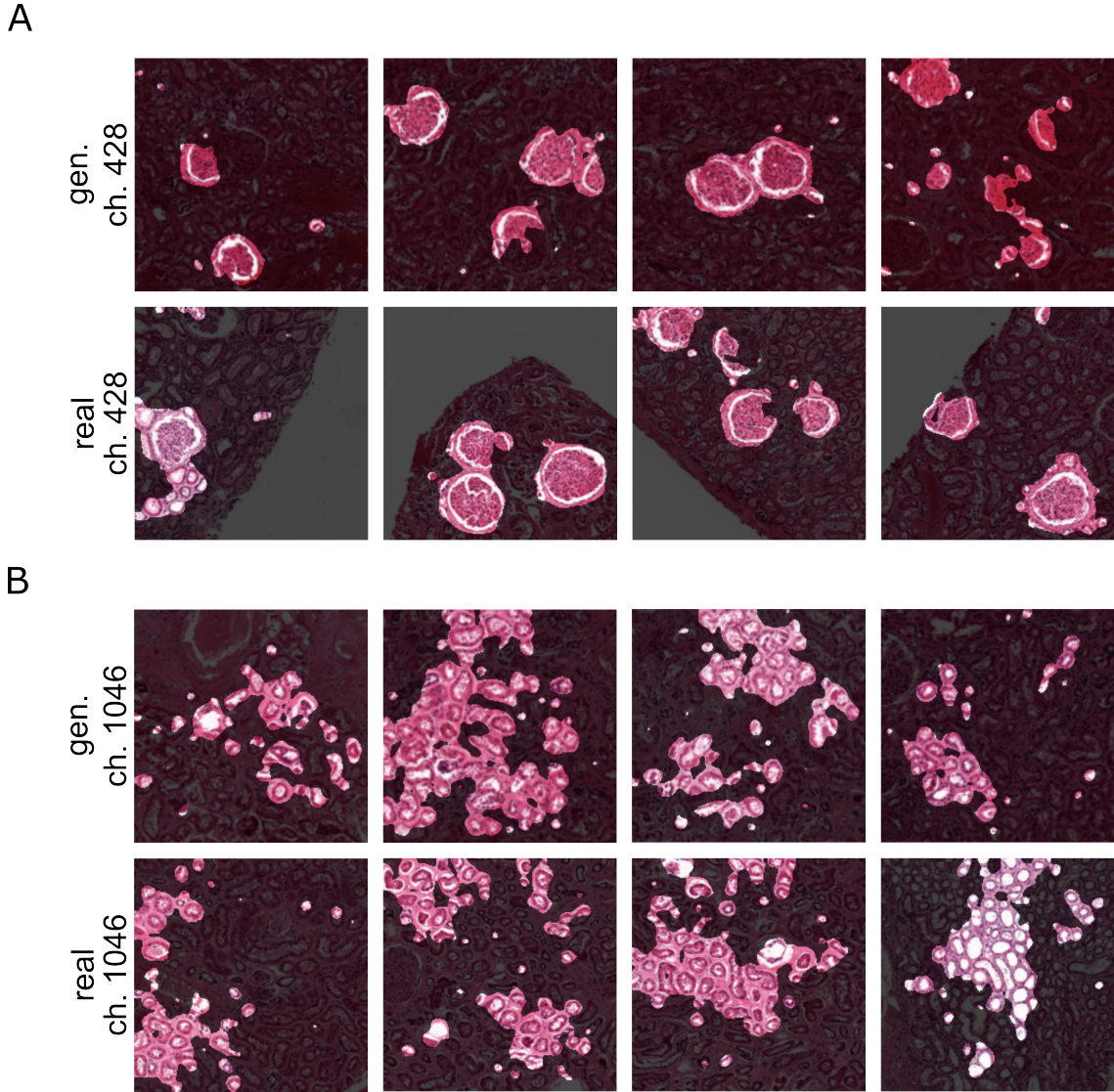


Figure 4.3: Concept relevance propagation visualization. Visualization of input pixels that activate the indicated channels. Each panel explores one network, either trained on real or synthetic data, and one concept. Each row examines the relevance of the channel either on the synthetic dataset (gen.) or on the real one (real). **(A)** ResNet trained on the real dataset. Input pixels that activate channel 428 in the last convolutional layer belong to the same tissue class and consistent morphological elements in both datasets. **(B)** ResNet trained on the synthetic dataset. Here, too, the tiles that activate the channel belong to the same class in both datasets, and input pixels belong to consistent morphological structures.

We also used the global approach in CRP to identify which classes are most activated by the same set of selected channels. The upper part of Supplementary Table S5 contains channels from the network trained on real images, while the lower part contains channels from the network trained on generated tiles. Upon inspecting the classes activated by each channel in both datasets, we found a strong agreement, suggesting that the networks behave similarly when it comes to tiles from both datasets. This evidence supports the performance metrics used to analyze the classifier’s performance.

Additionally, these results shed light on situations where the network makes

mistakes, particularly in cases of confusion between pancreas and lung. Channel 890 from the lower part of Supplementary Table S5 illustrates this, as although the most frequent class activated by that channel is Pancreas, the second most frequent class, with a substantial normalized relevance of 0.95 and 0.94 depending on the dataset, is the Lung class.

To gain a better understanding, we selected the most relevant concept for each tissue and visualized the pixels activating that concept in some example real and generated images using both the networks trained on the real dataset (Supplementary Figure S4) and the network trained on the generated dataset (Supplementary Figure S5). While it appears that the most important concept for tissues such as brain, kidney, lung, and uterus, to some extent, remains consistent between the networks trained on real and generated data, this isn't the case for pancreas. For the network trained on generated data, the most important concept for pancreas tissue, while descriptive enough (as indicated by the classifier's overall good performance), seems to resemble the most important concept for lung and uterus tissue, which explains the decrease in performance.

4.5.4 The Generated Data is Biologically Realistic in the Eyes of Experts

The pathologists' questionnaire results are summarized in Table 4.4 and depicted in Figure 4.4.

Table 4.4: First questionnaire results. The accuracy (ACC) and Matthews' correlation coefficient (MCC) values are reported for the tissue detection task using real and generated data, as well as for the real vs. generated discrimination task.

Task	Pathologist 1		Pathologist 2	
	ACC	MCC	ACC	MCC
Tissue detection with real images	100%	1	98%	0.9757
Tissue detection with generated images	100%	1	100%	1
Real or generated	47%	-0.061	61%	0.2278
Task	Pathologist 3		Combined	
	ACC	MCC	ACC	MCC
Tissue detection with real images	100%	1	98.33%	0.9917
Tissue detection with generated images	98%	0.9757	98.33%	0.9917
Real or generated	60%	0.2881	56%	0.1245

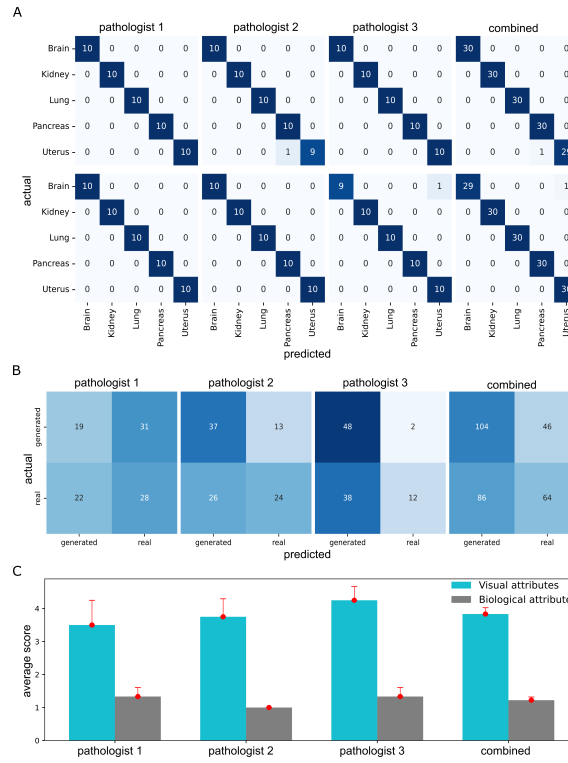


Figure 4.4: Pathologists' questionnaire results. **(A)** Presents the confusion matrices illustrating the pathologists' performance in tissue detection task in the first questionnaire, using both real (top row) and generated (bottom row) images. These matrices correspond to the first two rows of Table 4.3. **(B)** Displays the confusion matrices that depict the pathologists' performance in the task of discriminating between real and generated images during the first questionnaire. These matrices correspond to the last row of the results presented in Table 4.3. **(C)** Illustrates the average $\pm$ standard error of the mean of the pathologists' responses to the second questionnaire. The cyan bar represents the average score indicating the score to which pathologists relied on visual attributes to differentiate real images from generated ones, while the gray bar indicates the extent to which they utilized biological attributes as described in Materials and Methods.

The results from the first task of tissue detection demonstrate that pathologists are highly proficient in identifying the tissue type based on the given three tiles at the provided zoom level. Importantly, there is no significant difference in performance between real and generated images, indicating that the generated data accurately capture the morphology of the original tissues. The combined performance of the pathologists slightly exceeds chance level in detecting real vs. generated tiles (accuracy of 56%), suggesting that the generated tiles possess a biologically realistic appearance. When we examined the distribution of errors made for the generated tiles (generated tiles mistaken for real), as shown in Supplementary Figure S6, the brain images had slightly higher average number of mistakes, suggesting that they appeared more realistic to the pathologists' eyes and conversely, kidney images had the lowest average number of mistakes, indicating that pathologists found it easier to detect generated kidney images. Nevertheless, the overall differences across tissues were small and not significant. Notably, the results from the second questionnaire reveal that the pathologists primarily relied on visual attributes such as chromatic

features and blurriness to determine whether a tile was real or generated, as they did not detect any noticeable biological irregularities in the generated images. When combined, the average score given to visual attributes was 3.83 ± 0.19 , which was found to be statistically significantly different from the average score of 1.22 ± 0.01 given to biological attributes (two-tailed paired t-test, p-value = 0.0075). This finding highlights the potential value of the generated dataset for educational purposes, as it indicates that the generated images are biologically realistic to the extent that even expert pathologists are almost unable to distinguish them, except for some minor differences at the image visual quality level.

4.6 Discussion

The field of digital pathology, like many other disciplines, could greatly benefit from the ability to generate realistic synthetic data. These data could assist in addressing privacy concerns and enabling AI-assisted diagnosis by providing high-quality data to aid the training process. This holds promise in scenarios requiring pixel/tile-level annotations which can be time-consuming and susceptible to errors when executed by humans. Hence, in this article, we have introduced a comprehensive pipeline for generating and assessing synthetic data in digital pathology. We employed this pipeline to generate images from five distinct tissues with a wide field of view, employing a denoising diffusion model. The quality of the data was assessed through three distinct evaluation approaches: quantitative assessment employing well-established metrics, practical applicability evaluation through deep learning-based tissue classification and explainability, and qualitative evaluation involving questionnaires administered to pathologists. Our findings emphasize the fact that employing these varied evaluations provides us with complementary insights into data quality, highlighting that relying solely on a single approach would be insufficient.

Given the current trend towards analyzing digital pathology images at the whole slide level [165, 187, 188], it is desirable to generate synthetic data at the same level as well. In our study, we focused on generating tiles of wide field of view, to examine the ability of diffusion models in generating images at this resolution and as an early attempt to get closer to WSI image generation. Future research can build upon this by generating tiles from multiple fields of view and subsequently integrating them to generate complete whole slide images. Additionally, incorporating the adjacency of tiles in the generative model could assist in generating neighboring tiles to be able to generate WSIs. An example of this has been done with GANs in [189].

We explored the diffusion model’s capability to generate tiles from various tissues, each characterized by distinct morphological complexities. Our findings indicated that, although the overall performance was satisfactory, the model’s effectiveness varied across different tissues. An intriguing observation highlighting the importance of diverse evaluations emerged when examining each tissue individually.

In particular, when comparing class-wise FID scores, brain and pancreas showed similar results, both significantly higher than lung. However, classifier performance was notably lower for pancreas and lung, while the brain exhibited almost perfect performance. This trend was more closely mirrored in the results from the class-wise IL-NIQE, where pancreas, lung, and uterus showed lower quality. Furthermore,

the results from the pathologists’ questionnaire indicated that pathologists had the most difficulty in distinguishing between real and generated brain images, implying that the generated brain images are of the highest quality. Conversely, they had the easiest time distinguishing kidney generated images, which had relatively good FID and IL-NIQE scores.

Explainability analyses shed further light on these findings. They revealed that the most relevant histopathological features for detecting pancreas when the network is trained on generated data differ from that when trained on real data. Interestingly, these features closely resemble the most relevant concepts for lung tissue. It remains to be determined whether these differences are attributed to the limitations of the diffusion model’s classifier-free guidance pipeline in distinguishing between lung and pancreas tissues or whether they stem from inherent complexities within the pancreas tissue itself.

One limitation of utilizing diffusion models is their longer inference time, i.e. will take longer for the model to generate a synthetic image, compared to alternative generative methods like GANs. However, our primary focus in this paper is to underscore the importance of a multifaceted evaluation pipeline and the valuable insights it offers. This evaluation framework is applicable to various generative models, not limited to diffusion models. We opted for diffusion models due to their consistent ability to generate realistic medical and natural images, a critical consideration given the sensitive nature of medical data. Nonetheless, further studies can be conducted, assessing multiple generative models using the proposed evaluation pipeline, enabling a comprehensive comparison of their strengths and weaknesses.

Despite tissue classification not ranking among the foremost priorities in digital pathology, we harnessed its utility to assess the quality of generated data, their viability in replacing real tiles within deep learning-based tasks and to emphasize the significance of multifaceted evaluation. As anticipated, the performance of a model trained solely on generated data did not match that of a network trained on real data, yet the outcome remained promising and satisfactory. The generation of data to facilitate network training becomes particularly valuable when dealing with scenarios that require pixel-level annotations, which are labor-intensive. Recent studies have showcased generative models’ potential in producing annotated pairs of tissue images alongside their corresponding cell annotations’ mask [155, 190] for specific tissues. This lays the foundation for expanding our work, potentially involving the generation of data from various tissues, accompanied by their respective cell type annotations under normal and various pathological conditions unique to each tissue. The classifier-free guidance conditioning employed in the diffusion model, as demonstrated by [154], enables the encoding of multiple biological features, thereby paving the way to condition on more biologically significant factors.

Explainable artificial intelligence is a rapidly growing sub field within the machine learning community [191] . It has emerged as a response to the challenge of interpreting predictions made by black box models, where the underlying mathematical functions are often highly complex and difficult to understand, even for AI experts. Deep neural networks, although highly effective in their performance when dealing with unstructured data such as images, fall into this category of black box models. To leverage the power of deep learning methods while addressing the need for interpretability, post-hoc methods have been developed. In this study, we demonstrated the application of an explainability technique to enhance our under-

standing of data quality and to gauge whether, for the generated data, the model learns comparable features as when trained with real data. This holds a special significance, especially considering a recent study that highlighted how generated histopathological images can provide explanations beyond what is attainable with real data [192]. This is particularly important in the medical domain, where the demand for interpretable models remains crucial.

Finally, the pathologists’ relatively low performance in distinguishing between real and generated images reinforced our confidence in the realism of the generated images, thus promoting their utility for educational purposes. Notably, the feedback garnered from the subsequent questionnaire was intriguing. It emphasized that instances where pathologists successfully identified generated tiles were often attributable to discernible visual cues like blurriness or chromatic characteristics. Given the significantly high resolution of histopathology images, the fact that generated images would not consistently match this level of resolution was to be expected. This limitation could potentially be addressed in future studies by integrating super-resolution networks such as the one proposed in [193], to enhance the quality of generated images.

Overall, although synthetic data in digital pathology may not be ready to completely complement real images for research purposes, our findings, like other studies, highlight their promising potential in the future. What’s important to consider is the need to examine these generated data from different angles to fully understand their strengths and weaknesses. Our results emphasize the importance of thoroughly evaluating the generated dataset using various approaches, which should become a common practice in the field to better develop and use synthetic datasets effectively.

4.7 Data and Code Availability

All the data used in this paper is publicly accessible from GTEx Portal. The source code is publicly available on GitHub at: [GitHub Repository](#). ResNet and Diffusion model weights are publicly available on the HuggingFace model hub at: [HuggingFace Repository](#).

4.8 Additional Informations

4.8.1 Fundings

The study was partially funded under the National Plan for Complementary Investments to the NRRP, project “D34H—Digital Driven Diagnostics, prognostics and therapeutics for sustainable Health care” (project code: PNC0000001), Spoke 2: “Multilayer platform to support the generation of the Patients’ Digital Twin”, CUP: B53C22006170001, funded by the Italian Ministry of University and Research.

Chapter 5

Investigating TILs in ductal carcinoma in situ

5.1 Summary

DOI: PMLR 254:214-225

Introduction This paper explores the role of tumor-infiltrating lymphocytes (TILs) in ductal carcinoma in situ (DCIS), a non-invasive form of breast cancer. TILs are a critical component of the tumor microenvironment, influencing prognosis and therapeutic responses. The study investigates their detection, scoring, and correlation with pathologist’s computed TILs score.

Methods

1. Data Collection: Histopathological slides of DCIS cases are collected and digitized.
2. TIL Detection: Automated algorithms and pathologist evaluations are used to identify and quantify TILs in the tumor microenvironment.
3. Scoring Systems: Both manual and automated scoring approaches are developed to assess TIL density and distribution.
4. Statistical Analysis: Correlations between AI assisted and pathologist computed TILs score.

Results Automated and manual TIL detection methods achieve high concordance.

Discussion The study highlights the potential of incorporating TIL quantification into routine pathology workflows to refine risk stratification and guide therapeutic decisions. Future directions include integrating TIL scoring into AI-driven diagnostic pipelines and validating these approaches in larger, multicenter studies.

Scoring Tumor-Infiltrating Lymphocytes in breast DCIS: A guideline-driven artificial intelligence approach

Matteo Pozzi^{1,2}, Natalie Klubickova³, Michela Campora⁴, Frederique Meeuwssen³, Joey Spronck³, Carlijn Lems³, Michelle Stegeman³, Leslie Tessier³, Mattia Barbareschi⁴, Jeroen van der Laak³, Giuseppe Jurman¹, Francesco Ciompi³

¹ Fondazione Bruno Kessler, Data Science for Health Unit, Via Sommarive 18, 38123 Povo, Trento, Italy.

² Università degli Studi di Trento, Department for Computational and Integrative Biology, Via Sommarive 9, 38123 Povo, Trento, Italy.

³ Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands.

⁴ Pathology Unit, Department of Laboratory Medicine, Santa Chiara Hospital, APSS, Trento, Italy.

* Corresponding author: Giuseppe Jurman, Data Science for Health Unit, Fondazione Bruno Kessler, Via Sommarive 18, 38123 Povo, Trento, Italy, giuseppe.jurman@fbk.eu (+39) 0461-314523.

5.2 abstract

This study focuses on the assessment of Tumor-Infiltrating Lymphocytes (TILs) in Breast ductal carcinoma in situ (DCIS) by integrating artificial intelligence with international guidelines. DCIS is a non-invasive cancer with intrinsic potential to evolve to invasive breast cancer (IBC), making it critical to understand factors influencing this progression. TILs are a prognostic biomarker in IBC, but their role in DCIS remains under-explored. This work proposes an automated pipeline for computing TILs scores using deep learning for DCIS segmentation and TILs detection, following the guidelines of the International Immuno-Oncology Biomarker Working Group. We report the inter-observer variability at TILs scoring among Pathologists and show that the AI-based TILs scores have good concordance with human assessments. Future research will aim to reduce false positives in DCIS segmentation and detection, support the reference standard with immunohistochemical staining, and expand the dataset to enhance the robustness of the TILs detection algorithm. Ultimately, this method aims to aid Pathologists in assessing the risk associated with DCIS lesions

5.3 Introduction

Ductal Carcinoma In Situ (DCIS) is a non-invasive form of breast cancer that has garnered significant attention owing to its increasing incidence and potential to progress to invasive breast cancer (IBC). DCIS is characterized by the presence of neoplastic cells within the breast ducts, without invasion beyond the basement

membrane. DCIS is both morphologically and clinically heterogeneous precursor lesion which might be identified through mammographic breast cancer screening programs in some cases [194]. To improve patient outcomes and treatment strategies, it is crucial to understand the factors influencing the progression of DCIS to IBC. The neoplastic transition evolving to IBC involves a multistep process of progression through the stages of atypical hyperplasia, DCIS, and ultimately invasive carcinoma [195]. Studies have shown that a significant proportion of high grade DCIS cases may progress to invasive breast cancer within 30 years [196].

The management of DCIS poses challenges with questions surrounding the necessity of surgery for low-grade DCIS and the role of active surveillance [197]. The standard of care for DCIS typically involves either breast conservative surgery or mastectomy for patients treated with conservative surgery, ASCO guidelines suggest a clearance of margins greater than 2 mm [198] and possible omission of radiotherapy for selected low-risk women [199]. Recent study compared genomics profile from patients with an initial DCIS lesion and a later invasive recurrence, showing that in 18% of the cases the invasive recurrence was not genetically related to the DCIS, while in 75% of the cases a clonal relation between the two lesions was found, suggesting that tumor cells were not eliminated during the initial treatment of DCIS [200]. In order to avoid overtreatment and to move toward personalized medicine, it is important to develop new biomarkers that help assessing the exact invasive potential of the DCIS lesion.

Tumor-Infiltrating Lymphocytes (TILs) are a well known prognostic biomarker in IBC, with high TILs density associated with better response to adjuvant or neoadjuvant therapy and with a positive outcome [201, 202, 203]. In the context of DCIS, the assessment of TILs is an emerging area of research, while the role of different TILs in DCIS is still not well defined [203]. Several studies have investigated the quantification of TILs in DCIS and their correlation with clinical outcomes [204], their potential prognostic significant related to the underlying genomics instability [205], and their association with a second breast cancer event [206].

Automated analysis of the tumor micro environment and assessment of TILs through deep learning techniques is increasingly gaining interest in the immuno oncology community. In the field on DCIS researchers have focused on analyzing the spatial distribution of TILs between patients with pure DCIS and the ones with IBC adjacent to DCIS [207]. Generative Adversarial Networks (GAN) have also been used to segment DCIS and the resulting segmentation mask used to compute a TILs score [208].

In this study we have focused on identifying DCIS as well as other relevant morphological classes for an objective and reproducible TILs score computation according to the recommendations of the International Immuno-Oncology Biomarker Working Group on Breast Cancer. We collected and annotated hematoxylin and eosin (H&E) stained digital pathology Whole Slide Images (WSIs) from Santa Chiara Hospital and used the publicly available BRACS dataset as an external validation set [209]. Aim of the study is to provide Pathologist a tool to enable quantitative assesment of TILs abundance in DCIS.

5.4 Materials and Methods

5.4.1 Data collection

A total of 100 WSIs were used for this study. Data was obtained from two independent sources: Santa Chiara Trento Hospital, for a total of 55 WSIs used for training, and BRACS dataset, for the remaining 45 used for testing.

In our pipeline we did not use QuPath [210] annotations of DCIS areas available within the BRACS dataset. Pathologist annotated selected area of interests, using the same classes of the TIGER challenge [211], but including only dcis in the in-situ tumor. The class included in the annotations where the following: tumor_associated_stroma, inflammed_stroma, healthy_glands, invasive_tumor, necrosis, dcis, rest.

Annotation on both dataset were made by trained Pathologist which sparsely annotated the slides using QuPath. All slides from both datasets were converted into tif format and saved at a resolution of 0.5 micron per pixel.

5.4.2 DCIS segmentation

In our study we used a version of nnUNet [212] adapted for digital pathology applications [213] for multi-class tissue segmentation in whole-slide images. Training parameters encompass a square patch size of 512 pixels, a mini-batch size of 18 samples, and input resolution of 0.5 micro per pixels. The training dataset was partitioned into a 5 fold cross validation. Training have been performed for 1000 epochs and model checkpoints have been saved based on the overall DICE score performance on the internal validation set. Regions in the WSI which were not annotated, have been excluded from loss calculation. Classes other than DCIS and rest where not used for the following part of the analysis and where therefore ignored. DICE score limited to the DCIS class have been computed on the external BRACS dataset. We limited our evaluation on this class since it was the relevant one for the TILs score computation. During training we adopted a weighted sampling strategy to increase the frequency of sampling of DCIS regions (see weights in Table 5.1). An example of a segmentation mask is reported in Figure 5.1

Label	Weight
rest	0.125
tumor_associated_stroma	0.125
inflammed_stroma	0.125
healthy_glands	0.125
dcis	0.250
invasive_tumor	0.125
necrosis	0.125

Table 5.1: Annotation weights used during nnUNet training

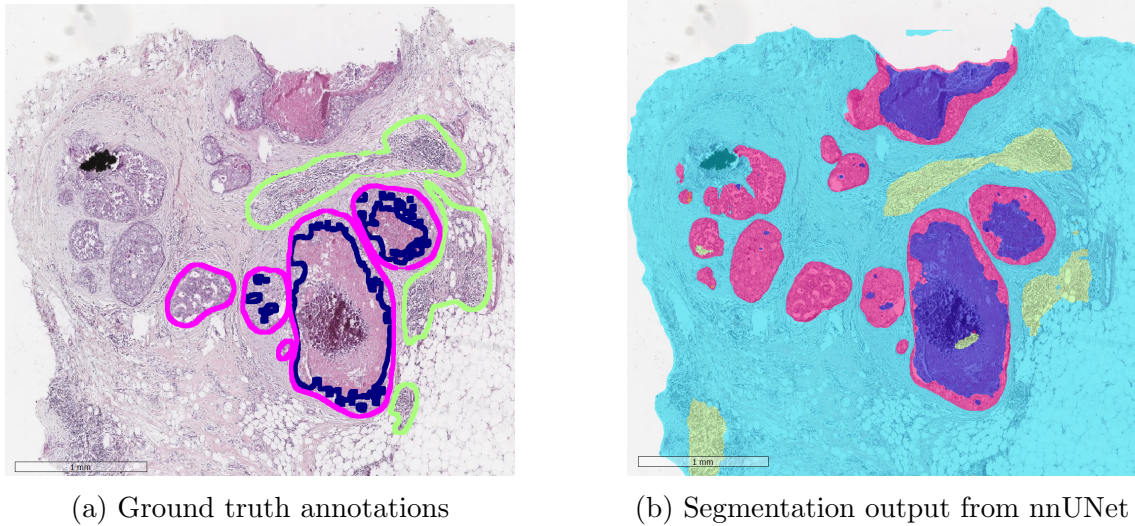


Figure 5.1: Example of the nnUNet segmentation output. DCIS is colored in pink, necrosis in blue, inflamed stroma in yellow, while stroma in light blue. Images are visualized using the ASAP software.

5.4.3 TILs detection

TILs detection (see Figure 5.2) was done using Biototem algorithm and pipeline¹, the best performing algorithm in the computer vision task of the Tumor Infiltrating lymphocytes in breast cancer (TIGER) challenge [211], hosted on the Grand Challenge Platform.

The model is based on UperNet with a Visual Attention Network (VAN) backbone, which enables long-range correlations in self-attention. The algorithm adapt the SFCN-OPI model, originally designed for nuclei detection, to improve TILs localization and reducing false positives.

We leveraged this model to identify TILs in BRACS slides. Detection results were then converted into xml format.

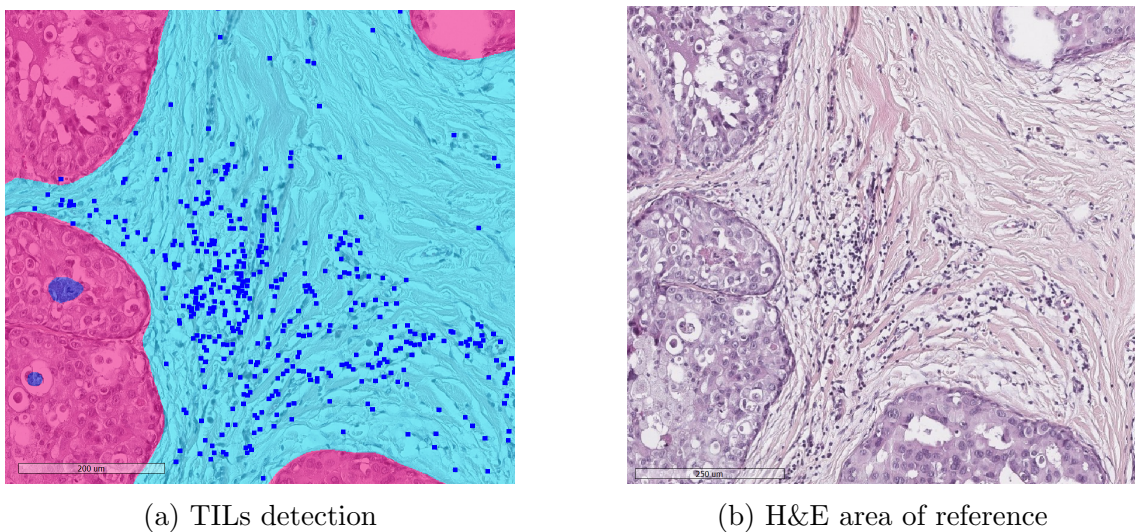


Figure 5.2: Biototem's TILs detection output

¹https://github.com/biototem/TIGER_challenge_2022

5.4.4 Pathologist TILs score

Two Pathologist provided TILs score for each WSI of the BRACS dataset following the International Immuno-Oncology Biomarker Working Group on Breast Cancer guidelines. Pathologist evaluated TILs score by assessing the stromal area around DCIS and reported a score per slide as a percentage of the stromal area infiltrated by lymphocytes. Pathologist scored the TILs ratio with 5% steps.

5.4.5 Automatic TILs score computation

DCIS segmentation mask derived from nnUNet model where converted into xml annotation files and merged with the TILs detection from Biototem algorithm. A buffer area of 250 micrometer where computed around each DCIS polygon annotation, merging buffers deriving from adjacent DCIS. The dimension of the buffer area is within the range suggested by the International guidelines, which suggest to inspect an area *up to* two high-power microscopic fields.

The buffer of a geometry is defined as the Minkowski sum of the geometry with a circle with radius equal to the absolute value of the buffer distance. A representative example is shown in Figure 5.3. We then counted the number of TILs detection within this buffer. We approximated the TILs area as being the one of a circle having a diameter of 8 micrometers [214]. Tils score for each WSI is therefore computed as follows:

$$\text{TILs_score} = \frac{\sum_{i=1}^n \text{TIL}_i \in (\text{B area} - \text{DCIS area}) \times 50 \mu\text{m}}{\text{B area} - \text{DCIS area}} \quad (5.1)$$

where B is the buffer derived from geometric operation. According to TILs scoring guidelines, immune hostspots identified by the class *inflammed_stroma* (i.e., regions with high-density clusters of lymphocytes within the stromal region) were included in the computation both by pathologist and by our computational approach.

5.4.6 Statistical Analysis

Performance of the DCIS segmentation model was evaluated using the DICE score. The correlation between the automatic TILs scores and the pathologists' scores was analyzed using Pearson and Spearman correlation coefficients to assess agreement.



Figure 5.3: The buffer area used to computed tils detection. The total buffer area is represented in red, while the pink area represent the DCIS identified by nnUNet

5.5 Results and Discussion

The DCIS DICE score achieved by the model was of 0.69 with a variance of 0.08.

The computed TILs score was compared with the one estimated by Pathologist, resulting in a Pearson correlation coefficient of 0.53 with Pathologist 1 and 0.65 with Pathologist 2. Interestingly the correlation between Pathologist is limited to 0.48, showing a stronger concordance with the computed TILs score.

Spearman correlation coefficient shows a correlation of 0.57 between Pathologist 1 and 2. The computed TILs score shows a Spearman correlation of 0.61 with Pathologist 1 and 0.74 with Pathologist 2.

Interestingly, this result is in contrast with what was highlighted in the study by Hagos et al [208] in which the concordance between Pathologist was higher than the concordance between Pathologist and AI-based computed score.

We developed a pipeline to automatically compute TILs scores following the guidelines for TILs assessment from the International Immuno-Oncology Biomarker Working Group. This study shows the discordance between Pathologists in computing density ratios with little to no quantitative guidance. On the contrary, computed

TILs score is shown to correlate better with both Pathologist.

It has to be noted that the computed TILs score tend to be always low, potentially showing a tendency in underestimating the TILs score Figure 5.4. However this dataset did not contain any cases with high TILs score, therefore this hypothesis should be tested in future work, by integrating a wider and more diverse dataset.

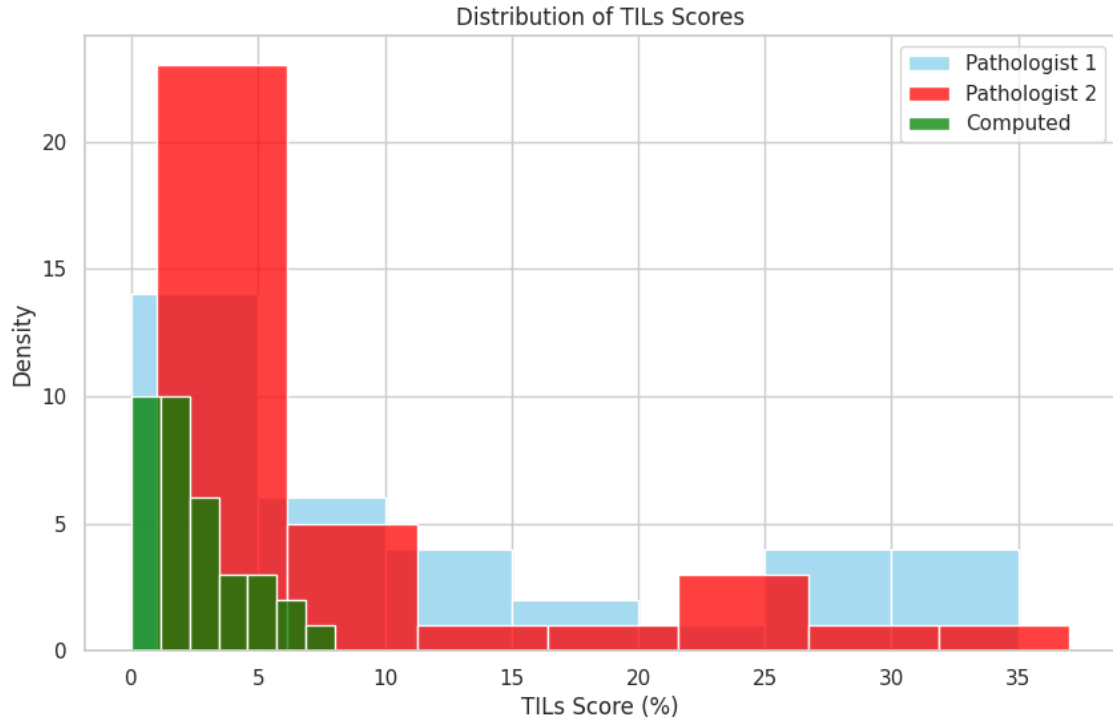


Figure 5.4: Distribution of TILs scores within Pathologist and the computed ones for the BRACS dataset

Future work will focus on reducing the number of false positives prediction in DCIS class as well as integrating immunohistochemistry staining to validate the predictions made by the TILs detection algorithm. Another point of action will be also increasing the size of the training dataset, with the goal to improve the segmentation performance of nnUNet. Additionally, in order to further investigate the robustness of these results, we plan to expand the number of pathologist who will assign TILs score for the BRACS validation dataset.

In conclusion we developed a pipeline to compute the TILs score in ductal carcinoma in situ using state of the art deep learning method for DCIS segmentation and TILs detection. Those output have been combined and exploited to compute TILs score following international guidelines. All results have been computed on an external and independent dataset, which has never been used in any part of the training protocol, to showcase the robustness of our method, with the aim to develop a tool able to assist Pathologist in DCIS risk assessment. Upon acceptance we aim to make our BRACS annotation publicly available, to faster the reproducibility of the study and to enlarge the number of high quality annotations, useful to develop large scale Deep Learning applications.

5.6 Acknowledgments

This work have been supported by the National Plan for Complementary Investments to the NRRP, project “D34H—Digital Driven Diagnostics, prognostics and therapeutics for sustainable Health care” (project code: PNC0000001), Spoke 2: “Multilayer platform to support the generation of the Patients’ Digital Twin”, CUP: B53C22006170001, funded by the Italian Ministry of University and Research.

Chapter 6

Discussion

The advent of artificial intelligence (AI) in medicine has ushered in transformative changes across various domains of healthcare. AI's capacity to analyze complex datasets and uncover patterns beyond human capabilities holds immense promise for diagnostics, treatment planning, and personalized medicine. This thesis has examined the integration of AI into medical practice, with a particular focus on computational pathology, aiming to enhance diagnostic accuracy and efficiency by leveraging existing computational techniques rather than developing entirely new algorithms. This approach not only streamlines the deployment of AI in clinical settings but also promotes reproducibility and generalizability, core attributes necessary to ensure reliable replication of results and application across diverse patient populations.

A foundational aspect of this research was the review of automation in pathology. Automation has emerged as a critical enabler for improving the robustness and consistency of pathology practices. By standardizing routine tasks and minimizing human error, automation lays the groundwork for integrating AI applications into clinical workflows. The review highlighted the ability of automated systems to process large volumes of data with high precision, thereby enabling pathologists to focus on more complex diagnostic tasks. This efficiency also mitigates biases such as data leakage and dataset shift, which can undermine the reliability of AI models if not properly addressed. By ensuring clean, well-curated input data, the foundation is set for models to deliver more reproducible and stable performance across various datasets.

Building on this, the study on bowel disease exemplified the application of AI in a clinical context. Explainable AI (XAI) techniques were employed to interpret bowel disease conditions, addressing the critical need for transparency in AI-assisted diagnostics. By providing insights into the decision-making processes of AI models, XAI fosters trust and accountability among clinicians. This study shows how XAI can be employed to bridge the gap between complex machine learning models and clinical acceptance, ultimately enhancing patient outcomes through more accurate and interpretable diagnostics.

Addressing the challenge of data scarcity, this thesis also explored the use of synthetic data in computational pathology. Data scarcity is a significant barrier to training effective AI models, particularly in medicine, where obtaining large, annotated datasets is often difficult due to privacy concerns or the rarity of certain conditions. In this study, synthetic histopathological images were generated to

augment existing datasets. XAI methods were used to evaluate the usability of this synthetic data, ensuring the generated images were representative and effective for training AI models. The findings emphasized that properly validated synthetic data can mitigate limitations related to data privacy and complement AI models in detecting and classifying pathological features. By integrating XAI with traditional quantitative metrics, a robust framework was established to assess and validate the quality of synthetic datasets.

The research on tumor-infiltrating lymphocytes (TILs) further demonstrated the practical applications of AI in computational pathology. Rather than developing a new segmentation model, existing semantic segmentation and object detection models were utilized to calculate TIL scores in accordance with international guidelines. Explainability remained central, ensuring that segmentation masks used for TIL quantification were accurate and interpretable. By aligning AI-generated outputs with standardized scoring systems, this study enhanced the reliability and clinical relevance of the findings, providing valuable insights into the tumor microenvironment, which is crucial for prognosis and tailoring immunotherapies.

Collectively, these studies contribute to a cohesive narrative aimed at advancing computational pathology through the integration of AI and XAI techniques. Each investigation addressed critical challenges—automation, explainability, data scarcity, standardization—that are essential for successfully deploying AI in medical practice. By systematically tackling these issues, the thesis worked toward the overarching goal of enhancing diagnostic processes and improving patient care.

6.1 Conclusion and Future Directions

In conclusion, this thesis has demonstrated the significant impact AI can have on medicine, particularly when integrated with explainable AI. The review on automation underscored the necessity of robust and efficient systems as a foundation for AI integration. The bowel disease study illustrated how XAI can make AI models more transparent and acceptable to clinicians. The exploration of synthetic data provided solutions to data scarcity, addressing a major hurdle in AI development. Lastly, the research on TILs highlighted the importance of aligning AI outputs with clinical guidelines to ensure practical applicability.

Looking ahead, two primary directions emerge for the field of AI in medicine. On one hand, cutting-edge research should focus on exploring advanced methodologies and foundational models that deepen clinical knowledge and push the boundaries of AI's potential in medical diagnostics. On the other hand, translational research must prioritize the development of clinically impactful tools that, while potentially less complex, can be rapidly integrated into routine practice to provide immediate support for clinicians at the point of care. Both trajectories require interdisciplinary collaboration, ensuring that tools are not only technologically robust but also clinically relevant.

To further enhance the impact of this work, future research should prioritize a broader examination of AI's implications for global healthcare systems, particularly its potential to address disparities in resource-limited settings. Additionally, there is a need to establish standardized protocols for validating AI applications, fostering wider clinical acceptance and integration. Expanding upon the principles

of explainability, reproducibility, and generalizability outlined in this thesis, future efforts can build a foundation for AI-driven tools that are effective, trustworthy, and seamlessly integrated into everyday medical practice.

One promising avenue emerging from the synthetic data paper involves generating entire Whole Slide Images (WSI) rather than just image tiles, as well as focusing on generating data for more rare and clinically interesting conditions, thereby extending the applicability of synthetic data beyond specific tissue types. Generating whole WSIs could significantly enhance the training and evaluation of AI models by providing more holistic and comprehensive datasets that closely mimic real-world clinical scenarios. Furthermore, synthetic data could be designed to simulate unique pathological features across a broader spectrum of disease states, enabling more robust model performance in diverse clinical environments. Additionally, synthetic data could be utilized to train new pathologists by progressively including complexity in the generated images. This approach would allow trainees to build foundational skills with simpler datasets and gradually advance to interpreting more intricate and challenging pathological patterns, thereby improving both educational outcomes and the practical readiness of emerging professionals in the field.

Moreover, a related future direction tied to the TILs score paper involves further characterizing the TIL population by exploiting immunohistochemical (IHC) markers to analyze and explore the differences in specific lymphocyte subtypes within the context of ductal carcinoma in situ (DCIS). This expanded characterization could uncover critical variations in the functional roles of lymphocyte subpopulations, providing insights into their interactions with the tumor microenvironment. Leveraging IHC markers would allow researchers to integrate more precise biological data into AI models, moving beyond merely quantifying TILs to exploring their heterogeneity and functional implications.

By combining these two approaches—expanding synthetic data generation to include full WSIs and advancing the characterization of TILs through IHC markers—future research can bridge the gap between computational techniques and biological insights. These efforts would enable AI systems to address more complex questions, such as the interplay between immune responses and disease progression in DCIS, while also creating datasets that push the boundaries of AI’s capabilities in pathology.

In sum, these directions underscore the transformative potential of AI in medicine, where advancing both computational techniques and biological understanding can lead to tools that are innovative, clinically impactful, and equitable.

Bibliography

- [1] F. Micoli, F. Bagnoli, R. Rappuoli, and D. Serruto. The role of vaccines in combatting antimicrobial resistance. *Nature Reviews. Microbiology*, 19:287 – 302, 2021.
- [2] M. Cook and Gerard D. Wright. The past, present, and future of antibiotics. *Science Translational Medicine*, 14, 2022.
- [3] A. Balmain, J. Gray, and B. Ponder. The genetics and genomics of cancer. *Nature Genetics*, 33 Suppl 1:238–244, 2003.
- [4] E. Lander. Initial impact of the sequencing of the human genome. *Nature*, 470:187–197, 2011.
- [5] S. Kraus, Francesco Schiavone, Anna Pluzhnikova, and Anna C. Invernizzi. Digital transformation in healthcare: Analyzing the current state-of-research. *Journal of Business Research*, 123:557–567, 2021.
- [6] P. Kostkova. Grand challenges in digital health. *Frontiers in Public Health*, 3, 2015.
- [7] Qiuhan Zheng, Le Yang, B. Zeng, Jiahao Li, Kaixin Guo, Yujie Liang, and G. Liao. Artificial intelligence performance in detecting tumor metastasis from medical radiology imaging: A systematic review and meta-analysis. *EClinicalMedicine*, 31, 2020.
- [8] Rachel Y L Kuo, C. Harrison, T. Curran, Benjamin Jones, Alexander Freethy, David Cussons, Max Stewart, G. Collins, and D. Furniss. Artificial intelligence in fracture detection: A systematic review and meta-analysis. *Radiology*, page 211785, 2022.
- [9] T. Sakellaropoulos, K. Vougas, Sonali Narang, F. Koinis, A. Kotsinas, A. Polyzos, T. Moss, S. Piha-Paul, Hua Zhou, Eleni Kardala, Eleni Dami-anidou, L. Alexopoulos, I. Aifantis, P. Townsend, M. Panayiotidis, P. Sfikakis, J. Bartek, R. Fitzgerald, D. Thanos, Kenna R. Mills Shaw, R. Petty, A. Tsirigos, and V. Gorgoulis. A deep learning framework for predicting response to therapy in cancer. *Cell reports*, 29 11:3367–3373.e4, 2019.
- [10] Yiwen Xu, A. Hosny, R. Zeleznik, C. Parmar, T. Coroller, I. Franco, R. Mak, and H. Aerts. Deep learning predicts lung cancer treatment response from serial medical imaging. *Clinical Cancer Research*, 25:3266 – 3275, 2019.

- [11] Taro Sakamoto, Tomoi Furukawa, Kris Lami, Hoa Hoang Ngoc Pham, Wataru Uegami, Kishio Kuroda, Masataka Kawai, Hidenori Sakanashi, Lee Alex Donald Cooper, Andrey Bychkov, and Junya Fukuoka. A narrative review of digital pathology and artificial intelligence: focusing on lung cancer. *Translational Lung Cancer Research*, 9(5):2255–2276, 2020.
- [12] Jeroen van der Laak, Geert Litjens, and Francesco Ciompi. Deep learning in histopathology: the path to the clinic. *Nature Medicine*, 27(5):775–784, May 2021.
- [13] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.
- [14] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):e0130140, 2015.
- [15] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization, 2015.
- [16] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 2019.
- [17] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018.
- [18] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. *iŕThisj/iŕ Looks like iŕThatj/iŕ: Deep Learning for Interpretable Image Recognition*. Curran Associates Inc., Red Hook, NY, USA, 2019.
- [19] Syed Ashar Javed, Dinkar Juyal, Harshith Padigela, Amaro Taylor-Weiner, Limin Yu, and Aaditya Prakash. Additive mil: Intrinsic interpretability for pathology, 2022.
- [20] Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. From "where" to "what": Towards human-understandable explanations through concept relevance propagation, 2022.
- [21] B. Markl, L. Fuzesi, R. Huss, Svenja Bauer, and T. Schaller. Number of pathologists in germany: comparison with european countries, usa, and canada. *Virchows Archiv*, 478:335 – 341, 2020.
- [22] L. Velez-Lapao, M. Maia, and G. Dussault. ehealth solutions to regional shortages in the health workforce. *European Journal of Public Health*, 2019.

- [23] R. Colling, Helen Pitman, K. Oien, N. Rajpoot, Philip Macklin, D. Snead, Tony Sackville, and C. Verrill. Artificial intelligence in digital pathology: a roadmap to routine use in clinical practice. *The Journal of Pathology*, 249, 2019.
- [24] Koen Salemink, D. Strijker, and G. Bosworth. Rural development in the digital age: A systematic literature review on unequal ict availability, adoption, and use in rural areas. *Journal of Rural Studies*, 54:360–371, 2017.
- [25] M. Niazi, A. Parwani, and M. Gurcan. Digital pathology and artificial intelligence. *The Lancet. Oncology*, 20 5:e253–e261, 2019.
- [26] Xi Wang, Hao Chen, Caixia Gan, Huangjing Lin, Q. Dou, E. Tsougenis, Qitao Huang, Muyan Cai, and P. Heng. Weakly supervised deep learning for whole slide lung cancer image analysis. *IEEE Transactions on Cybernetics*, 50:3950–3962, 2020.
- [27] Gabriele Campanella, M. Hanna, Luke Geneslaw, Allen P. Mirafior, Vitor Werneck Krauss Silva, K. Busam, E. Brogi, V. Reuter, D. Klimstra, and Thomas J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25:1301 – 1309, 2019.
- [28] Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data efficient and weakly supervised computational pathology on whole slide images. *Nature biomedical engineering*, 2021.
- [29] Pushpak Pati, Guillaume Jaume, A. Foncubierta, Florinda Feroce, A. Anniciello, G. Scognamiglio, N. Brancati, M. Fiche, E. Dubruc, D. Riccio, M. Bonito, G. Pietro, G. Botti, J. Thiran, Maria Frucci, O. Goksel, and M. Gabrani. Hierarchical graph representations in digital pathology. *Medical image analysis*, 75:102264, 2021.
- [30] Richard J Chen and Rahul G Krishnan. Self-supervised vision transformers learn visual concepts in histopathology. *Learning Meaningful Representations of Life, NeurIPS 2021*, 2021.
- [31] Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16144–16155, June 2022.
- [32] Gabriele Campanella, Ricky Kwan, Eugene Fluder, Jennifer Zeng, A. Stock, Brandon Veremis, A. Polydorides, Cyrus Hedvat, Adam Schoenfeld, Chad Vanderbilt, P. Kovatch, Carlos Cordon-Cardo, and Thomas J. Fuchs. Computational pathology at health system scale - self-supervised foundation models from three billion images. *ArXiv*, abs/2310.07033, 2023.
- [33] J. Ludwig and J. Weinstein. Biomarkers in cancer staging, prognosis and treatment selection. *Nature Reviews Cancer*, 5:845–856, 2005.

- [34] M. Lyngbakken, P. Myhre, H. Røsjø, and T. Omland. Novel biomarkers of cardiovascular disease: Applications in clinical practice. *Critical Reviews in Clinical Laboratory Sciences*, 56:33 – 60, 2018.
- [35] Robert M Califf. Biomarker definitions and their applications. *Experimental Biology and Medicine*, 243(3):213–221, February 2018.
- [36] Eugene Vorontsov, A. Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Siqu Liu, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, Eric Robert, Yi Kan Wang, Jeremy D. Kun, Matthew C. H. Le, Jan Bernhard, Ran Godrich, Gerard Oakley, Ewan Millar, Matthew Hanna, J. Retamero, W. A. Moye, Razik Yousfi, Christopher Kanan, D. Klimstra, B. Rothrock, and Thomas J. Fuchs. Virchow: A million-slide digital pathology foundation model. *ArXiv*, abs/2309.07778, 2023.
- [37] Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, David Klimstra, Razik Yousfi, Thomas Fuchs, Nicolo Fusi, Siqu Liu, and Kristen Severson. Virchow2: Scaling self-supervised mixed magnification models in pathology, 2024.
- [38] Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, March 2024.
- [39] Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, March 2024.
- [40] Jingwei Zhang, Ke Ma, Saarthak Kapse, Joel Saltz, Maria Vakalopoulou, Prateek Prasanna, and Dimitris Samaras. Sam-path: A segment anything model for semantic segmentation in digital pathology, 2023.
- [41] Zhi Huang, Federico Bianchi, Mert Yuksekogul, Thomas J. Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, 29(9):2307–2316, August 2023.
- [42] Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Melissa Zhao, Aaron K. Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, Amr Soliman, Chengkuan Chen, Tong Ding, Judy J. Wang, Georg Gerber, Ivy Liang, Long Phi Le, Anil V. Parwani, Luca L. Weishaupt, and Faisal Mahmood. A multimodal generative ai copilot for human pathology. *Nature*, 634(8033):466–473, June 2024.

- [43] Jana Lipkova, Richard J. Chen, Bowen Chen, Ming Y. Lu, Matteo Barbieri, Daniel Shao, Anurag J. Vaidya, Chengkuan Chen, Luoting Zhuang, Drew F.K. Williamson, Muhammad Shaban, Tiffany Y. Chen, and Faisal Mahmood. Artificial intelligence for multimodal data integration in oncology. *Cancer Cell*, 40(10):1095–1110, October 2022.
- [44] P. Moulin, K. Grünberg, Erio Barale-Thomas, and Jeroen van der Laak. Imi—bigpicture: A central repository for digital pathology. *Toxicologic Pathology*, 49:711 – 713, 2021.
- [45] S Rao, S Masilamani, and S et al Sundaram. Quality measures in pre-analytical phase of tissue processing: understanding its value in histopathology. *Journal of Clinical Diagnostic Research*, 10:EC07–EC11, 2016.
- [46] College of American Pathologist. *Laboratory accreditation manual*, 2018. Accessed 10 Oct 2023.
- [47] Italian Ministry of Health. Linee guida tracciabilità, raccolta, trasporto, conservazione e archiviazione di cellule e tessuti per indagini diagnostiche, 2015. Accessed 10 Oct 2023.
- [48] AA Evangelopoulos, M Dalamaga, K Panoutsopoulos, and K Dima. Nomenclature and basic concepts in automation in the clinical laboratory setting: a practical glossary. *Clinical Laboratory*, 59:1197–1214, 2013.
- [49] JR Srigley, M Judge, and T et al Helliwell. The international collaboration on cancer reporting (iccr): a decade of progress towards global pathology standardisation and data interoperability. *Histopathology*, 79:897–901, 2021.
- [50] V Baxi, R Edwards, M Montalto, and S Saha. Digital pathology and artificial intelligence in translational medicine and clinical practice. *Modern Pathology*, 35:23–32, 2022.
- [51] J van der Laak, G Litjens, and F Ciompi. Deep learning in histopathology: the path to the clinic. *Nature Medicine*, 27:775–784, 2021.
- [52] F Muscarella and Bonis De et al. Automated workflow for somatic and germline next generation sequencing analysis in routine clinical cancer diagnostics. *Cancers (Basel)*, 11:1691, 2019.
- [53] RL Sirota. Defining error in anatomic pathology. *Archives of Pathology and Laboratory Medicine*, 130:604–606, 2006.
- [54] RJ Zarbo, FA Meier, and SS Raab. Error detection in anatomic pathology. *Archives of Pathology and Laboratory Medicine*, 129:1237–1245, 2005.
- [55] RE Nakhleh. Core components of a comprehensive quality assurance program in anatomic pathology. *Advances in Anatomic Pathology*, 16:418–423, 2009.
- [56] L Pantanowitz, AC Mackinnon, and JH Sinard. Tracking in anatomic pathology. *Archives of Pathology and Laboratory Medicine*, 137:1798–1810, 2013.

- [57] MG Hanna and L Pantanowitz. Bar coding and tracking in pathology. *Surgical Pathology Clinics*, 8:123–135, 2015.
- [58] T Carll, C Fuja, and T et al Antic. Tissue contamination during transportation of formalin-fixed, paraffin-embedded blocks. *American Journal of Clinical Pathology*, 158:96–104, 2022.
- [59] MR Wick. The hematoxylin and eosin stain in anatomic pathology-an often-neglected focus of quality assurance in the laboratory. *Seminars in Diagnostic Pathology*, 36:303–311, 2019.
- [60] JD Martina, C Simmons, and DM Jukic. High-definition hematoxylin and eosin staining in a transition to digital pathology. *Journal of Pathology Informatics*, 2:45, 2011.
- [61] D Ferreira, J Vale, and M et al Curado. The impact of different coverslipping methods in the quality of the whole slide images used for diagnosis in pathology. *Journal of Pathology Informatics*, 13:100098, 2022.
- [62] H Herbst and T R”udiger. Automatisierung und der einsatz von robotern im pathologielaor. *Die Pathologie*, 43:56–63, 2022.
- [63] S Marletta, L Pantanowitz, and D et al Malvi. Validation of portable tablets for transplant pathology diagnosis according to the college of american pathologists guidelines. *Academic Pathology*, 9:100047, 2022.
- [64] C Mercan, M Balkenhol, and R et al Salgado. Deep learning for fully-automated nuclear pleomorphism scoring in breast cancer. *npj Breast Cancer*, 8:120, 2022.
- [65] J-M Bokhorst, ID Nagtegaal, and I et al Zlobec. Semi-supervised learning to automate tumor bud detection in cytokeratin-stained whole-slide images of colorectal cancer. *Cancers (Basel)*, 15:2079, 2023.
- [66] S Jarkman, M Karlberg, and M et al Pocevičiūtė. Generalization of deep learning in digital pathology: experience in breast cancer metastasis detection. *Cancers (Basel)*, 14:5424, 2022.
- [67] K Kartasalo, W Bulten, and B et al Delahunt. Artificial intelligence for diagnosis and gleason grading of prostate cancer in biopsies-current status and next steps. *European Urology Focus*, 7:687–691, 2021.
- [68] W Bulten, M Balkenhol, and JJA et al Belinga. Artificial intelligence assistance significantly improves gleason grading of prostate biopsies by pathologists. *Modern Pathology*, 34:660–671, 2021.
- [69] CN Heinz, A Echle, and S et al Foersch. The future of artificial intelligence in digital pathology - results of a survey across stakeholder groups. *Histopathology*, 80:1121–1127, 2022.
- [70] E Munari, FR Mariotti, and L et al Quatrini. Pd-1/pd-l1 in cancer: pathophysiological, diagnostic and therapeutic aspects. *International Journal of Molecular Sciences*, 22:5123, 2021.

- [71] S Lu, JE Stein, and DL et al Rimm. Comparison of biomarker modalities for predicting response to pd-1/pd-l1 checkpoint blockade: a systematic review and meta-analysis. *JAMA Oncology*, 5:1195–1204, 2019.
- [72] S Park, CY Ock, and H et al Kim. Artificial intelligence-powered spatial analysis of tumor-infiltrating lymphocytes as complementary biomarker for immune checkpoint inhibition in non-small-cell lung cancer. *Journal of Clinical Oncology*, 40:1916–1928, 2022.
- [73] I Girolami, L Pantanowitz, and S et al Marletta. Artificial intelligence applications for pre-implantation kidney biopsy pathology practice: a systematic review. *Journal of Nephrology*, 35:1801–1808, 2022.
- [74] F Neri, A Eccher, and P et al Rigotti. Advantages of using a web-based digital platform for kidney preimplantation biopsies. *Journal of Pathology Informatics*, 12:41, 2021.
- [75] A Eccher, I Girolami, and M et al Brunelli. Digital pathology for second opinion consultation and donor assessment during organ procurement: review of the literature and guidance for deployment in transplant practice. *Transplant Reviews*, 34:100562, 2020.
- [76] J Ho, SM Ahlers, and C et al Stratman. Can digital pathology result in cost savings? a financial projection for digital pathology implementation at a large integrated health care organization. *Journal of Pathology Informatics*, 5:33, 2014.
- [77] J Temprana-Salvador, P López-García, and J et al Castellví Vives. Digipatics: digital pathology transformation of the catalan health institute network of 8 hospitals—planning, implementation, and preliminary results. *Diagnostics*, 12:852, 2022.
- [78] F Fraggetta, A Caputo, and R et al Guglielmino. A survival guide for the rapid transition to a fully digital workflow: the "caltagirone example". *Diagnostics*, 11:1916, 2021.
- [79] F Fraggetta, S Garozzo, and GF et al Zannoni. Routine digital pathology workflow: the catania experience. *Journal of Pathology Informatics*, 8:51, 2017.
- [80] C Eloy, J Vale, and M et al Curado. Digital pathology workflow implementation at ipatimup. *Diagnostics*, 11:2111, 2021.
- [81] LA Hassell, SF Absar, and C et al Chauhan. Pathology education powered by virtual and digital transformation: now and the future. *Archives of Pathology and Laboratory Medicine*, 147:474–491, 2023.
- [82] K Bera, KA Schalper, and DL et al Rimm. Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. *Nature Reviews Clinical Oncology*, 16:703–715, 2019.

- [83] B Sch"omig-Markiefka, A Pryalukhin, and W et al Hulla. Quality control stress test for deep learning-based diagnostic model in digital pathology. *Modern Pathology*, 34:2098–2108, 2021.
- [84] TR Helliwell, MJ Judge, and GG et al Birdsong. The international collaboration on cancer reporting (iccr): 10 years progress in the development of cancer pathology datasets. *International Journal of Gynecological Pathology*, 41:S3–S7, 2022.
- [85] DW Ellis. Surgical pathology reporting at the crossroads: beyond synoptic reporting. *Pathology*, 43:404–409, 2011.
- [86] DW Ellis and J Srigley. Does standardised structured reporting contribute to quality in diagnostic pathology? the importance of evidence-based datasets. *Virchows Archiv*, 468:51–59, 2016.
- [87] M Casparie, ATMG Tiebosch, and G et al Burger. Pathology databanking and biobanking in the netherlands, a central role for palga, the nationwide histopathology and cytopathology data network and archive. *Cell Oncology*, 29:19–24, 2007.
- [88] A Hojat, B Wei, and MG et al Olson. Procurement and storage of surgical biospecimens. *Methods in Molecular Biology*, 1897:65–76, 2019.
- [89] S Kapila, K Boaz, and S Natarajan. The post-analytical phase of histopathology practice: storage, retention and use of human tissue specimens. *International Journal of Applied and Basic Medical Research*, 6:3, 2016.
- [90] A Eccher, A Scarpa, and AP Dei Tos. Impact of a centralized archive for pathology laboratories on the health system. *Pathology - Research and Practice*, 245:154488, 2023.
- [91] A Eccher, AP Dei Tos, and A et al Scarpa. Cost analysis of archives in the pathology laboratories: from safety to management. *Journal of Clinical Pathology*, 76:659–663, 2023.
- [92] T Kalinski, H Hofmann, DS Franke, and A Roessner. Digital imaging and electronic patient records in pathology using an integrated department information system with pacs. *Pathology Research and Practice*, 198:679–684, 2002.
- [93] M Ali, H Evans, and P et al Whitney. Using systemised nomenclature of medicine (snomed) codes to select digital pathology whole slide images for long-term archiving. *Journal of Clinical Pathology*, 76:349–352, 2023.
- [94] A Huisman, A Looijen, SM van den Brink, and PJ van Diest. Creation of a fully digital pathology slide archive by high-volume tissue slide scanning. *Human Pathology*, 41:751–757, 2010.
- [95] L Pantanowitz, A Sharma, and AB et al Carter. Twenty years of digital pathology: an overview of the road travelled, what is on the horizon, and the emergence of vendor-neutral archives. *Journal of Pathology Informatics*, 9:40, 2018.

- [96] C Gaudet-Blavignac, V Fouf, M Bjelogrljic, and C Lovis. Use of the systematized nomenclature of medicine clinical terms (snomed ct) for processing free text in health care: systematic scoping review. *Journal of Medical Internet Research*, 23:e24594, 2021.
- [97] Lucian Negreanu, Theodor Voiosu, Monica State, Andrei Voiosu, Andreea Bengus, and Bogdan Radu Mateescu. Endoscopy in inflammatory bowel disease: from guidelines to real life. *Therapeutic Advances in Gastroenterology*, 12, 2019.
- [98] Guihua Chen and Jun Shen. Artificial Intelligence Enhances Studies on Inflammatory Bowel Disease. *Frontiers in Bioengineering and Biotechnology*, 9:570, 2021.
- [99] Gian Eugenio Tontini, Alessandro Rimondi, Marta Venero, Helmut Neumann, Maurizio Vecchi, Cristina Bezzio, and Flaminia Cavallaro. Artificial intelligence in gastrointestinal endoscopy for inflammatory bowel disease: a systematic review and new horizons. *Therapeutic Advances in Gastroenterology*, 14, 2021.
- [100] Georg Semmler, Sarah Wernly, Bernhard Wernly, Behrooz Mamandipoor, Sebastian Bachmayer, Lorenz Semmler, Elmar Aigner, Christian Datz, and Venet Osmani. Machine learning models cannot replace screening colonoscopy for the prediction of advanced colorectal adenoma. *Journal of Personalized Medicine*, 11(10), 2021.
- [101] G Semmler, S Wernly, B Wernly, B Mamandipoor, S Bachmayer, L Semmler, E Aigner, C Datz, and V Osmani. Gastroenterologist against the machine - opportunities and limitations of machine learning models for prediction of advanced adenoma. In *Zeitschrift für Gastroenterologie*, volume 59, 2021.
- [102] Ryan W Stidham, Wenshuo Liu, Shrinivas Bishu, Michael D Rice, Peter DR Higgins, Ji Zhu, Brahmajee K Nallamotheu, and Akbar K Waljee. Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with ulcerative colitis. *JAMA network open*, 2(5):e193963–e193963, 2019.
- [103] Kento Takenaka, Kazuo Ohtsuka, Toshimitsu Fujii, Mariko Negi, Kohei Suzuki, Hiromichi Shimizu, Shiori Oshima, Shintaro Akiyama, Maiko Motobayashi, Masakazu Nagahori, et al. Development and validation of a deep neural network for accurate evaluation of endoscopic images from patients with ulcerative colitis. *Gastroenterology*, 158(8):2150–2157, 2020.
- [104] Amir Klein, Yoav Mazor, Amir Karban, Ofer Ben-Itzhak, Yehuda Chowers, and Edmond Sabo. Early histological findings may predict the clinical phenotype in crohn’s colitis. *United European Gastroenterology Journal*, 5(5):694–701, 2017.
- [105] Akbar K Waljee, Beth I Wallace, Shirley Cohen-Mekelburg, Yumu Liu, Boang Liu, Kay Sauder, Ryan W Stidham, Ji Zhu, and Peter DR Higgins. Development and validation of machine learning models in prediction of remis-

- sion in patients with moderate to severe crohn disease. *JAMA network open*, 2(5):e193721–e193721, 2019.
- [106] Akbar K Waljee, Rachel Lipson, Wyndy L Wiitala, Yiwei Zhang, Boang Liu, Ji Zhu, Beth Wallace, Shail M Govani, Ryan W Stidham, Rodney Hayward, et al. Predicting hospitalization and outpatient corticosteroid use in inflammatory bowel disease patients using machine learning. *Inflammatory bowel diseases*, 24(1):45–53, 2018.
- [107] Shirley Cohen-Mekelburg, Sameer Berry, Ryan W Stidham, Ji Zhu, and Akbar K Waljee. Clinical applications of artificial intelligence and machine learning-based methods in inflammatory bowel disease. *Journal of Gastroenterology and Hepatology*, 36(2):279–285, 2021.
- [108] Pietro Cinaglia, Luciano Caroprese, Giuseppe Lucio Cascini, Francesco Dattola, Pasquale Iaquinta, Miriam Iusi, Pierangelo Veltri, and Ester Zumpano. Bioinformatics solutions for image data processing. In Robert Koprowski, editor, *Medical and Biological Image Analysis*, chapter 4. IntechOpen, Rijeka, 2018.
- [109] Suneha Sundaram, Tenzin Choden, Mark C. Mattar, Sanjal Desai, and Madhav Desai. Artificial intelligence in inflammatory bowel disease endoscopy: current landscape and the road ahead. *Therapeutic Advances in Gastrointestinal Endoscopy*, 14, 2021.
- [110] 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 (CVPR)*, pages 770–778. IEEE, 2016.
- [111] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical Evaluation of Rectified Activations in Convolutional Network. arXiv:1505.00853, 2015.
- [112] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.
- [113] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.
- [114] Hee E Kim, Alejandro Cosa-Linan, Nandhini Santhanam, Mahboubeh Janesari, Mate E Maros, and Thomas Ganslandt. Transfer learning for medical image classification: a literature review. *BMC medical imaging*, 22(1):1–13, 2022.
- [115] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. arXiv:1412.6980, Proceedings of the 3rd International Conference for Learning Representations (ICLR), 2015.
- [116] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

- [117] L. N. Smith. Cyclical Learning Rates for Training Neural Networks. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472. IEEE, 2017.
- [118] D. Chicco and G. Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21:6, 2020.
- [119] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.
- [120] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.
- [121] M Ellrichmann, P Wietzke-Braun, S Dhar, S Nikolaus, A Arlt, J Bethge, T Kuehbacher, L Wintermeyer, K Balschun, W Klapper, et al. Endoscopic ultrasound of the colon for the differentiation of crohn’s disease and ulcerative colitis in comparison with healthy controls. *Alimentary pharmacology & therapeutics*, 39(8):823–833, 2014.
- [122] Pietro Cinaglia, Giuseppe Tradigo, Giuseppe L. Cascini, Ester Zumpano, and Pierangelo Veltri. A framework for the decomposition and features extraction from lung dicom images. In *Proceedings of the 22nd International Database Engineering & Applications Symposium, IDEAS 2018*, page 31–36, New York, NY, USA, 2018. Association for Computing Machinery.
- [123] A. Caputo, V. L’Imperio, F. Merolla, I. Girolami, E. Leoni, V. Della Mea, F. Pagni, and F. Fraggetta. The slow-paced digital evolution of pathology: lights and shadows from a multifaceted board. *Pathologica*, 115(3):127, June 2023.
- [124] Y. Makhoulf, M. Salto-Tellez, J. James, P. O’Reilly, and P. Maxwell. General roadmap and core steps for the development of ai tools in digital pathology. *Diagnostics*, 12(5):1272, May 2022.
- [125] S.W. Jahn, M. Plass, and F. Moinfar. Digital pathology: advantages, limitations and emerging perspectives. *Journal of Clinical Medicine*, 9(11):3697, November 2020.
- [126] B.E. Bejnordi, M. Veta, P.J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J.A.W.M. Van Der Laak, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318:2199–2210, 2017.
- [127] N. Coudray, P.S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A.L. Moreira, N. Razavian, and A. Tsirigos. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24:1559–1567, 2018.

- [128] M. Shaban, R. Awan, M.M. Fraz, A. Azam, Y.-W. Tsang, D. Snead, and N.M. Rajpoot. Context-aware convolutional neural network for grading of colorectal cancer histology images. *IEEE Transactions on Medical Imaging*, 39:2395–2405, 2020.
- [129] D. Bychkov, N. Linder, R. Turkki, S. Nordling, P.E. Kovanen, C. Verrill, M. Walliander, M. Lundin, C. Haglund, and J. Lundin. Deep learning based tissue analysis predicts outcome in colorectal cancer. *Scientific Reports*, 8:3395, 2018.
- [130] S. Graham, Q.D. Vu, S.E.A. Raza, A. Azam, Y.W. Tsang, J.T. Kwak, and N. Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, 2019.
- [131] K. Xu, M. Jahanifar, S. Graham, and N. Rajpoot. Accurate segmentation of nuclear instances using a double-stage neural network. In *Medical Imaging 2023: Digital and Computational Pathology*, 2023.
- [132] Y. Makhoulf, V.K. Singh, S. Craig, A. McArdle, D. French, M.B. Loughrey, N. Oliver, J.B. Acevedo, P. O'Reilly, J.A. James, and P. Maxwell. True-t–improving t-cell response quantification with holistic artificial intelligence based prediction in immunohistochemistry images. *Computational and Structural Biotechnology Journal*, 23:174–185, December 2024.
- [133] J. Cho, K. Lee, E. Shin, G. Choy, and S. Do. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? *arXiv preprint arXiv:1511.06348*, 2015.
- [134] W.N. Price and I.G. Cohen. Privacy in the age of medical big data. *Nature Medicine*, 25:37–43, 2019.
- [135] R.J. Chen, M.Y. Lu, T.Y. Chen, D.F.K. Williamson, and F. Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5:493–497, 2021.
- [136] N. Savage. Synthetic data could be better than real data. *Nature*, 2023.
- [137] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63:139–144, 2020.
- [138] A.F. Frangi, S.A. Tsafaris, and J.L. Prince. Simulation and synthesis in medical imaging. *IEEE Transactions on Medical Imaging*, 37:673–679, 2018.
- [139] T. Zhou, H. Fu, G. Chen, J. Shen, and L. Shao. Hi-net: hybrid-fusion network for multi-modal mr image synthesis. *IEEE Transactions on Medical Imaging*, 39:2772–2781, 2020.
- [140] Y. Tang, Y. Tang, Y. Zhu, J. Xiao, and R.M. Summers. A disentangled generative model for disease decomposition in chest x-rays via normal image synthesis. *Medical Image Analysis*, 67:101839, 2021.

- [141] A. Kapil, A. Meier, A. Zuraw, K.E. Steele, M.C. Rebelatto, G. Schmidt, and N. Brieu. Deep semi supervised generative learning for automated tumor proportion scoring on nslc tissue needle biopsies. *Scientific Reports*, 8:17343, 2018.
- [142] F. Mahmood, D. Borders, R.J. Chen, G.N. McKay, K.J. Salimian, A. Baras, and N.J. Durr. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *IEEE Transactions on Medical Imaging*, 39:3257–3267, 2019.
- [143] A.B. Levine, J. Peng, D. Farnell, M. Nurse, Y. Wang, J.R. Naso, H. Ren, H. Farahani, C. Chen, D. Chiu, et al. Synthesis of diagnostic quality cancer pathology images by generative adversarial networks. *The Journal of Pathology*, 252:178–188, 2020.
- [144] K. Falahkheirkhah, S. Tiwari, K. Yeh, S. Gupta, L. Herrera-Hernandez, M.R. McCarthy, R.E. Jimenez, J.C. Cheville, and R. Bhargava. Deepfake histologic images for enhancing digital pathology. *Laboratory Investigation*, 103:100006, 2023.
- [145] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020.
- [146] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794, 2021.
- [147] G. Müller-Franzes, J.M. Niehues, F. Khader, S.T. Arasteh, C. Haarbuerger, C. Kuhl, T. Wang, T. Han, T. Nolte, S. Nebelung, et al. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Scientific Reports*, 13:12098, 2023.
- [148] W.H.L. Pinaya, P.-D. Tudosiu, J. Dafflon, P.F. Da Costa, V. Fernandez, P. Nachev, S. Ourselin, and M.J. Cardoso. Brain imaging generation with latent diffusion models. In *MICCAI Workshop on Deep Generative Models*, 2022.
- [149] Z. Dorjsembe, S. Odonchimed, and F. Xiao. Three-dimensional medical image synthesis with denoising diffusion probabilistic models. In *Medical Imaging with Deep Learning*, 2022.
- [150] D.J.E. Waibel, E. Röoell, B. Rieck, R. Giryes, and C. Marr. A diffusion model predicts 3d shapes from 2d microscopy images. In *arXiv preprint arXiv:2208.14125*, 2022.
- [151] B. Kim and J.C. Ye. Diffusion deformable model for 4d temporal medical image generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2022.
- [152] P. Chambon, C. Bluethgen, C.P. Langlotz, and A. Chaudhari. Adapting pretrained vision-language foundational models to medical imaging domains. In *arXiv preprint arXiv:2210.04133*, 2022.

- [153] A. Kazerouni, E.K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacıhaliloglu, and D. Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, page 102846, 2023.
- [154] P.A. Moghadam, S. Van Dalen, K.C. Martin, J. Lennerz, S. Yip, H. Farahani, and A. Bashashati. A morphology focused diffusion probabilistic model for synthesis of histopathology images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.
- [155] A. Shrivastava and P.T. Fletcher. Nasdm: Nuclei-aware semantic histopathology image generation using diffusion models. *arXiv preprint arXiv:2303.11477*, 2023.
- [156] X. Xu, S. Kapse, R. Gupta, and P. Prasanna. Vit-dae: Transformer-driven diffusion autoencoder for histopathology image analysis. *arXiv preprint arXiv:2304.01053*, 2023.
- [157] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, volume 29, 2016.
- [158] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- [159] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- [160] M.F. Naeem, S.J. Oh, Y. Uh, Y. Choi, and J. Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, 2020.
- [161] L. Zhang, L. Zhang, and A.C. Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015.
- [162] R. Achibat, M. Dreyer, I. Eisenbraun, S. Bosse, T. Wiegand, W. Samek, and S. Lapuschkin. From "where" to "what": Towards human-understandable explanations through concept relevance propagation. 2022.
- [163] L.J. Carithers, K. Ardlie, M. Barcus, P.A. Branton, A. Britton, S.A. Buia, C.C. Compton, D.S. DeLuca, J. Peter-Demchok, E.T. Gelfand, et al. A novel approach to high-quality postmortem tissue procurement: the gtex project. *Biopreservation and Biobanking*, 13:311–319, 2015.
- [164] L.J. Carithers and H.M. Moore. The genotype-tissue expression (gtex) project. *Biopreservation and Biobanking*, 13:307–308, 2015.
- [165] C.L. Srinidhi, O. Ciga, and A.L. Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67:101813, 2021.

- [166] A. Bizzego, N. Bussola, M. Chierici, V. Maggio, M. Francescatto, L. Cima, M. Cristoforetti, G. Jurman, and C. Furlanello. Evaluating reproducibility of ai algorithms in digital pathology with dapper. *PLoS Computational Biology*, 15:e1006269, 2019.
- [167] A. Marcolini, N. Bussola, E. Arbitrio, M. Amgad, G. Jurman, and C. Furlanello. histolab: A python library for reproducible digital pathology preprocessing with automated testing. *SoftwareX*, 20:101237, 2022.
- [168] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9:62–66, 1979.
- [169] C. Redies, S.A. Amirshahi, M. Koch, and J. Denzler. Phog-derived aesthetic measures applied to color photographs of artworks, natural scenes and objects. In *Computer Vision–ECCV 2012. Workshops and Demonstrations: Florence, Italy, October 7–13, 2012, Proceedings, Part I*, 2012.
- [170] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* 18, 2015.
- [171] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [172] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE, 2009.
- [173] S. Barratt and R. Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.
- [174] T. Kynkäänniemi, T. Karras, M. Aittala, T. Aila, and J. Lehtinen. The role of imagenet classes in fréchet inception distance. *arXiv preprint arXiv:2203.06026*, 2022.
- [175] Y. LeCun, Y. Bengio, et al. Convolutional networks for images, speech, and time series. In *The Handbook of Brain Theory and Neural Networks*, volume 3361, page 1995. 1995.
- [176] N. Bussola, A. Marcolini, V. Maggio, G. Jurman, and C. Furlanello. Ai slipping on tiles: Data leakage in digital pathology. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part I*, 2021.
- [177] A. Madabhushi and G. Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, 33:170–175, 2016.
- [178] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

- [179] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.
- [180] B.W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405:442–451, 1975.
- [181] P. Baldi, S. Brunak, Y. Chauvin, C.A.F. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16:412–424, 2000.
- [182] G. Jurman, S. Riccadonna, and C. Furlanello. A comparison of mcc and cen error measures in multi-class prediction. *PLOS*, 2012.
- [183] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS One*, 10:e0130140, 2015.
- [184] M.D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013.
- [185] Z. Qin, F. Yu, C. Liu, and X. Chen. How convolutional neural networks see the world — a survey of convolutional neural network visualization methods. *Mathematical Foundations of Computing*, 1:149–180, 2018.
- [186] E. McAlpine, P. Michelow, E. Liebenberg, and T. Celik. Is it real or not? toward artificial intelligence-based realistic synthetic cytology image generation to augment teaching and quality assurance in pathology. *Journal of the American Society of Cytopathology*, 11:123–132, 2022.
- [187] Y. Guan, J. Zhang, K. Tian, S. Yang, P. Dong, J. Xiang, W. Yang, J. Huang, Y. Zhang, and X. Han. Node-aligned graph convolutional network for whole-slide image representation and classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
- [188] S. Javed, A. Mahmood, T. Qaiser, N. Werghi, and N. Rajpoot. Unsupervised mutual transformer learning for multi-gigapixel whole slide image classification. In *arXiv preprint arXiv:2305.02032*, 2023.
- [189] S. Deshpande, F. Minhas, S. Graham, and N. Rajpoot. Safron: Stitching across the frontier network for generating colorectal cancer histology images. *Medical Image Analysis*, 77:102337, April 2022.
- [190] S. Deshpande, F. Minhas, and N. Rajpoot. Synthesis of annotated colorectal cancer tissue images from gland layout. *arXiv preprint arXiv:2305.05006*, 2023.
- [191] W. Samek, G. Montavon, S. Lapuschkin, C.J. Anders, and K.-R. Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109:247–278, 2021.

- [192] J.M. Dolezal, R. Wolk, H.M. Hieromnimon, F.M. Howard, A. Srisuwananukorn, D. Karpeyev, S. Ramesh, S. Kochanny, J.W. Kwon, M. Agni, et al. Deep learning generates synthetic cancer histology for explainability and education. *NPJ Precision Oncology*, 7:49, 2023.
- [193] C. Dong, C.C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:295–307, 2015.
- [194] L. Wiechmann and H. M. Kuerer. The molecular journey from ductal carcinoma in situ to invasive breast cancer. *Cancer*, 112:2130–2142, 2008.
- [195] G. Chen, X. F. Ding, K. Pressley, H. Bouamar, B. Wang, G. Zheng, L. E. Broome, A. Nazarullah, A. Brenner, V. Kaklamani, I. Jatoi, and L. Sun. Everolimus inhibits the progression of ductal carcinoma in situ to invasive breast cancer via downregulation of mmp9 expression. *Clinical Cancer Research*, 26:1486–1496, 2020.
- [196] E. J. Groen, L. E. Elshof, L. L. Visser, E. J. Rutgers, H. A. O. Winter-Warnars, E. H. Lips, and J. Wesseling. Finding the balance between over- and under-treatment of ductal carcinoma in situ (dcis). *The Breast*, 31:274–283, 2017.
- [197] W. B. Coleman. Breast ductal carcinoma in situ. *The American Journal of Pathology*, 189:942–945, 2019.
- [198] Monica Morrow, Kimberly J. Van Zee, Lawrence J. Solin, Nehmat Houssami, Mariana Chavez-MacGregor, Jay R. Harris, Janet Horton, Shelley Hwang, Peggy L. Johnson, M. Luke Marinovich, Stuart J. Schnitt, Irene Wapnir, and Meena S. Moran. Society of surgical oncology–american society for radiation oncology–american society of clinical oncology consensus guideline on margins for breast-conserving surgery with whole-breast irradiation in ductal carcinoma in situ. *Practical Radiation Oncology*, 6(5):287–295, September 2016.
- [199] Eileen Rakovitch, Sharon Nofech-Mozes, Wedad Hanna, Rinku Sutradhar, Sumei Gu, Cindy Fong, Alan Tuck, Bruce Youngson, Naomi Miller, Susan J. Done, Martin C. Chang, Sandip Sengupta, Leela Elavathil, Prashant A. Jani, Michel Bonin, Nafisha Lalani, and Lawrence Paszat. Omitting radiation therapy after lumpectomy for pure dcis does not reduce the risk of salvage mastectomy. *The Breast*, 37:181–186, 2018.
- [200] Esther H. Lips, Tapsi Kumar, Anargyros Megalios, Lindy L. Visser, Michael Sheinman, Angelo Fortunato, Vandna Shah, Marlous Hoogstraat, Emi Sei, Diego Mallo, Maria Roman-Escorza, Ahmed A. Ahmed, Mingchu Xu, Alexandra W. van den Belt-Dusebout, Wim Brugman, Anna K. Casasent, Karen Clements, Helen R. Davies, Liping Fu, Anita Grigoriadis, Timothy M. Hardman, Lorraine M. King, Marielle Krete, Petra Kristel, Michiel de Maaker, Carlo C. Maley, Jeffrey R. Marks, Brian A. Menegaz, Lennart Mulder, Frank Nieboer, Salpie Nowinski, Sarah Pinder, Jelmar Quist, Carolina Salinas-Souza, Michael Schaapveld, Marjanka K. Schmidt, Abeer M. Shaaban, Rana Shami, Mathini Sridharan, John Zhang, Hilary Stobart, Deborah Collyar,

- Serena Nik-Zainal, Lodewyk F. A. Wessels, E. Shelley Hwang, Nicholas E. Navin, P. Andrew Futreal, P. Andrew Futreal, E. Shelley Hwang, Jos Jonkers, Jacco, Fariba Behbod, Daniel Rea, Proteeti Bhattacharjee, Donna Pinto, Ellen Verschuur, Marja van Oirsouw, Alastair M. Thompson, Jelle Wesseling, and Elinor J. Sawyer. Genomic analysis defines clonal relationships of ductal carcinoma in situ and recurrent invasive breast cancer. *Nature Genetics*, 54(6):850–860, June 2022.
- [201] Ezzeldin M. Ibrahim, Meteb E. Al-Foheidi, Mubarak M. Al-Mansour, and Ghieth A. Kazkaz. The prognostic value of tumor-infiltrating lymphocytes in triple-negative breast cancer: a meta-analysis. *Breast Cancer Research and Treatment*, 148(3):467–476, November 2014.
- [202] Carsten Denkert, Gunter von Minckwitz, Jan C. Brase, Bruno V. Sinn, Stephan Gade, Ralf Kronenwett, Berit M. Pfitzner, Christoph Salat, Sherene Loi, Wolfgang D. Schmitt, Christian Schem, Karin Fisch, Silvia Darb-Esfahani, Keyur Mehta, Christos Sotiriou, Stephan Wienert, Peter Klare, Fabrice André, Frederick Klauschen, Jens-Uwe Blohmer, Kristin Krappmann, Marcus Schmidt, Hans Tesch, Sherko Kümmel, Peter Sinn, Christian Jackisch, Manfred Dietel, Toralf Reimer, Michael Untch, and Sibylle Loibl. Tumor-infiltrating lymphocytes and response to neoadjuvant chemotherapy with or without carboplatin in human epidermal growth factor receptor 2–positive and triple-negative primary breast cancers. *Journal of Clinical Oncology*, 33(9):983–991, March 2015.
- [203] Maria Vittoria Dieci, Nina Radosevic-Robin, Susan Fineberg, Gert van den Eynden, Nils Ternes, Frederique Penault-Llorca, Giancarlo Pruneri, Timothy M. D’Alfonso, Sandra Demaria, Carlos Castaneda, Joselyn Sanchez, Sunil Badve, Stefan Michiels, Veerle Bossuyt, Federico Rojo, Baljit Singh, Torsten Nielsen, Giuseppe Viale, Seong-Rim Kim, Stephen Hewitt, Stephan Wienert, Sybille Loibl, David Rimm, Fraser Symmans, Carsten Denkert, Sylvia Adams, Sherene Loi, and Roberto Salgado. Update on tumor-infiltrating lymphocytes (tils) in breast cancer, including recommendations to assess tils in residual disease after neoadjuvant therapy and in carcinoma in situ: A report of the international immuno-oncology biomarker working group on breast cancer. *Seminars in Cancer Biology*, 52:16–25, October 2018.
- [204] M. K. Komforti, S. Badve, B. Harmon, Y. Lo, and S. Fineberg. Tumour-infiltrating lymphocytes in ductal carcinoma in situ (dcis)—assessment with three different methodologies and correlation with oncotype dx dcis score. *Histopathology*, 77:749–759, 2020.
- [205] Michael S. Toss, Asima Abidi, Dorothea Lesche, Chitra Joseph, Sakshi Mahale, Hugo Saunders, Tanjina Kader, Islam M. Miligy, Andrew R. Green, Kylie L. Gorringer, and Emad A. Rakha. The prognostic significance of immune microenvironment in breast ductal carcinoma in situ. *British Journal of Cancer*, 122(10):1496–1506, March 2020.
- [206] A. Farolfi, E. Petracci, L. Serra, A. Ravaioli, S. Bravaccini, S. Ravaioli, M. M. Tumedei, P. Ulivi, M. Canale, M. Puccetti, F. Falcini, S. Folli, A. Curcio, and

- A. Rocca. Tumor-infiltrating lymphocytes (tils) and risk of a second breast event after a ductal carcinoma in situ. *Frontiers in Oncology*, 10, 2020.
- [207] P. Narayanan, S. E. A. Raza, A. Hall, J. R. Marks, L. King, R. B. West, L. Hernández, N. Guppy, M. Dowsett, B. A. Gusterson, C. C. Maley, E. S. Hwang, and Y. Yuan. Unmasking the immune microecology of ductal carcinoma in situ with deep learning. *NPJ Breast Cancer*, 7, 2021.
- [208] Yeman Brhane Hagos, Faranak Sobhani, Simon P. Castillo, Allison H. Hall, Khalid AbdulJabbar, Roberto Salgado, Bryan Harmon, Kristalyn Gallagher, Mark Kilgore, Lorraine M. King, Jeffrey R. Marks, Carlo Maley, Hugo M. Horlings, Robert West, E. Shelley Hwang, and Yinyin Yuan. *DCIS AI-TIL: Ductal Carcinoma In Situ Tumour Infiltrating Lymphocyte Scoring Using Artificial Intelligence*, page 164–175. Springer Nature Switzerland, 2022.
- [209] Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierto, Gerardo Botti, Maria Gabrani, Florinda Feroce, and Maria Frucci. BRACS: A Dataset for BReAst Carcinoma Subtyping in H &E Histology Images. *Database*, 2022:baac093, 10 2022.
- [210] Peter Bankhead, Maurice B. Loughrey, José A. Fernández, Yvonne Dombrowski, Darragh G. McArt, Philip D. Dunne, Stephen McQuaid, Ronan T. Gray, Liam J. Murray, Helen G. Coleman, Jacqueline A. James, Manuel Salto-Tellez, and Peter W. Hamilton. Qupath: Open source software for digital pathology image analysis. *Scientific Reports*, 7(1), December 2017.
- [211] Radboudumc Computational Pathology Group. Tiger challenge, 2022.
- [212] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, December 2020.
- [213] Joey Spronck, Thijs Gelton, Leander van Eekelen, Joep Bogaerts, Leslie Tessier, Mart van Rijnthoven, Lieke van der Woude, Michel van den Heuvel, Willemijn Theelen, Jeroen van der Laak, and Francesco Ciompi. nnUNet meets pathology: bridging the gap for application to whole-slide images and computational biomarkers. In *Medical Imaging with Deep Learning*, 2023.
- [214] Zaneta Swiderska-Chadaj, Hans Pinckaers, Mart van Rijnthoven, Maschenka Balkenhol, Margarita Melnikova, Oscar Geessink, Quirine Manson, Mark Sherman, Antonio Polonia, Jeremy Parry, Mustapha Abubakar, Geert Litjens, Jeroen van der Laak, and Francesco Ciompi. Learning to detect lymphocytes in immunohistochemistry with deep learning. *Medical Image Analysis*, 58:101547, 2019.