UNIVERSIDAD NACIONAL DE COLOMBIA

# A Deep Learning model for automatic grading of prostate cancer histopathology images

**Sebastian Medina Carrillo**

Universidad Nacional de Colombia
Departamento de Ingeniería de Sistemas e Industrial
Bogotá, Colombia
2024

# A Deep Learning model for automatic grading of prostate cancer histopathology images

**Sebastian Medina Carrillo**

Submitted to the Engineering School of the National University of Colombia, in partial fulfillment of the requirements for the degree of:
**Master of Science**
**Systems and Computer Engineering**

Advisor:
Fabio A. González PhD.

Co-advisor:
Ángel A. Cruz Roa PhD.

Research area:
Intelligent Systems

Research Group:
MindLab Research Group

Universidad Nacional de Colombia
Departamento de Ingeniería de Sistemas e Industrial
Bogotá, Colombia
2024

# Abstract

**Title: A Deep Learning model for automatic grading of prostate cancer histopathology images**

Gleason grading is recognized as the standard method for diagnosing prostate cancer. However, it is subject to significant inter-observer variability due to its reliance on subjective visual assessment. Current deep learning approaches for grading often require exhaustive pixel-level annotations and are generally limited to patch-level predictions, which do not incorporate slide-level information. Recently, weakly-supervised techniques have shown promise in generating whole-slide label predictions using pathology report labels, which are more readily available. However, these methods frequently lack visual and quantitative interpretability, reinforcing the black box nature of deep learning models, hindering their clinical adoption. This thesis introduces WiSDoM, a novel weakly-supervised and interpretable approach leveraging attention mechanisms and Kernel Density Matrices for the grading of prostate cancer on whole slides. This method is adaptable to varying levels of supervision. WiSDoM facilitates multi-scale interpretability through several features: detailed heatmaps that provide granular visual insights by highlighting critical morphological features without requiring tissue annotations; example-based phenotypical prototypes that illustrate the internal representation learned by the model, aiding in clinical verification; and visual-quantitative measures of model uncertainty, which enhance the transparency of the model's decision-making process, a crucial factor for clinical use. WiSDoM has been validated on core-needle biopsies from two different institutions, demonstrating robust agreement with the reference standard (quadratically weighted $\kappa$ of 0.93). WiSDoM achieves state-of-the-art inter-observer agreement performance on the PANDA Challenge publicly available dataset while being clinically interpretable.

**Keywords: Prostate Cancer, Histopathology, Deep Learning, Cancer Grading, Density Matrix, Interpretability**

# Resumen

**Título: Modelo de Deep Learning para la gradación automática de imágenes histopatológicas de cáncer de próstata**

La clasificación de Gleason se reconoce como el método estándar para diagnosticar el cáncer de próstata. Sin embargo, está sujeto a una variabilidad significativa entre observadores debido a su dependencia de la evaluación visual subjetiva. Los enfoques actuales de aprendizaje profundo a menudo requieren anotaciones exhaustivas a nivel de píxeles y generalmente se limitan a predicciones a nivel de parche, que no incorporan información a nivel de lámina. Recientemente, las técnicas débilmente supervisadas se han mostrado prometedoras a la hora de generar predicciones de etiquetas de láminas completas utilizando etiquetas de informes de patología, que están más fácilmente disponibles. Sin embargo, estos métodos frecuentemente carecen de interpretabilidad visual y cuantitativa, lo que refuerza la naturaleza de caja negra de los modelos de aprendizaje profundo y dificulta su adopción clínica. Esta tesis introduce WiSDoM, un enfoque novedoso interpretable y débilmente supervisado que aprovecha los mecanismos de atención y las matrices de densidad para gradar cáncer de próstata en láminas completas. Este método se adapta a distintos niveles de supervisión. WiSDoM facilita la interpretabilidad a múltiples escalas a través de varias características: mapas de calor detallados que brindan información visual granular al resaltar características morfológicas críticas sin requerir anotaciones de tejido; prototipos fenotípicos basados en ejemplos que ilustran la representación interna aprendida por el modelo, ayudando en la verificación clínica; y medidas visual-cuantitativas de incertidumbre del modelo, que mejoran la transparencia del proceso de toma de decisiones, un factor crucial para el uso clínico. WiSDoM se ha validado en biopsias de dos instituciones diferentes, lo que demuestra una sólida concordancia con el estándar de referencia ($\kappa$ ponderado cuadráticamente de 0,93). WiSDoM logra un rendimiento del estado del arte de acuerdo entre observadores en el conjunto de datos PANDA Challenge además de ser clínicamente interpretable.

**Palabras clave: Cáncer de prostata, Histopatología, Aprendizaje automático, Gradación de cáncer, Matriz de densidad, Interpretabilidad**

Esta tesis de maestría se sustentó el 19 de Abril de 2024 a las 4:00 pm., y fue evaluada por
los siguientes jurados:

**Eduardo Romero, MD, PhD**
Facultad de Medicina
Universidad Nacional de Colombia

**Reinel Tabares Soto, PhD**
Departmento de Sistemas e Informática
Universidad de Caldas

# Acknowledgements

Research reported in this publication was partially supported by projects BPIN 2019000100-060 ”Implementation of a Network for Research, Technological Development and Innovation in Digital Pathology (RedPat) supported by Industry 4.0 technologies” from FCTeI of SGR resources, approved by OCAD of FCTeI and MinCiencias, and project 110192092354, entitled ”Program for the Early Detection of Premalignant Lesions and Gastric Cancer in urban, rural and dispersed areas in the Department of Nariño” of call No. 920 of 2022 of MinCiencias.

# Content

# List of Figures

# List of Tables

# 1 Introduction

Prostate cancer (PCa) is the second most common cancer in men worldwide, ranking just behind lung cancer, with 1,276,106 new cases and 358,989 deaths (3.8% of all male cancer fatalities) reported in 2018 and the average age at diagnosis is 66 years [2].

Prostate cancer in its early stages often presents no symptoms and may require no treatment. The disease is frequently detected through elevated levels of prostate-specific antigen (PSA > 4 ng/mL), a glycoprotein typically expressed by prostatic tissue. However, due to the presence of increased PSA in men without cancer, tissue biopsy is recommended for confirming malignancy [3].

The diagnosis of prostate cancer heavily relies on the microscopic analysis of prostate tissue obtained through a needle biopsy. Using transrectal ultrasonography to guide the biopsy extraction, 10 to 12 tissue samples are gathered, are formalin-fixed and paraffin embedded and stained with Hematoxylin and Eosin (H&E). A pathologist then examines these samples based on their microscopic architecture and cellular appearance (see figure **1-1**), assigning a primary and secondary Gleason grade [4], each on a scale of 1 to 5. The pathologist determines the Gleason score, which in biopsies is the sum of the predominant pattern and the secondary one, such as 4+3. This traditional classification of prostate cancer into low, intermediate, or high risk is based on the combined Gleason patterns, prostate-specific antigen (PSA) level, and clinical stage [5]. Recognizing the heterogeneity within each risk group and the potential for inter- and intra-observer variability (see figure **1-2**), the International Society of Urological Pathology (ISUP) revised the pathological grading into five categories in 2014 [6] (see table **1-1**). However, despite these modifications, inter- and intra-observer variability remains an issue, limiting the grading system's applicability and reproducibility across different geographic locations and experience levels among uropathologists [7].

Nevertheless, the tissue biopsy examination and pathologist-performed grading using the Gleason/ISUP grade group system remains the gold standard for prostate cancer diagnosis [8]. This system stands as the most crucial prognostic factor [9] and plays a significant role in determining the appropriate treatment strategies [10], making it invaluable for guiding clinical decisions and managing prostate cancer.

## 1.1 Problem Statement

With an annual incidence of 1,2 million new cases of prostate cancer worldwide, a high mortality rate, and the risk of overdiagnosis and overtreatment [11], there is an urgent need

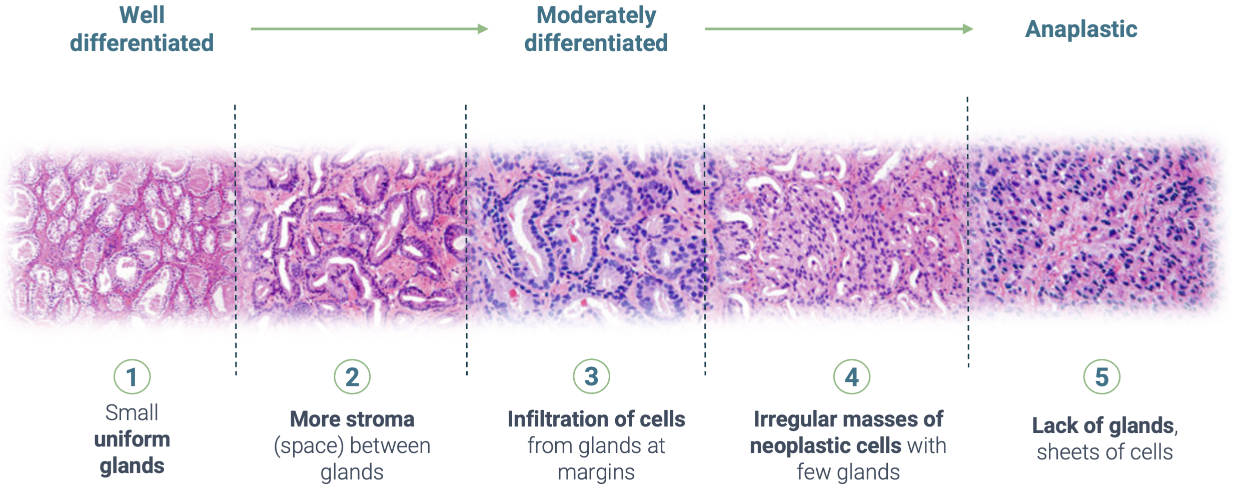
Well differentiated
Moderately differentiated
Anaplastic
1
Small **uniform glands**
2
**More stroma** (space) between glands
3
**Infiltration of cells** from glands at margins
4
**Irregular masses of neoplastic cells** with few glands
5
**Lack of glands**, sheets of cells

**Figure 1-1**: Progressive architectural changes in prostate tissue from Gleason Grade 1 to 5, illustrating the evolution from well-differentiated, organized glandular structures in Grade 1, through increasingly disorganized and poorly differentiated cellular patterns, to the nearly complete loss of glandular architecture and highly irregular, invasive cell formations characteristic of Grade 5.

**Table 1-1**: Summary of the Gleason and ISUP grade group prostate cancer grading systems.

| **Grade Group** | **Description** |
|---|---|
| Grade Group 1 (Gleason score $\leq 6$) | Only individual discrete well-formed glands |
| Grade Group 2 (Gleason score 3+4=7) | Predominantly well-formed glands with a lesser component of poorly-formed/fused/cribriform glands |
| Grade Group 3 (Gleason score 4+3=7) | Predominantly poorly-formed/fused/cribriform glands with a lesser component of well-formed glands |
| Grade Group 4 (Gleason score 8) | Only poorly-formed/fused/cribriform glands or predominantly well-formed glands with a lesser component lacking glands or predominantly lacking glands with a lesser component of well-formed glands |
| Grade Group 5 (Gleason scores 9-10) | Lacks gland formation (or with necrosis) with or without poorly-formed/fused/cribriform glands |

for accurate and reproducible assessment of prognostic survival as well as appropriate treatment of the patient that supports pathologists in making decisions. The Gleason/ISUP score [10, 6], assigned by a pathologist after examining the biopsy, is the standard grading system for determining the prognosis of prostate cancer patients and prescribing the most effective treatment. However, it suffers from significant inter- and intra-observer variability

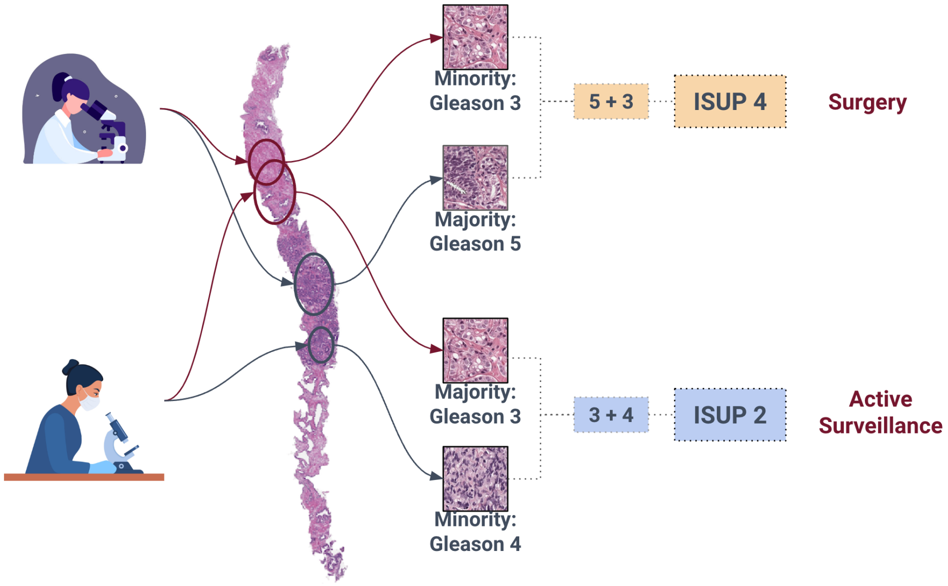

**Figure 1-2**: Illustration of the conventional grading process for a prostate whole-slide image and a depiction of the inter-observer variability when multiple pathologists evaluate the slide. As the quantification of the extent of the two most aggressive patterns relies solely on visual inspection, the diagnosis often differs significantly among pathologists, indicating a notable risk of inconsistency in diagnosis and treatment options.

[12], making it an intriguing problem for reproduction by automatic diagnostic and prognostic support systems that produce results within the range of expert pathologists' variability and agreement. Typically, pathologists with decades of experience have increased agreement rates of $\kappa \approx 0.7$ [13], but this experience of subspecialized pathologists in specific organs and cancer types is uncommon, particularly in developing nations where the majority are general pathologists. Due to the nature of the whole-slide histopathology images (size, intricacy of tissue structures, differences in processing and staining, absence of labels and annotations, and scarcity), the development of an automatic diagnostic and prognostic support system presents a computational challenge [14]. While the normal glands and tissues of the prostate have their natural anatomical variability, the degree of severity associated with the progression and severity of the heterogeneous, poorly differentiated, difficult-to-define, and variable tumor, for which reason its quantification by conventional methods is problematic, varies with the tumor's aggressiveness and severity. The characterization of cellular or tissue morphometry is an additional challenge. In addition to the aforementioned obstacles, the adoption of Deep Learning-based automatic diagnostic support systems is low due to their opaque nature and lack of interpretability. To have confidence in their decisions, pathologists anticipate being able to interpret the decisions made by automated models. The ability to explain and interpret system decisions would encourage their acceptance and regulation in

the medical industry [15]. A robust, interpretable Gleason/ISUP grading by a deep learning-based computer diagnostic support (CAD) system that achieves the same level of agreement as expert pathologists could be advantageous for prostate cancer diagnosis. Deep Learning has demonstrated the ability to perform pathological diagnoses [16, 9, 17]. Therefore, it is desirable to investigate its potential in automated Gleason/ISUP grading, with a focus on the interpretability of the model decisions, i.e. the ability to provide the pathologist with, visual cues, perspectives, and objective information regarding the decision-making process.

## 1.2 Objective

To develop a robust and interpretable Deep Learning model for prostate cancer grading from histopathological images.

### 1.2.1 Specific Objectives

- To integrate a collection of heterogeneous histopathological image data, including its preparation, preprocessing, and representation.
- To design a thorough architecture that enables the image's pathological characteristics to be captured, thereby enhancing the interpretability of decisions made from histopathological images.
- To design a robust and accurate Gleason/ISUP score prediction component for prostate cancer histopathological images.
- To systematically evaluate the efficacy of the proposed method using pathologist-labeled data and compare it using problem-appropriate metrics.

## 1.3 Contributions

This thesis presents a novel approach in computational pathology with the introduction of WiSDoM: **I**nterpretable **W**eakly-**S**upervised Kernel **D**ensity **M**atrices (refer to figure **4-1**), an interpretability-constrained, probabilistic deep learning framework. WiSDoM integrates interpretability through attention mechanisms and a novel method to aggregate information to predict a whole-slide label using Kernel Density Matrices (KDM)[18]. It is specifically designed to tackle the dual challenges of supervision and interpretability in medical image analysis.
At the core of WiSDoM is its capability to:

- Predict posterior probability distributions for Gleason and ISUP grades under various levels of supervision, distinguishing it from traditional probabilistic methods by generating explicit discrete distributions, in contrast to Gaussian processes' continuous distributions.

- Aggregate posterior distributions into a comprehensive distribution for an entire whole-slide image, outperforming methodologies like Multiple Instance Learning (MIL) and conventional aggregation techniques.
- Extend the applicability of KDM to weakly-supervised frameworks, utilizing its robust interpretability features. This is particularly important in contexts where supervision is limited, but the need for transparency and interpretability is high.
- Address the challenge of extensive pixel-level annotations in computational pathology by providing detailed modeling of label probability distributions within an ordinal regression framework. This approach portrays cancer progression as a continuum and offers valuable confidence measures.
- Generate detailed heatmaps in both fully and weakly supervised scenarios, essential for identifying diagnostically significant regions and patterns. Additionally, it aids clinical decision-making by producing phenotypic prototypes that reveal deep insights into the decision-making process.

Additionally, the code implementation of WiSDoM is publicly available (see 4.4.1).

## 1.4 Thesis outline

This thesis is structured as follows: Chapter 2 provides an overview of recent advancements and related work in histopathological image analysis, weakly-supervised learning for histopathology and prostate cancer grading, along with interpretability in deep learning models. In Chapter 3, we introduce WiSDoM in a fully-supervised learning task for patch Gleason grading, explains the method and results. Chapter 4 describes WiSDoM in a weakly-supervised, interpretable setting in a whole-slide ISUP grading of prostate biopsies, results, interpretability and discussion. Chapter 5 concludes the thesis and discusses potential avenues for future work.

# 2 Background and Related Work

## 2.1 Histopathology image analysis

Histopathology, the microscopic evaluation of tissue abnormalities, serves as the basis of cancer diagnosis, playing a pivotal role in both diagnosing and treating the disease [19]. Despite its integral role in the detection and analysis of cancerous formations, it is afflicted with challenges, notably its time-consuming nature and susceptibility to variances and errors[20].

Traditional clinical practices often necessitate pathologists to conduct histological diagnoses. This process involves visually identifying, semi-quantifying, and integrating a multitude of morphological attributes of the test sample with respect to the underlying disease mechanism. Through rigorous systematic training, pathologists can detect predominant morphological patterns corresponding to predefined criteria and existing clinical presentations, facilitating the classification of their observations. Typically, a histopathological diagnosis serves as the end-product of this process, which is then conveyed to treating physicians in a comprehensive written report (see figure **2-1**). Even though systematic training and adherence to standard guidelines can bolster the analytical process and diagnostic precision, histopathology is inherently challenged by certain limitations. These drawbacks primarily stem from its subjective nature and the discrepancies in visual perception, data amalgamation, and variability among different observers [21]. Interestingly, even amongst pathologists with similar training, variations in interpretations can occur, leading to diagnostic inconsistencies that could hinder optimal patient care [22]. Moreover, the increasing use of non-invasive or minimally invasive procedures for diagnostic sample collection has significantly reduced the size and quality of collected specimens. This not only burdens pathologists but also amplifies the difficulty, given the rising demand for diagnostics that include reporting variables with prognostic or predictive value [14].

The complexities mentioned above are further exacerbated by the variability in companion diagnostic tests used to make treatment decisions. This variability often emanates from a lack of standardization, as well as the spatial and temporal biological heterogeneity in samples [23].

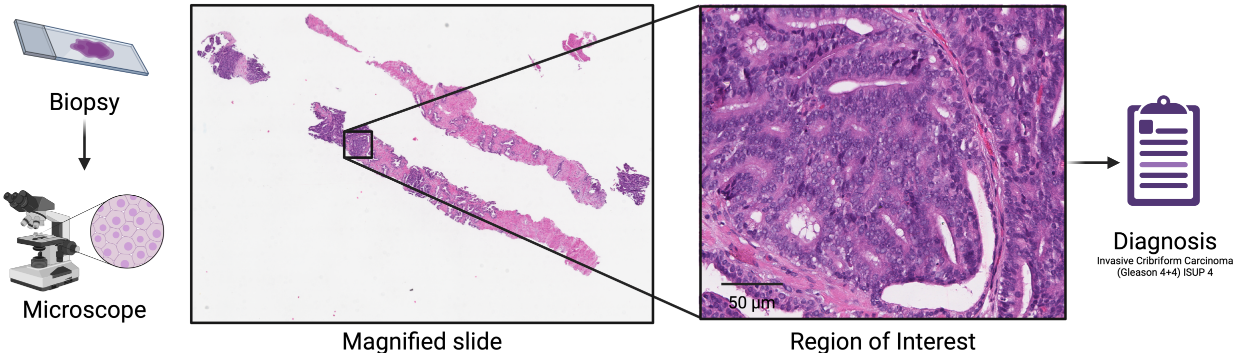

**Figure 2-1**: A pathologist's detailed examination process of a biopsy, starting with the H&E staining of the biopsy slide to highlight cellular and tissue structures. The slide is then placed under a microscope, enabling the pathologist to evaluate the tissue's morphology, including the architecture of glands and the nuclear characteristics of cells. This analysis culminates in a diagnostic report that integrates observations on glandular patterns, cellular differentiation, and potential malignancy indicators, leading to a final diagnosis.

## 2.2 Computational Pathology

To mitigate challenges, the advent of artificial intelligence (AI)-based image analysis techniques are providing remarkable advancements in the realms of pathology and oncology [17, 24, 25]. These innovative tools, designed to augment diagnostic accuracy, serve as a tool for pathologists and oncologists, who are the principal end-users of these ground-breaking technological developments.

A promising development in this area is digital pathology, which capitalizes on advancements in computing power to digitize histopathological specimens using whole-slide scanners and computationally analyze the resulting digital whole-slide images (WSI) [14]. However, the immense complexity and sheer size of these WSI, coupled with their large amount of information, make human interpretation challenging and intrinsically subjective.

Computational pathology emerges as a potential solution to these issues. It is born out of the integration of AI and Machine Learning (ML) tools in pathology and can effectively handle the complexity of WSI [26].

Fundamentally, computational pathology can be divided into two primary branches: feature engineering and deep learning. Feature engineering revolves around creating domain-inspired or domain-agnostic features to enhance the efficiency of machine learning algorithms, while deep learning capitalizes on the learning capabilities of advanced algorithms to interpret complex data. Recently, there has been a perceptible shift towards deep learning due to its ability for autonomous feature extraction, diminished reliance on domain knowledge, and the increased accuracy it provides. This incorporation of deep learning into computational pathology is driving an evolution in the field, setting the stage for enhanced diagnosis and

treatment strategies in medicine [27].
Feature engineering involves constructing domain-inspired algorithms that typically target a specific cancer or tissue type, focusing on particular features that are not universally applicable. These include quantitative counting of mitosis, textural heterogeneity measurements, or domain-agnostic features like graph-based approaches that quantify tissue architecture, shape, or spatial relationships. These feature-based approaches are being developed for various cancers, including prostate, breast, lung, and oral cavity, among others [28, 29, 30].
The shift from manually crafted features to deep learning was motivated by the observation that hierarchical feature representations yielded by these algorithms significantly outperform traditional image analysis methods [31]. This reduces the reliance on extensive domain knowledge and ushers in a more refined and reliable process of cancer detection and treatment. This paradigm shift could represent a breakthrough, introducing a new era of cancer diagnosis characterized by increased precision, consistency, and efficiency. Deep learning reduces implementation time and complexity while improving diagnostic performance [26].
Regardless of the branch, computational pathology has been used for various image processing and classification tasks, from low-level tasks like the detection and segmentation of tissues, nuclei, or glands to high-level tasks like predicting disease diagnosis and prognosis [24, 32]. It is primarily used to automate time-consuming tasks [33, 34, 35, 36], thus enabling pathologists to focus on complex decision-making tasks. Furthermore, it assists oncologists by creating prognostic tools for evaluating disease severity, predicting outcomes, and forecasting response to therapy [37, 38, 39]. The progress and applications of computational pathology are also leaving significant imprints on the field of prostate cancer diagnosis and treatment, which stands at the core of this thesis.

## 2.3 Deep Learning in Prostate Cancer

Prostate cancer diagnosis has particularly benefited from the advancements in computational pathology. Deep learning methodologies have increasingly been applied to the task of grading prostate cancer severity, utilizing the Gleason score and ISUP grade groups. This scoring mechanism plays a crucial role in evaluating disease severity and designing personalized treatment strategies. However, it is subject to both inter- and intra-observer variability [40]. Pathologists with significant expertise often attain a medium degree of agreement while Gleason grading, typically exhibiting an average Kappa $\kappa$ of 0.34 to 0.43 [7]. However, such expertise remains rare due to a limited number of pathologists specializing in organ-specific and cancer-specific areas. General pathologists who lack this specific specialization might only achieve lower Kappa levels. The recent integration of computational pathology techniques has improved accuracy and efficiency, delivering significant results ($\kappa \approx 0.7$ with genitourinary pathologists) [41, 42]. Beyond detection, the realm of grading cancer has witnessed a profound impact. Computational pathology's efficacy in achieving prostate cancer grading surpasses expert pathologist levels (ranging from $\kappa = 0.7$ to $\kappa = 0.9$), indicating its significant potential for more precise and efficient diagnoses [43, 9, 16, 44, 45, 46, 47].

Deep learning has significantly advanced the field of prostate cancer diagnosis, reaching a level comparable to that of pathologists. This progress has been driven by the availability of large datasets of WSI of prostate cancer [48], the initiation of public competitions and challenges [49, 50, 1], and the creation of advanced algorithms [17, 51]. These developments have transformed the way prostate cancer is diagnosed.

Although the strides in prostate cancer diagnosis and grading are immense, they do not stop at detection and classification tasks. The field is evolving to explore methodologies that improve visual interpretability and semantic understanding of learned representations. By broadening the scope beyond classification and venturing into interpretability and semantic comprehension, the stage is being set for more insightful and efficient cancer diagnostics. This focus on interpretability is particularly crucial given the 'black box' nature of deep learning methods, which can be a significant hindrance in their broader acceptance and use in clinical settings [52]. Consequently, research is increasingly focusing on enhancing the interpretability of deep learning models within computational pathology to shed light on these systems' internal operations such as CLAM[17], UNI[51], This Looks Like That[53], among others[54, 55, 56, 57, 58]. This heightened interpretability can build trust in these systems by offering more transparency and the ability to identify and address any potential errors or biases.

Computational pathology has shown the potential to enhance diagnostics, particularly in the field of prostate cancer. Its effectiveness in grading cancer aligns with the performance of urological pathologists with decades of experience [1]. With the promise of improved visual interpretability and semantic comprehension, we are welcoming a new era where diagnoses could provide more in-depth and comprehensive insights, setting the stage for personalized and efficient treatment strategies. However, while the technology offers significant benefits, it is also essential to address its inherent limitations. Doing so will ensure successful integration into the broader clinical ecosystem and pave the way for unlocking the full potential of computational pathology powered by deep learning and AI methods in prostate cancer diagnosis and treatment.

## 2.4 Weakly supervised learning

Despite the recent success of computational pathology applications, significant challenges persist. WSI present a complex landscape with challenges for automated analysis. Deep learning techniques, despite their efficacy and ease of development, often necessitate labor-intensive, pixel-level annotations by pathologists for supervised learning. Weakly supervised learning models have garnered substantial attention in computational pathology due to their proficiency in learning from limited or imprecisely labeled data. This feature is particularly valuable in scenarios where acquiring detailed annotations is challenging or not feasible, a frequent situation in pathology. It has been especially useful for slide-level tasks such as

Gleason grading in prostate cancer[44, 9], subtyping of renal cell carcinoma and non-small-cell lung cancer[17], and sentinel lymph node metastases slide-level classification [59]. These methods typically rely on a single label derived from a pathology report, reducing the need for exhaustive manual annotations.

The primary approach used for weakly supervised learning in computational pathology is Multiple Instance Learning (MIL). In MIL, a WSI is divided into numerous small patches, each inheriting the slide-level label. The model is trained on these patches, and then predictions are aggregated into a single slide-level decision using techniques such as max-pooling. However, this approach under-utilizes WSI data since MIL updates the model's parameters by considering only the signal from the max-pooled instance in the WSI, leading to a significant loss of context and information. Additionally, these models, along with most deep learning models, are often perceived as 'black boxes' due to the opaque nature of their decision-making processes, which involve navigating complex networks with millions of parameters. Therefore, it is crucial to enhance the interpretability of these processes or to develop tools that assist in understanding the decisions made by these models. Ensuring interpretability or providing explanatory tools is a necessary step for the successful integration of deep learning models into routine clinical workflows[60].

## 2.5 Interpretability

Interpretability in machine learning encompasses the capacity to render the operations of these computational models comprehensible to humans. This attribute holds particular significance in computational pathology, medical images are analyzed. Ensuring that the workings of these algorithms are accessible and trustworthy is imperative, given their implications for patient care.

ML systems have limitations in the formulation of problems. These limitations may arise from various sources, including gaps in scientific knowledge, safety considerations, ethical dilemmas, or conflicts between differing objectives [61]. In the medical domain, clarity regarding the rationale behind an algorithm's decision is essential to ascertain that it is operating in a manner that is safe, unbiased, and aligned with ethical norms.

In the transition from traditional machine learning to deep learning, the challenge of interpretability becomes even more pronounced due to the "black box" nature of deep learning models. Deep learning architectures, especially those used in computational pathology, such as convolutional neural networks, can achieve remarkable performance in tasks like tumor detection, classification, and prognosis. However, their complex layers and non-linear transformations make it difficult to understand how they arrive at their decisions. This opacity can hinder clinicians' and researchers' trust in these systems, despite their potential to revolutionize pathology through enhanced accuracy and efficiency.

Attention-guided weakly-supervised MIL methods have emerged as a solution that addresses two significant challenges in computational pathology: the sub-optimal usage of data inherent in traditional MIL and the pervasive lack of interpretability in deep learning models. By incorporating attention mechanisms, these methods not only enhance the data aggregation process, allowing for more effective utilization of the rich information in whole-slide images, but also improve the qualitative interpretability of the models. One approach has been the integration of attention heatmaps [62, 63]. These mechanisms aim to highlight the regions within the data that the model deems significant for making a diagnosis. For instance, in models like clustering-constrained-attention multiple-instance learning (CLAM)[17], attention mechanisms help elucidate the areas within tissue samples that are influential in the model's decision-making process. However, while these efforts in qualitative interpretability contribute valuable insights, they do not comprehensively address the transparency of these models.

In the context of enhancing deep learning models for clinical use in computational pathology, it is essential to address three intertwined aspects of model interpretability and transparency:

1. Visual Interpretability: This involves using methods like classification and attention heatmaps to highlight important areas in diagnostic images. These methods help clinicians see what the model focuses on, making it easier for them to trust and understand the model's outputs.

2. Prototypical Interpretability: Providing explainability of how the model works internally by showing previously diagnosed 'prototypes' or key examples that the model used in its decision-making. It is important for clinicians to see these prototypes to grasp why the model makes certain predictions, ensuring these are based on patterns and phenomena they recognize in their clinical practice.

3. Uncertainty Quantification: It's vital for models to do more than just identify key areas and prototypes; they must also accurately measure and communicate the confidence in their predictions. This is especially important in medical diagnostics, where it helps manage inherent complexities and ambiguities. By highlighting its own limitations and potential errors, a model can greatly assist clinicians in making better-informed decisions. This aspect of model transparency has been largely overlooked in deep learning research within the medical field.

While weakly supervised learning models offer significant advantages in computational pathology, their full clinical adoption hinges on overcoming the 'black box' nature through enhanced interpretability. This involves visual and qualitative insights and a deeper understanding of the model's internal representations through prototypical interpretability and the incorporation of uncertainty quantification. Collectively, these dimensions contribute to building robust, transparent, and clinically viable diagnostic tools.

## 2.6 Kernel Density Matrices

This thesis presents WiSDoM (see Figure **4-1**), a framework based on Kernel Density Matrices[18] to model joint probability distributions. WiSDoM is tailored as a probabilistic deep learning framework specifically for automated grading of prostate whole-slide images. It operates in a fully and weakly supervised manner, placing a strong emphasis on interpretability.

The application of KDM in medical imaging has already been proven effective in domains such as diabetic retinopathy analysis and prostate cancer tissue grading [64]. Its success stems from its unique ability to integrate the robust feature representation of deep convolutional neural networks with a differentiable probabilistic regression model. This integration enables KDM to offer a sophisticated representation of label probability distributions within an ordinal regression framework. Such a framework is particularly adept at modeling cancer progression as a continuum. A key strength of KDM lies in its ability to predict posterior probability distributions, which allows for the precise quantification of the uncertainty in its predictions.

### 2.6.1 Density Matrix

In quantum mechanics, the state of a pure quantum system is encapsulated by a wave function, denoted as $|\psi\rangle$, within a Hilbert space $H$. To determine the likelihood of the system being in a particular state, one computes the square of the magnitude of the wave function's projection onto that state [65].

On the other hand, quantum systems can also exhibit classical uncertainty, characterized by mixed states. These mixed states are conceptualized as a statistical blend of various pure states, each represented by $|\psi_i\rangle$ and associated with a distinct probability $p_i$. The set of probabilities $\{p_i\}$ adheres to the normalization condition $\sum_i^N p_i = 1$. For such mixed states, the formalism of quantum mechanics employs a density matrix, denoted by $\rho$, to represent the statistical properties of the system:

$$\rho = \sum_i^N p_i \left|\psi_i\right\rangle \left\langle\psi_i\right| \tag{2.1}$$

where $\langle\psi_i|$ represents the conjugate transpose of $|\psi_i\rangle$.The Born rule [65] can be extended to compute the probability of finding a system with the state represented by $\rho$ in a state $|\psi\rangle$ after a measurement:

$$p(|\psi\rangle \mid \rho) = \operatorname{Tr}(|\psi\rangle\langle\psi|\rho) = \langle\psi|\rho|\psi\rangle = \sum_i^N p_i \left|\langle\psi \mid \psi_i\rangle\right|^2 \tag{2.2}$$

Density matrices serve as powerful instruments for representing probability distributions and facilitating a range of computations with efficiency. They are pivotal in predicting the results of quantum measurements and in calculating expected values, among other applications [65, 18].

The concept of a Kernel Density Matrix (KDM) emerges as a specialized form of a density matrix, formulated within a Hilbert space shaped by a kernel function. KDMs excel in efficiently encapsulating joint probability distributions and are instrumental in tasks like inference, generation, and sampling. A key attribute of KDM is the differentiability of their internal operations, which allows for seamless integration into deep learning frameworks.

The formal definition of KDM is as follows, as originally stated by González et al [66]:

**Definition 1** (Kernel Density Matrix)**.** *A Kernel Density Matrix over a set $\mathbb{X}$ is a triplet $\rho = (\boldsymbol{C}, \boldsymbol{p}, k_\theta)$ where $\boldsymbol{C} = \left\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\right\} \subseteq \mathbb{X}, \boldsymbol{p} = (p_1, \ldots, p_m) \in \mathbb{R}^m$ and $k_\theta : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, such that $\forall \boldsymbol{x} \in \mathbb{X}, k(\boldsymbol{x}, \boldsymbol{x}) = 1, \forall i p_i \geq 0$ and $\sum_{i=1}^{n} p_i = 1$*

The elements of $C$ are the components of the KDM, and the $p_i$ value represents the mixture weight, or probability, of the component $\boldsymbol{x}_i$. If $\phi : \mathbb{R}^n \to \mathcal{H}$ is the mapping to the RKHS $\mathcal{H}$ associated to the kernel $k_\theta, \rho$ represents a density matrix defined as in 2.1 with components $|\psi_i\rangle = \left|\phi\left(\boldsymbol{x}^{(i)}\right)\right\rangle$. The projection function associated to a KDM $\rho$ is defined as:

$$f_\rho(\boldsymbol{x}) = \sum_{\boldsymbol{x}^{(i)} \in \boldsymbol{C}} p_i k_\theta^2 \left(\boldsymbol{x}, \boldsymbol{x}^{(i)}\right) \tag{2.3}$$

The projection function in 2.3 can be transformed into a probability density function (PDF) by multiplying it by a normalization constant that depends on the kernel of the KDM:

$$\hat{f}_\rho(\boldsymbol{x}) = \mathcal{M}_k f_\rho(\boldsymbol{x}) \tag{2.4}$$

## 2.6.2 Inference with Kernel Density Matrices

Inference is the process of deducing unknown output variables from known input variables and the parameters of a model. This process can be approached probabilistically, where the relationship between inputs and outputs is expressed through a probability distribution, denoted as $p(\mathbf{x}', \mathbf{y}')$. This distribution reflects the inherent uncertainty in the data generation process. In making predictions about output variables, it is crucial to consider and integrate both types of uncertainty into the output probability distribution, $p(\mathbf{y})$. Inference transforms the probability distribution of input variables, $p(\mathbf{x})$, into a distribution for the output variables, $p(\mathbf{y})$, by leveraging the joint probability distribution of inputs and outputs, $p(\mathbf{x}', \mathbf{y}')$. In the context of KDM, these probability distributions are represented as follows:

$$\rho_{\mathbf{x}} = \left(\left\{\boldsymbol{x}^{(i)}\right\}_{i=1\ldots m}, (p_i)_{i=1\ldots m}, k_{\mathbb{X}}\right) \tag{2.5}$$

$$\rho_{\mathbf{x}', \mathbf{y}'} = \left(\left\{\left(\boldsymbol{x}'^{(i)}, \boldsymbol{y}'^{(i)}\right)\right\}_{i=1\ldots m'}, (p'_i)_{i=1\ldots m'}, k_{\mathbb{X}} \otimes k_{\mathbb{Y}}\right) \tag{2.6}$$

$$\rho_{\mathbf{y}} = \left( \left\{ \boldsymbol{y}'^{(i)} \right\}_{i=1\ldots m'}, (p''_i)_{i=1\ldots m'}, k_{\mathbb{Y}} \right) \tag{2.7}$$

The probabilities of $\rho_y$ after the inference procedure are given by the following expression:

$$p''_i = \sum_{\ell=1}^{m} \frac{p_\ell p'_i \left( k_x \left( \boldsymbol{x}^{(\ell)}, \boldsymbol{x}'^{(i)} \right) \right)^2}{\sum_{j=1}^{m'} p'_j \left( k_x \left( \boldsymbol{x}^{(\ell)}, \boldsymbol{x}'^{(j)} \right) \right)^2}, \text{ for } i = 1 \ldots m' \tag{2.8}$$

This inference procedure uses a kernel function to assign local weights to each output training sample according to the input sample's similarity with each learned prototype $\boldsymbol{x}'^{(i)}$ in the KDM $\rho_{\mathbf{x}',\mathbf{y}'}$. We obtain a full probability distribution represented as an output KDM $\rho_y$.

In quantum mechanics, the density matrix $\rho_{\mathbf{x}',\mathbf{y}'}$ represents the state of a bipartite system. The inference process in this context is a measurement operation, where the $\mathbf{x}$ subsystem collapses, resulting in a state represented by $\rho_{\mathbf{x}}$. As a consequence of this collapse, the state of the $\mathbf{y}$ subsystem is altered and becomes $\rho_{\mathbf{y}}$ [66].

The parameters of the inference model correspond to the parameters of the KDM $\rho_{\mathbf{x}',\mathbf{y}'}$. These parameters can be estimated in a non-parametric way which does not scale well to large datasets, discriminative learning by performing gradient-based optimization minimizing a suitable loss function like cross-entropy loss or mean square error depending on the output variable type and task, and maximum likelihood learning which estimates the parameters by maximizing the probability density of the training dataset assigned by $\rho_{\mathbf{x}',\mathbf{y}'}$.

For training, set $C_{x'y'}$ is initialized with a randomly selected subsample from the training dataset $D$. Each pair $(\mathbf{x}'^{(i)}, y'^{(i)})$ in $C_{x'y'}$ is selected to represent the initial state of KDM. The algorithm proceeds by assigning an initial equal probability $p_{x'y'}$ to each component in $C_{x'y'}$. Subsequently, a KDM $\rho^{(i)}_{x'y'}$ is constructed for each sample, integrating the kernel functions $k_{X,\sigma}$ and $k_Y$ through tensor product operations. Iterative optimization is carried out through a gradient-based method, where the objective is to minimize the loss function $\mathcal{L}$ with respect to the predicted label distributions $\rho^{(i)}_Y$ and the actual labels $y^{(i)}$. This process continues until a predefined stopping criterion is met, resulting in a KDM $\rho_{\mathbf{x}',\mathbf{y}'}$ that encapsulate the learned distributional characteristics of the data, allowing for the predictive modeling of labels within an ordinal regression framework and quantifying the uncertainty in predictions. A graphical representation of KDM $\rho_{\mathbf{x}',\mathbf{y}'}$ and the inference procedure is shown in figure **2-2**.

The key strength of KDM lies in its ability to provide uncertainty quantification and model explainability. In this study, KDM will be extended to tackle two main tasks: prostate tissue Gleason grading and whole-slide ISUP grading.

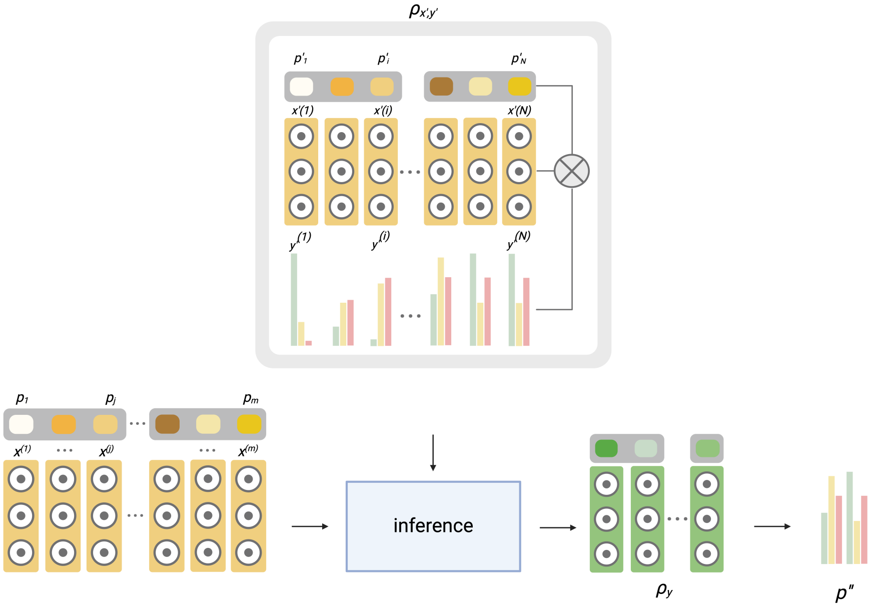

**Figure 2-2**: Graphical representation of KDM $\rho_{\mathbf{x}',\mathbf{y}'}$ containing a set of prototypes or learned representations of the input data $\boldsymbol{x}'$, a label distribution $\boldsymbol{y}'$, and a set of probabilities $p'$ correlating to the importance or probability of each learned representation and the inference procedure involving input KDM $\rho_x$, and joint KDM $\rho_{\mathbf{x}',\mathbf{y}'}$ yielding a KDM $\rho_y$ from which label probabilities $p''$ can be calculated.

# 3 Fully-supervised WiSDoM

In this chapter, the introduction of WiSDoM is followed by an in-depth description of its training and inference pipeline, applicable in fully-supervised learning environments. The fully-supervised task operates similarly to conventional deep learning approaches, where a deep neural network is used as a feature extractor, and then a classification is made based on those features. However, WiSDoM is able to provide uncertainty quantification by treating the task as an ordinal regression problem where expected value and variance are obtained for every prediction as a measure of uncertainty, while modeling cancer progression as a continuum. Additionally, learned internal parameters of the KDM within WiSDoM allow prototypes to be obtained as examples of the internal learned representations, providing an additional layer of interpretability for clinical trust. The discussion also includes an exploration of the uncertainty quantification and interpretability features.

## 3.1 Method

We investigate the application of WiSDoM for Gleason pattern grading in a fully-supervised setting, a critical task in the diagnostic process of prostate cancer. Our primary focus is on the use of WiSDoM to effectively distinguish among various Gleason patterns depicted in histopathological images of prostate tissue. These patterns constitute a vital determinant in assessing the severity and progression of prostate cancer.

While WiSDoM in a fully-supervised task works similarly to a traditional deep learning model, we aim to provide additional layers of trust and interpretability to predictions. Our model aims to replicate the diagnosis procedure employed by pathologists by visually quantifying the extent of each pattern. The purpose extends beyond the classification of these patterns; it attempts to accurately quantify their prevalence. It is noteworthy that this approach mirrors the visual analysis performed by pathologists in examining tissue samples, which ultimately yields a Gleason score. We create an automated process that closely reflects a real-world clinical setting, thereby enhancing interpretability.

WiSDoM encodes WSI patches into 128-dimensional feature vector representations $x \in \mathbb{R}^{128}$, using a deep learning model as the backbone. First, patch feature vectors are represented as a KDM $\rho_{\mathbf{x}}$ with $m = 1$ components and $p' = 1$, KDM $\rho_{\mathbf{x}',\mathbf{y}'}$ is initialized by an arbitrary set of encoded patch-label pairs $C_{x'y'}$ from training dataset $D$ (see figure **3-1**).

The inference procedure involves using the KDM $\rho_{\mathbf{x}',\mathbf{y}'}$ and input KDM $\rho_{\mathbf{x}}$, and performing an inference operation (see eq. 2.8). The resulting KDM $\rho_y$ contains a discrete probability

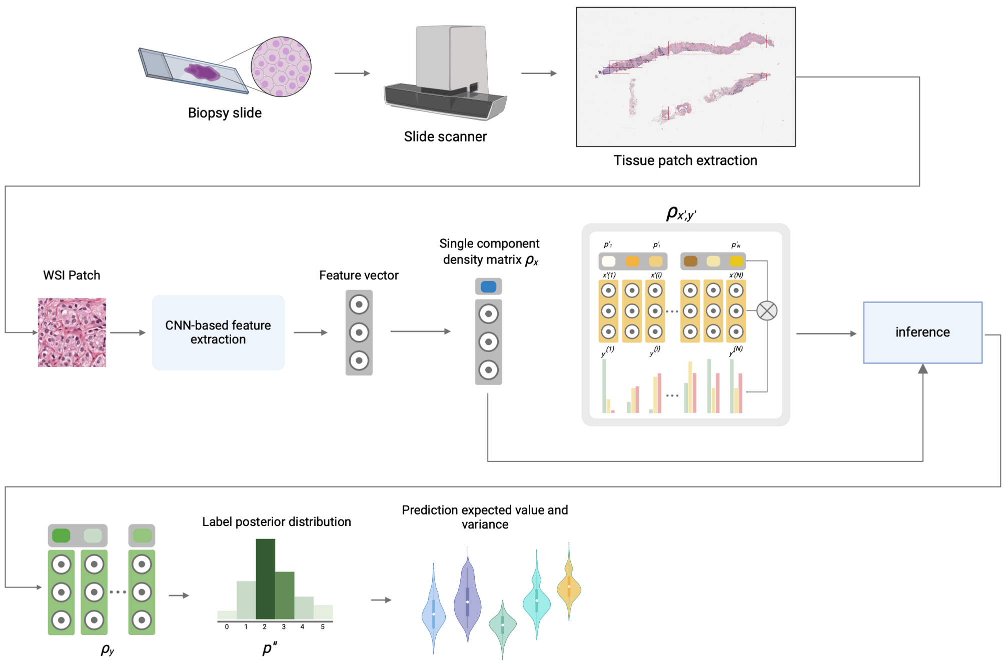

**Figure 3-1**: **Fully supervised WiSDoM architecture** The process begins with extracting patch bags from the WSI. These bags are encoded into a feature space by a CNN (Convolutional Neural Network). Every patch feature vector is modeled as a single component density matrix. An output distribution of labels is then derived from the joint KDM $\rho_{\mathbf{x}',\mathbf{y}'}$ of weighted prototypes and their labels obtained after the training process. The output includes a patch-level label posterior distribution $p''$, from which an expected value and variance can be computed.

distribution of output labels $p'' = (p''_1, p''_2, \ldots, p''_n)$, where each $p''_i$, represents the probability associated with the $i$-th label. When performing a classification task, we select the most probable label from the distribution. When performing ordinal regression, we slightly modify the inference procedure: First, we convert categorical labels to continuous labels in the range $[0, 1]$, the conversion operation from a categorical label to an ordinal label is simply achieved by normalizing the categorical label of each sample by the total number of possible labels as follows (Eq. 3.1).

$$y_{\text{ordinal}} = \frac{y_{\text{categorical}}}{N_{\text{labels}}} \tag{3.1}$$

From the probability distribution obtained from $\rho_{\mathbf{y}}$ we can calculate the expected value and variance.
Given a density matrix $\rho_{\mathbf{y}}$, represented by a vector $p'' = (p''_1, p''_2, \ldots, p''_n)$, where each $p''_i$ represents the probability associated with the $i$-th label, the expected value and variance are computed as follows:

A weighted sum of the probabilities and the values associated with each label. The values associated with each label are evenly distributed between 0 and 1. Therefore, the expected value ($E[\hat{y}]_i$) is calculated as follows:

$$E[\hat{y}]_i = \sum_{j=1}^{m} p''_{ij} \cdot y'_j \tag{3.2}$$

Where ${y'_j}_{j=1\ldots m}$ are the values associated with each label.
The variance is calculated as the expected value of the squares minus the square of the expected value. This is given by:

$$Var[\hat{y}]_i = E[\hat{y}^2]_i - (E[\hat{y}]_i)^2 \tag{3.3}$$

where $E[\hat{y}^2]_i$ is calculated similarly to $E[\hat{y}]$ but using the square of the values ($y'^2_j$):

$$E[\hat{y}^2] = \sum_{i=j}^{m} p''_{ij} \cdot y'^2_j \tag{3.4}$$

The expected value and variance are then output for each input patch. Algorithms 1 2 summarize the training and prediction procedure of fully-supervised WiSDoM.

## 3.2 Experimental Design

This section presents the experimental design of WiSDoM in the fully-supervised task. It begins with a description of the dataset characteristics, detailing the specific splits used for training, validation, and testing. This is followed by an overview of the performance measures adopted for assessing WiSDoM effectiveness. We then discuss the baselines selected for comparative analysis.

**Algorithm 1:** Fully-supervised WiSDoM Training Algorithm

```
Input:
$D = \{(x_i, y_i)\}_{i=1\ldots N}$: Training dataset.
$m$: number of components (encoded set of training patches) of KDM $\rho_{\mathbf{x}',\mathbf{y}'}$
$Z$: Deep learning backbone
  1. KDM $\rho_{\mathbf{x}',\mathbf{y}'}$ is initialized with a sample of size $m$ from dataset $D$
  2. for each $(x_i, y_i) \in D$:
         $\rho_y = (\{y'_i\}_{i=1\ldots m}, (p''_i)_{i=1\ldots m}, K_Y) = \text{predict}(x_i, \rho_{\mathbf{x}',\mathbf{y}'}, Z)$ (see algorithm 2)
  3. If task = classification:
       a) $\hat{y} = y'_{\arg\max(p'')}$
       b) Minimize $L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$
  4. If task = ordinal regression:
       a) Calculate $E[\hat{y}]_i = \sum_{j=1}^{m} p''_{ij} \cdot y'_j$ and $Var[\hat{y}]_i = E[\hat{y}^2]_i - (E[\hat{y}]_i)^2$, where
          $E[\hat{y}^2]_i = \sum_{j=1}^{m} p''_{ij} \cdot y'^2_j$
       b) Minimize $L = \frac{1}{N} \sum_{i=1}^{N} (E[\hat{y}]_i - y_i)^2 + \alpha \cdot Var[\hat{y}]_i$, where $\alpha$ is a penalization
          parameter for variance.
  5. Update all backbone weights $\mathbf{w}$, and KDM $\rho_{\mathbf{x}',\mathbf{y}'}$ parameters using gradient descent.
  6. Return $(Z, \rho_{\mathbf{x}',\mathbf{y}'})$
```

**Algorithm 2:** Fully-supervised WiSDoM prediction procedure

```
Input:
$\rho_{\mathbf{x}',\mathbf{y}'} = \{(x'_i, y'_i)\}_{i=1\ldots m}, (p_i)_{i=1\ldots m}, k_{\mathbb{X}} \otimes k_{\mathbb{Y}}$: joint KDM
$x$: input patch
$Z$: Deep Learning backbone
  1. Encode patch $x$ using $Z$: $z = Z(x)$
  2. Create $\rho_x = (\{z\}, (1), k_{\mathbb{X}})$
  3. Calculate probabilities $p''$ for output KDM $\rho_y$ using $\rho_x$ and $\rho_{\mathbf{x}',\mathbf{y}'}$ with eq. 2.8
  4. $\rho_y = (\{y'_i\}_{i=1\ldots m}, (p''_i)_{i=1\ldots m}, K_Y)$
  5. Return $\rho_y$
```

**Table 3-1**: Patch dataset distribution, G = Gleason grade

| Dataset | Total | Stroma | Healthy | G 3 | G 4 | G 5 |
|---|---|---|---|---|---|---|
| Training set | 1’039.873 | 655.467 | 85.354 | 105.741 | 160.868 | 32.443 |
| Validation set | 343.114 | 217.987 | 28.308 | 33.350 | 54.516 | 8.953 |
| Test set | 344.472 | 218.870 | 28.703 | 36.224 | 52.826 | 7.849 |

### 3.2.1 Dataset

The effectiveness of WiSDoM at the patch level was assessed using the Prostate Cancer Grade Assessment (PANDA) dataset[1], one of the largest publicly available collections of prostate WSI. This dataset comprises 10,600 digital whole-slide images of H&E-stained biopsies from Radboud University Medical Center (D1) and Karolinska Institutet (D2).

The data provided by the two centers vary. Both D1 and D2 include WSI-level diagnostics for both Gleason and ISUP grades, along with annotation masks indicating the presence of Gleason patterns on the slides. However, not all WSIs are annotated. D2 includes regions annotated as cancerous and non-cancerous, whereas D1 provides detailed pixel-level annotations for Gleason patterns 3, 4, 5, stroma, and healthy tissue.

Furthermore, the centers employed different data acquisition methods. D1 scanned all slides at a 20x magnification (pixel resolution 0.24 $\mu m$) using a 3DHistech Pannoramic Flash II 250 scanner. In contrast, D2 used a 20x magnification as well, but with pixel resolutions of 0.45202 $\mu m$ and 0.5032 $\mu m$ for slides scanned with Hamamatsu C9600-12 and Aperio ScanScope AT2, respectively.

The annotation process for both centers is subject to variability due to the subjective nature of Gleason grading. Annotations in D1 were determined by a consensus of medical students experienced in pathology, while D2 relied on a single expert pathologist for annotations.

For our study, the data from D1 was divided into training, validation, and testing sets to assess the model’s performance in patch-level Gleason grading. Patch extraction was exclusively conducted on data from D1 due to the lack of tissue annotations in D2. In D1, patches are labeled into specific Gleason grades based on the predominant tissue type in the corresponding annotation masks, with a patch assigned to a particular class if it contains over 25% of that tissue type. The split of patches was consistently maintained at the slide level to avoid information leakage. The specifics of the patch dataset are detailed in Table **3-1**. It is important to note that the patch data originates solely from D1.

### 3.2.2 Performance Measures

The efficacy of WiSDoM in both classification and ordinal regression tasks is assessed using two metrics: Cohen’s Kappa ($\kappa$, see equation 3.5), Accuracy (see equation: 3.6), and Mean Absolute Error (MAE, see equation: 3.7). Cohen’s Kappa ($\kappa$) is a statistical metric that quantifies the agreement between two outcomes. This metric generally ranges from 0, indicating random agreement, to 1, denoting complete agreement. We chose this metric for its relevance in measuring inter-observer agreement with the Prostate Cancer Grade Assessment (PANDA) reference standard. Additionally, since the $\kappa$ is the primary metric used in the PANDA challenge, it allows for a direct and coherent assessment of WiSDoM performance compared to established benchmarks [1]. The MAE, on the other hand, provides a direct measure of the average magnitude of errors in WiSDoM predictions.

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}} \tag{3.5}$$

where $O_{i,j}$ is the observed agreement between the model's predictions and the actual ground truth, $E_{i,j}$ is the expected agreement by chance, and $w_{i,j}$ is the quadratic weight given to the disagreement between prediction $i$ and ground truth $j$. The $\kappa$ metric ranges between 0 and 1, where 1 signifies perfect agreement, providing a measure of the model's agreement with the ground truth.

Accuracy is defined as a metric to assess the ratio of correct predictions by the model:

$$\mathit{Accuracy} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(\hat{y}_i = y_i) \tag{3.6}$$

Where $n$ is the total number of samples, $\hat{y}_i$ is the predicted label for the $i$-th sample, $y_i$ is the ground truth label for the $i$-th sample, and $\mathbf{1}(\cdot)$ is the indicator function, which is 1 if $\hat{y}_i = y_i$ and 0 otherwise. This calculates the proportion of correct predictions over all predictions made.

The MAE is defined as:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3.7}$$

where $y_i$ is the true value, $\hat{y}_i$ is the predicted value, and $n$ is the total number of observations. MAE provides a straightforward measure of how close the predictions are to the actual values, irrespective of direction, a suitable metric for regression problems.

### 3.2.3 Baseline

Since, to the best of our knowledge, no work has been published on patch Gleason grading with the PANDA dataset, we use as a baseline a traditional CNN trained on the same problem as WiSDoM on a fully-supervised fashion (see table **3-2**.)

### 3.2.4 Training Details

For patch-level, fully-supervised training, we process each WSI to extract all patches that contain tissue. The selection of patches for training is guided by an automated tissue detection algorithm that identifies regions within the slide that contain significant tissue content. To ensure the relevance and quality of the training data, we exclusively select patches that exhibit more than 25% of a specific Gleason pattern, as determined by the annotation masks provided in the Radboud UMC cohort (D1). Patches are extracted at a 20x magnification level, with each patch sized at 192x192 pixels.

The encoding of patches into a latent feature space is accomplished using an ImageNet pretrained ConvNeXT[67] model $Z$. We employ ConvNeXT due to its position as a state-of-the-art convolutional neural network. It balances complexity and parameter count, providing high-quality feature representations. This allows for efficient representation acquisition without the substantial computational power demand associated with Vision Transformers. Before training, the model is subjected to a preliminary warm-up phase. During this phase, lasting for 2 epochs, the encoder processes individual patches in a straightforward classification scenario to adjust its weights for optimal performance in the patch-level context.

For the initialization of the KDM $\rho_{\mathbf{x}',\mathbf{y}'}$, which requires establishing a set of patch-label pairs $C_{x',y'}$ and corresponding importance weights $p'$, we compile a balanced collection of patch-label pairs from the training dataset. These pairs, after being processed through the encoder to transform them into latent space representations, serve as the basis for initializing $\mathbf{x}'$ and $\mathbf{y}'$ within the KDM $\rho_{\mathbf{x}',\mathbf{y}'}$. Lacking initial importance weights, $p'$ is uniformly set, assigning equal significance $p'_i = 1/m$ to all $m$ pairs at the start.

Hyperparameter tuning led to the selection of 216 initialization pairs or 'prototypes', making sure that each possible class is well represented in this initial set of pairs. Following warmup and initialization, training is started. Optimization is carried out using the Adam optimizer, with a learning rate of 0.0001 and parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We used a mini-batch size of 36, with a gradual warm-up scheduler applied for the first epoch followed by cosine annealing for the subsequent epochs. Categorical cross-entropy loss is used for classification, training is conducted over 50 epochs with early stopping after five epochs without improvement in validation loss. For ordinal regression, the setup is similar, using continuous labels between 0 and 1. The loss function is Mean Squared Error with a variance penalty ($\alpha$), maintaining consistent optimizer and batch size settings as in classification.

## 3.3 Results

WiSDoM demonstrated notable performance in the five-class Gleason pattern classification task, achieving a $\kappa$ of 0.896 and an accuracy of 0.901. This task encompassed the classification of stromal, benign epithelium, and Gleason grades 3, 4, and 5 patches. When the grading was framed as an ordinal regression task, WiSDoM achieved a $\kappa$ of 0.906, an accuracy of 0.890, and a MAE of 0.13. These results highlight the model's adaptability and effectiveness in addressing the grading challenge, both as a multi-class classification problem and as an ordinal regression task. The latter approach is particularly significant as it aligns with the clinical understanding of Gleason grades, representing them as a continuum

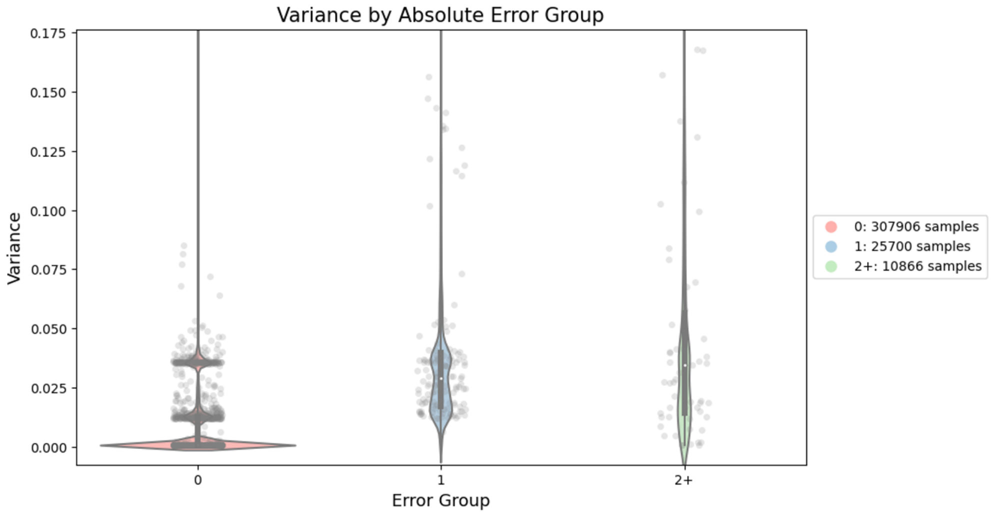

**Figure 3-2**: Violin plots of predictions variance grouped by absolute error, as seen in the plot, correct predictions have less variance while error 1 and 2 groups tend to have larger variance.

indicative of cancer progression.

### Uncertainty Quantification

A key feature of the ordinal regression model is its ability to quantify uncertainty. Each prediction generated by the model includes a confidence level, expressed as variance. This aspect of the model is crucial as it allows for the exclusion of predictions with high levels of uncertainty, enhancing the model's precision and reliability. Through the analysis of various variance thresholds, it was observed that prioritizing predictions with lower variance ($\sigma^2 < 0.05$) improved the model's inter-observer agreement and accuracy. Specifically, under these parameters, the model attained a $\kappa$ of 0.924, an accuracy of 0.910, and maintained an MAE of 0.13. This relationship between prediction accuracy and variance is visually represented in Figure **3-2**, which displays violin plots of test set predictions. The plots are categorized according to absolute error and variance values, demonstrating that correct predictions typically align with lower variance. Predictions off by an error of 1 (for example, predicting Gleason score 3 when the actual score was 2 or 4) exhibited higher variance values. Predictions with an error of 2 or more (e.g., predicting a Gleason score of 3 when the true classification was Benign or Gleason score 5) displayed even greater variance.

### Interpretability

WiSDoM delivers fine-grained heatmaps (Figure **4-3** presents various examples with varied ISUP grade, including an instance of a healthy biopsy. In the case of the healthy biopsy, the model is expected not to highlight any regions but healthy epithelium, illustrating its

**Table 3-2**: Performance Metrics for fully-supervised Gleason pattern grading

| **WiSDoM** | $\kappa$ | **Accuracy** | **MAE** |
|---|---|---|---|
| Five-Class Classification | 0.896 | 0.901 | - |
| Ordinal Regression | 0.906 | 0.890 | 0.13 |
| Ordinal Regression with Variance Threshold ($\sigma^2 < 0.05$) | **0.924** | **0.910** | **0.13** |
| Baseline (EfficientNet-v2[68]) | 0.892 | 0.899 | 0.158 |

ability to discern between pathological and non-pathological samples accurately.) that reveal near-pixel-level detail of the Gleason patterns identified across the whole slide. This feature is particularly valuable as a visual aid for pathologists in quantifying the extent of Gleason patterns, a fundamental aspect of the Gleason grading protocol. The heatmaps serve as a supplementary tool, aiding pathologists in identifying potential blind spots and providing a secondary perspective on regions that may be difficult to diagnose.

### Gleason pattern area quantification

Following the results of WiSDoM in Gleason pattern classification and ordinal regression tasks, we expanded our analysis to encompass an element of pathologist-performed prostate cancer (PCa) grading: the quantification of Gleason pattern areas. Traditionally, pathologists grade prostate cancer by identifying and measuring the extent of different Gleason patterns present in tissue samples, with the Gleason score being determined from the two most prevalent patterns.

Employing WiSDoM patch-level classification capabilities on WSI, we were able to replicate this traditional method of area quantification. This approach not only allows for a direct comparison between the performance of our model and conventional pathological methods but also introduces an additional dimension of interpretability to the model's outputs.

We conducted a detailed comparison between the area estimations generated by WiSDoM and those derived from pathologist annotations. This analysis was stratified according to ISUP Grade Group. It involved calculating the average difference in area quantification between pathologist annotations and the model's predictions at the patch level for all slides within each ISUP grade group. The MAE was employed as the primary metric for this assessment, offering a quantifiable measure of the model's accuracy in area quantification when compared to the evaluations made by pathologists. The MAE values corresponding to each ISUP Grade Group, which are detailed in table **3-3**, provide insights into the extent to which the model's Gleason pattern area quantifications are in concordance with those of experienced pathologists.

**Table 3-3**: Quantification of average Gleason pattern extension in whole-slides grouped by ISUP grade group. True extension is measured from each healthy epithelium, stroma, and Gleason pattern available tissue annotations. Predicted extension is calculated by performing inference at a patch level with a high overlapping ratio of patches. The difference in extend is then quantified using MAE per tissue pattern and then averaged across all slides in each ISUP grade group.

| **ISUP Grade Group** | **Gleason Pattern** | **True extension %** | **Predicted %** | **MAE** |
|---|---|---|---|---|
| **GG 0** | Stroma | 78.0 | 76.5 | 0.0059 |
| | Healthy Epithelium | 22.0 | 22.8 | |
| | Gleason 3 | 0.0 | 0.04 | |
| | Gleason 4 | 0.0 | 0.02 | |
| | Gleason 5 | 0.0 | 0.01 | |
| **GG 1** | Stroma | 64.4 | 62.3 | 0.0131 |
| | Healthy Epithelium | 10.7 | 13.4 | |
| | Gleason 3 | 24.9 | 23.7 | |
| | Gleason 4 | 0.0 | 0.05 | |
| | Gleason 5 | 0.0 | 0.0 | |
| **GG 2** | Stroma | 55.6 | 51.9 | 0.0171 |
| | Healthy Epithelium | 5.4 | 7.3 | |
| | Gleason 3 | 27.1 | 29.1 | |
| | Gleason 4 | 11.9 | 11.3 | |
| | Gleason 5 | 0.0 | 0.04 | |
| **GG3** | Stroma | 57.2 | 54.0 | 0.0165 |
| | Healthy Epithelium | 5.7 | 7.4 | |
| | Gleason 3 | 9.7 | 11.2 | |
| | Gleason 4 | 27.4 | 26.5 | |
| | Gleason 5 | 0.0 | 0.09 | |
| **GG 4** | Stroma | 62.8 | 60.6 | 0.0117 |
| | Healthy Epithelium | 5.1 | 6.1 | |
| | Gleason 3 | 2.6 | 3.5 | |
| | Gleason 4 | 25.9 | 25.2 | |
| | Gleason 5 | 3.5 | 4.7 | |
| **GG 5** | Stroma | 62.8 | 59.3 | 0.0174 |
| | Healthy Epithelium | 3.0 | 3.4 | |
| | Gleason 3 | 0.0 | 0.08 | |
| | Gleason 4 | 21.6 | 20.7 | |
| | Gleason 5 | 12.6 | 15.8 | |
| | **Overall** | | | 0.0136 |

# 4 Weakly-supervised WiSDoM

WiSDoM is a probabilistic deep learning framework tailored for weakly supervised classification and ordinal regression tasks in computational pathology that integrates deep learning, KDM, and local-global attention. On a weakly supervised task, WiSDoM operates on the principle that each WSI in the training set is an individual data point with an established slide-level diagnosis yet lacks specific pixel or region-level annotations, we then view each WSI as a collection of numerous smaller segments or patches, similar to MIL. Traditionally, MIL focuses on binary classification, discerning positive from negative classes under the assumption that the presence of one positive patch classifies the entire slide as positive. This approach typically employs a max-pooling aggregation function, choosing the patch with the highest probability of the positive class for slide-level classification. However, this method is not suitable for multiclass classification or binary classification without clear positive/negative annotations.

WiSDoM differentiates itself by not using the standard max-pooling or other conventional aggregation functions like average pooling, generalized mean, or log-sum-exp, which are limited in terms of problem-specific adaptability and interpretability. Instead, WiSDoM integrates an attention-guided KDM for aggregating information from patches. This allows for a better integration of patch-level data into a unified WSI prediction or representation, offering enhanced interpretability and adaptability for various classification problems.

## 4.1 Method

Our prior approach relied on tissue annotations to select relevant patches and to gather Gleason pattern labels for patches across an entire slide. This strategy facilitated an interpretable method for quantifying the extent of each Gleason pattern on the whole slide, similar to how a pathologist would conduct their diagnosis, all centered around WiSDoM classifying patches into specific Gleason patterns.

However, given the high cost and relative unavailability of tissue annotation masks in real-world clinical scenarios, we aim to eliminate the necessity for tissue annotations during training. In this section, we propose a novel method that requires only a whole-slide diagnosis, typically an ISUP grade group for prostate biopsies, which can be easily obtained from pathology reports.

Competitive performance can be achieved in line with current state-of-the-art methodologies for whole-slide grading, which only needs a weak label for training while being constrained by providing interpretability.

We extend WiSDoM probabilistic deep learning framework for weakly supervised, interpretable ordinal regression and classification. It operates on the principle that each WSI in the training set is an individual data point with an established slide-level diagnosis yet lacks specific pixel or region-level annotations. The framework adopts a similar approach to MIL, viewing each WSI as a collection of numerous smaller segments or patches (see figure **4-1**).

Traditionally, MIL focuses on binary classification, discerning positive from negative classes under the assumption that the presence of one positive patch classifies the entire slide as positive. This approach typically employs a max-pooling aggregation function, choosing the patch with the highest probability of the positive class for slide-level classification. However, this method is unsuitable for multiclass or binary classifications without explicit positive/negative annotations.

WiSDoM differentiates itself by not using the standard max-pooling or other conventional aggregation functions like average pooling, generalized mean, or log-sum-exp, which are limited in terms of problem-specific adaptability and interpretability. Instead, WiSDoM integrates an attention-guided KDM for aggregating information from patches. This method allows for a more nuanced integration of patch-level data into a unified WSI prediction or representation, offering enhanced interpretability and adaptability for various classification problems.

Following tissue detection and patch extraction, WiSDoM involves encoding the $N$ patches constituting a WSI into a feature vector representation $x_n \in \mathbb{R}^{128}$ utilizing a deep learning backbone.

In our study on ISUP grading, we adopt a novel approach by using a collection of sample instances, known as 'bags,' instead of labeling each sample individually. This method is particularly suited to scenarios where patches from a whole slide collectively form a specific ISUP grade group, but individual Gleason patterns at the patch level remain unknown, a common occurrence in real-world settings.

We interpret a specific collection of patches from a WSI as a 'bag.' The challenge for the model is to learn to assign accurate labels to each patch within these bags and then synthesize this information to make a comprehensive prediction at the whole-slide level. This approach inherently involves uncertainties, especially regarding the individual characteristics of each patch within a bag. Our objective is to model these uncertainties effectively. This integration allows for a more accurate and reliable prediction process, closely mirroring the complexities encountered in actual pathological assessments.

During the training process, WiSDoM takes bags of training samples $\boldsymbol{X}^{(i)} = \left(\boldsymbol{x}^{(i)j}\right)_{j=1\dots m_i}$. The training dataset corresponds to a set of pairs $\boldsymbol{D} = \left(\boldsymbol{X}^{(i)}, \boldsymbol{y}^{(i)}\right)_{i=1\dots\ell}$, where each $\boldsymbol{y}^{(i)}$ is a vector expressing the label proportions of the i-th sample. Each input sample is represented by a KDM $\rho_x$ with $m_i$ components. For our specific problem, where the goal is to obtain a whole-slide ISUP grade group from a patch bag, we model a variant of the original implemen-

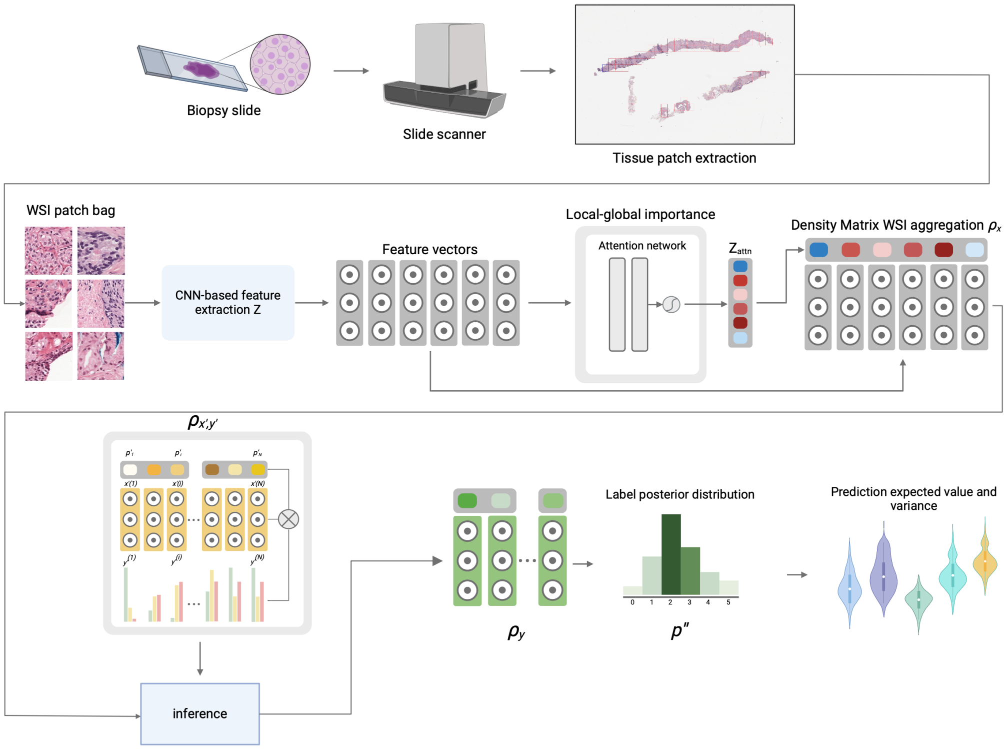

**Figure 4-1**: **Weakly supervised WiSDoM architecture.** The process begins with extracting patch bags from the WSI. These bags are encoded into a feature space by a CNN. The patch feature vectors are then processed through an attention network. This step aggregates local and global information, thereby weighing the importance of each patch. An inference operation is then performed using KDM $\rho_{\mathbf{x}',\mathbf{y}'}$ and $\rho_x$. An output distribution of labels $p''$ is then derived from the joint probability distribution of weighted prototypes and their labels. The output includes a whole slide-level label posterior distribution, from which an expected value and variance can be computed.

tation of KDM[18]. It receives training sample bags as a set of pairs $\boldsymbol{D} = \left(\boldsymbol{X}^{(i)}, \boldsymbol{y}^{(i)}\right)_{i=1\ldots\ell}$, where each $\boldsymbol{y}^{(i)}$ is the ISUP grade group of the bag, perceived as the whole-slide 'weak' label. Furthermore, considering the density matrix representation inherent to the KDM, which ascribes a probability to each possible label, we can model the significance or contribution of each instance within a bag towards the overall bag's label. To accomplish this, we employ a local-global attention method, as shown in [69]. This method assigns a weight, or a contribution factor, to each instance within the bag. Its application to natural images has proven to be useful, as it not only enhances performance but is also able to pinpoint regions of interest (ROIs), providing an additional layer of interpretability. This contribution of each patch to the bag class can be modeled into the probability $p_i$ of each KDM $\rho_x$ component $\boldsymbol{x}^{(i)j}$. By incorporating this additional information, we enhance the weakly-supervised learning process by compelling the model to assign greater importance to certain instances within the bags over others.

This attention module receives patches from a bag and processes it using two multi-layer perceptrons (MLPs), which form the means to extract attention weights from these patch bags. The initial MLP is charged with computing the local context, which essentially encapsulates the local information available in each patch $\mathbf{x}_j$. This is accomplished by passing the input through the first MLP, defined as $MLP_1$, which yields $\mathbf{z}_j^{\text{local}}$.

$$\mathbf{z}_j^{\text{local}} = \text{MLP}_1(\mathbf{x}_j) \tag{4.1}$$

Subsequently, a global context is obtained by aggregating the local context across all patches.

$$\mathbf{z}^{\text{global}} = \frac{1}{k}\sum_{j=1}^{k} \mathbf{z}_j^{\text{local}} \tag{4.2}$$

Where $k$ is the number of patches in the bag. This step provides understanding of the information present in the input data and forms the basis for the subsequent attention distribution. The local ($\mathbf{z}_j^{\text{local}}$) and global ($\mathbf{z}^{\text{global}}$) information are then combined, and this representation of both local and global information, are local-global embeddings that are fed to the second MLP $MLP_2$, yielding another set of weights, $\mathbf{z}$ which are the importance of each patch in the bag. The raw attention weights, $\mathbf{z}$, are then passed through a Softmax operation.

$$\mathbf{z}_j^{\text{attn}} = \text{Softmax}(\text{MLP}_2((\mathbf{z}_j^{\text{local}}, \mathbf{z}_{\text{global}}))) \tag{4.3}$$

The final attention weights are $\mathbf{z}_j^{\text{attn}}$ for each patch in the bag. This Softmax operation normalizes these weights such that they all lie between 0 and 1 and their total sum equals 1. The application of this operation allows the model to weigh each patch based on both the unique contribution of each patch and the global context, enhancing the model's performance by considering both individual and collective factors.

The primary differentiation in this approach, in comparison to the previous experiment, resides in the KDM $\rho_x$ creation process. Instead of uniformly distributing weights by assigning $\frac{1}{m_i}$ to every patch in the bag, where $m_i$ is the total number of patches, this novel approach

utilizes the attention mechanism to determine these weights. This inclusion allows for a more informative weight assignment that takes into account both individual patch contributions and their collective influence: we assign $p = \mathbf{z}^{\text{attn}}$ in $\rho_x$. Each MLP is configured with 64 neurons and uses a ReLU activation function. This configuration, with the number of neurons being half of the input's dimension, is chosen based on the feature vector size of 128 neurons, effectively reducing the input dimensionality by half, balancing between model complexity and computational efficiency, ensuring that the model is capable of learning a rich set of features without being prohibitively expensive to train on top of the encoder and KDM $\rho_{\mathbf{x}',\mathbf{y}'}$ parameters. The training is conducted in an end-to-end manner, optimizing the parameters across all components of the model. This includes the patch encoder, the global-local attention mechanism, and the KDM $\rho_{\mathbf{x}',\mathbf{y}'}$.
Additionally, the trained local-global attention layer of our model can be extended to provide qualitative interpretability of the decisions, not only providing a way to visualize the most important patches in the patch bag but effectively showing the most significant patches in the slide for accurately prediction its ISUP grade group.
The core of the slide-level classification task is the inference process using the KDM $\rho_{\mathbf{x}',\mathbf{y}'}$ and input KDM $\rho_x$ in the same fashion as the fully-supervised case using eq. 2.8.
The density matrix $\rho_y$ is then translated into a discrete probability distribution over the classes. A vector of probabilities is computed from the components of $\rho_y$, where the weights and vectors are denoted by $p''$ and $\mathbf{y}'$, respectively. Both are normalized, $p'' = \frac{p''}{\sum p''}$ and $y' = \frac{y'}{\|y'\|}$, and the probability distribution is obtained as $p'' = \sum_j p''_j y'^2_{ji}$. This probability vector represents the likelihood of the WSI belonging to each class, forming the basis for the slide-level classification or ordinal regression task. For the ordinal regression task, we add a final regression layer that takes the probability distribution of labels $p''$ as input. This layer computes the expected value and variance for predictions. Algorithms 3 and 4 show a summary of the training and prediction procedure of WiSDoM in a weakly-supervised setting.

## 4.2 Experimental Design

This section presents a the experimental design of WiSDoM in the weakly-supervised task. It begins with a description of the dataset characteristics, detailing the specific splits used for training, validation, and testing. This is followed by an overview of the performance measures adopted for assessing WiSDoM effectiveness. We then discuss the baselines selected for comparative analysis.

### 4.2.1 Dataset

To evaluate the performance and interpretability features of WiSDoM, we established the research scenario of weakly supervised ordinal regression task where we aim to predict the

**Algorithm 3:** Weakly-supervised WiSDoM training algorithm

```
Input:
$D = \{(X^{(i)}, y^{(i)})\}_{i=1...N}$: Training dataset, with $X^{(i)} = \{X_j^{(i)}\}_{j=1..k}$ a WSI with $k$ patches
$m$: number of components of KDM $\rho_{\mathbf{x}',\mathbf{y}'}$
$Z$: Deep learning backbone
  1. KDM $\rho_{\mathbf{x}',\mathbf{y}'}$ is initialized with a sample of size $m$ from dataset $D$
  2. for each $(X^{(i)}, y^{(i)}) \in D$:
         $\rho_y = (\{y_i'\}_{i=1...m}, (p_i'')_{i=1...m}, k_{\mathbb{Y}}) = \text{predict}(X^{(i)}, \rho_{\mathbf{x}',\mathbf{y}'}, Z)$ (see algorithm 4)
  3. If task = classification:
     a) $\hat{y} = y'_{\arg\max(p'')}$
     b) Minimize $L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$
  4. If task = ordinal regression:
     a) Calculate $E[\hat{y}]_i = \sum_{j=1}^{m} p''_{ij} \cdot y'_j$ and $Var[\hat{y}]_i = E[\hat{y}^2]_i - (E[\hat{y}]_i)^2$, where
        $E[\hat{y}^2]_i = \sum_{j=1}^{m} p''_{ij} \cdot y_j'^2$
     b) Minimize $L = \frac{1}{N} \sum_{i=1}^{N} (E[\hat{y}]_i - y_i)^2 + \alpha \cdot Var[\hat{y}]_i$, where $\alpha$ is a penalization
        parameter for variance.
  5. Update backbone, MLP$_1$ and MLP$_2$ weights $\mathbf{w}$, and KDM $\rho_{\mathbf{x}',\mathbf{y}'}$ parameters using
     gradient descent.
  6. Return $(Z, \rho_{\mathbf{x}',\mathbf{y}'})$
```

**Algorithm 4:** Weakly-supervised WiSDoM prediction procedure

```
Input:
$X = \{x_j\}_{j=1..k}$: input WSI with $k$ patches
$\rho_{\mathbf{x}',\mathbf{y}'} = \{(x_i', y_i)_{i=1...m}, (p_i)_{i...m}, k_{\mathbb{X}} \otimes k_{\mathbb{Y}}\}$: joint KDM
$Z$: Deep Learning backbone
  1. $\mathbf{z}_j^{\text{local}} = \text{MLP}_1(\mathbf{x}_j)$
  2. $\mathbf{z}^{\text{global}} = \frac{1}{k} \sum_{j=1}^{k} \mathbf{z}_j^{\text{local}}$
  3. $\mathbf{z}_j^{\text{attn}} = \text{Softmax}(\text{MLP}_2((\mathbf{z}_j^{\text{local}}, \mathbf{z}_{\text{global}})))$
  4. Encode patches using $Z$: $z_j = Z(x_j)$
  5. Create $\rho_x = (\{z_j\}_{j=1...k}, (\mathbf{z}_j^{\text{attn}})_{j=1...k}, k_{\mathbb{X}})$
  6. Calculate probabilities $p''$ from output KDM $\rho_y$ using $\rho_x$ and $\rho_{\mathbf{x}',\mathbf{y}'}$ with eq. 2.8
  7. $\rho_y = (\{y_i'\}_{i=1...m}, (p_i'')_{i=1...m}, K_Y)$
  8. Return $\rho_y$
```

**Table 4-1**: PANDA dataset description (Percentage), GG = ISUP Grade Group

| Source | No. of biopsies | Nontumor | GG 1 | GG 2 | GG 3 | GG 4 | GG 5 |
|---|---|---|---|---|---|---|---|
| Radboud UMC (D1) | 5160 | 967 (19) | 852 (17) | 675 (13) | 925 (18) | 768 (15) | 973 (19) |
| Karolinska Institutet (D2) | 5456 | 1925 (35) | 1814 (33) | 668 (12) | 317 (6) | 481 (9) | 251 (5) |

ISUP grade group for the entire whole-slide image (WSI) using selected bags of patches from each WSI.

The effectiveness of WiSDoM at the slide level was assessed using the Prostate Cancer Grade Assessment (PANDA) dataset[1] like with the fully-supervised model.

For our study, the data from D1 and D2 were divided into training, validation, and testing sets to assess the model's performance in whole-slide ISUP grading. The composition of the PANDA dataset are detailed in Table **4-1**. It is important to note that while the patch data used in chapter 3 originates solely from D1, the WSI dataset is a composite of data from both D1 and D2, following the same split to ensure data consistency.

### 4.2.2 Performance Measures

The efficacy of WiSDoM in both classification and ordinal regression tasks in a weakly-supervised scenario is assessed using the same three metrics as in the fully-supervised setting: Cohen's Kappa ($\kappa$, see equation 3.5), Accuracy (see equation 3.6), and Mean Absolute Error (MAE, see equation: 3.7).

### 4.2.3 Baseline

On the weakly supervised task, we compare the performance of WiSDoM with the publicly available dataset scores of the winning teams from the PANDA Challenge consortium [1] (see table **4-3**). Most strategies from the winning teams included weakly-supervised learning with CNNs on a mosaic of patches representing the whole slide, ensemble networks, and label denoising, i.e., removing from the training set during k-fold cross-validation, samples that are difficult for the network to predict.

### 4.2.4 Training details

During training, our process involves extracting a set of patches, referred to as a 'bag,' from each slide. The aim is to select patches containing substantial useful information. This is achieved using an automatic tissue detection process to identify areas rich in tissue within the slides. From these, the top $\boldsymbol{x}^{(i)j}$ patches, each with more than 90% tissue content, are randomly selected. This method ensures the training patches are both informative and diverse. Given computational capacity constraints and considering that prostate core-needle biopsies are relatively small, we selected 36 patches at 20× magnification, each measuring

192×192 pixels, for each whole-slide image. We employ a deep learning encoder, specifically a pre trained ConvNeXT [67], for mapping patch bags to latent space. The network undergoes a warm-up for 2 epochs by processing patches in a classification task before attaching this backbone to the KDM with adjusted weights post-warm-up. For KDM $\rho_{\mathbf{x}',\mathbf{y}'}$ initialization, which requires patch-label pairs $x'$ and $y'$ and importance weights $p'$, we select a balanced set of patch-label pairs from the training set. These pairs, termed trainable prototypes, are processed through the warmed-up deep learning backbone. The pairs in the latent space are then used to initialize $x'$ and $y'$. Without pre-training importance information, $p'$ is initialized with equal weights $1/m_i$ for all pairs. Systematic hyperparameter tuning led to the selection of 216 prototypes, with 36 patch-label pairs representing each ISUP grade group. After initializing KDM $\rho_{\mathbf{x}',\mathbf{y}'}$, the deep learning backbone, attention module, and KDM $\rho_{\mathbf{x}',\mathbf{y}'}$ are trained end-to-end. We use the Adam [70] optimizer with a learning rate of 0.0001, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. A gradual warm-up scheduler with a factor of 10 is applied for 1 epoch, followed by cosine annealing for the remaining epochs. The mini-batch size is set to 4 bags. For the loss function, we use categorical cross-entropy for the classification task. The model is trained for 50 epochs with an early-stopping callback to prevent overfitting, stopping training after 5 epochs without validation loss improvement.
The warm-up and KDM $\rho_{\mathbf{x}',\mathbf{y}'}$ initialization are the same for the ordinal regression task. However, we use real-valued labels in the range [0,1] instead of one-hot encoded labels during training. The loss function is modified to Mean Squared Error with an additional penalization $\alpha$ for high variance predictions. The same Adam optimizer settings are employed, along with a gradual warm-up scheduler and a mini-batch size of 4 bags.

## 4.3 Results

The flexibility of WiSDoM for both full and weak supervision, enables its application in grading entire slides without the necessity for pixel-level annotations. This feature is particularly valuable due to the high cost and scarcity of detailed annotations. We evaluated WiSDoM on the PANDA dataset, focusing this time exclusively on whole-slide labels. The objective was to classify entire slides according to the ISUP grading system, which includes six grades, each linked to a corresponding prior Gleason score (GS): ISUP 0 (benign), ISUP 1 (GS 3+3), ISUP 2 (GS 3+4), ISUP 3 (GS 4+3), ISUP 4 (GS 4+4, 3+5, 5+3), and ISUP 5 (GS 4+5, 5+4, 5+5). In this grading task, the model achieved a $\kappa$ of 0.898 and an accuracy rate of 0.663 (see tables **4-2** and **4-3**, aligning with the reference standard [1].

### Uncertainty Quantification

Similar to the fully supervised patch model approach, WiSDoM was also adapted to treat the grading challenge as an ordinal regression problem in a weakly supervised context. In this setting, the model recorded a $\kappa$ of 0.900, an accuracy of 0.660, and a MAE of 0.173.

Remarkably, when utilizing the variance of the ordinal regression model as an indicator of the model's confidence (with a threshold of $\sigma^2 < 0.05$), improvements were observed across the metrics, with a $\kappa$ of 0.930, accuracy of 0.73, and MAE of 0.073. Figure **4-2** presents violin plots of the test set predictions, categorized by absolute error and variance values, illustrating that correct predictions typically correspond to lower variance, even under weakly supervised training conditions.

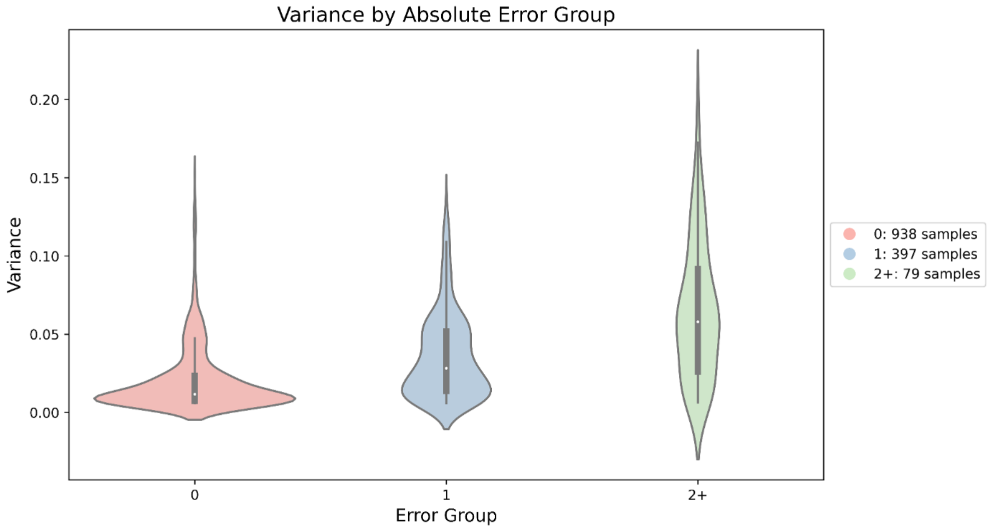

**Figure 4-2**: Violin plots of slide-level weakly-supervised predictions variance grouped by absolute error, as seen in the plot, correct predictions have less variance while error 1 and 2 groups tend to have a much larger variance

Table **4-3** presents a comparison of the leading solutions from the PANDA Challenge [1]. Alongside the winning team scores, the table includes an additional entry to compare the performance metrics of our model on the public test set. This test set comprises a subset of data that was openly accessible for model development and validation.

**Table 4-2**: Performance Metrics for Whole-Slide Grading

| WiSDoM | $\kappa$ | **Accuracy** | **MAE** |
|---|---|---|---|
| Whole-Slide Classification | 0.898 | 0.663 | - |
| Ordinal Regression (Whole-Slide) | 0.900 | 0.660 | 0.173 |
| Ordinal Regression with Variance Threshold ($\sigma^2 < 0.05$) | **0.930** | **0.73** | **0.073** |

### Interpretability

The interpretability of deep learning classifiers, particularly in a weakly supervised context, plays a crucial role in validating their predictive accuracy and aligning with established

**Table 4-3**: Performance comparison with PANDA consortium teams using public dataset scores[1]. The open development dataset, accessible for research, was the only source used for our model evaluation and comparison.

| **PANDA Consortium Team** | $\kappa$ |
|---|---|
| Dmitry A. Grechka | 0.8861 |
| KovaLOVE v2 | 0.8889 |
| ctrasd123 | 0.8948 |
| Manuel Campos | 0.898 |
| Kiminya | 0.9007 |
| ChienYiChi | 0.9086 |
| rähmä.ai | 0.9096 |
| PND | 0.9108 |
| BarelyBears | 0.9118 |
| iafoss | 0.9179 |
| NS Pathology | 0.9180 |
| Save The Prostate | 0.9209 |
| **WiSDoM (Ours)** | **0.9300** |

morphological criteria used in pathology. Moreover, this interpretability is instrumental in analyzing cases where the model may not perform as expected. In clinical applications, heatmaps generated at the whole-slide level offer valuable support for AI-assisted diagnoses. WiSDoM provides multifaceted interpretability, which is instrumental in enhancing clinician trust in automated diagnostic support tools. The model's capability for visual interpretability is evident through its generation of heatmaps that emphasize regions of high diagnostic importance.

In the weakly supervised setting, although the heatmaps lack the explicit detail of Gleason patterns per patch due to the absence of such information during training, they still effectively identify regions crucial for ISUP grade group classification. Notably, there is a consistent correlation between the regions highlighted in both fully and weakly supervised heatmaps. This overlap demonstrates that even with coarser granularity in the weakly supervised setting, the heatmaps remain effective in underscoring regions contributing significantly to the ISUP grade group classification. Furthermore, a comparison of these heatmaps with pathologist annotations (see figure **4-3**) reveals their efficacy in pinpointing morphological features pertinent to human pathology assessments, such as fused cell sheets indicative of Gleason 5 patterns and cribriform patterns for Gleason 4.

Additionally, these heatmaps prove valuable in analyzing misclassified slides. For instance, challenging cases were noted where the model identified regions as low-aggressive Gleason patterns but failed to accurately classify them. This finding mirrors the complexities encountered in Gleason grading, particularly in differentiating benign from Gleason 3 patterns, a

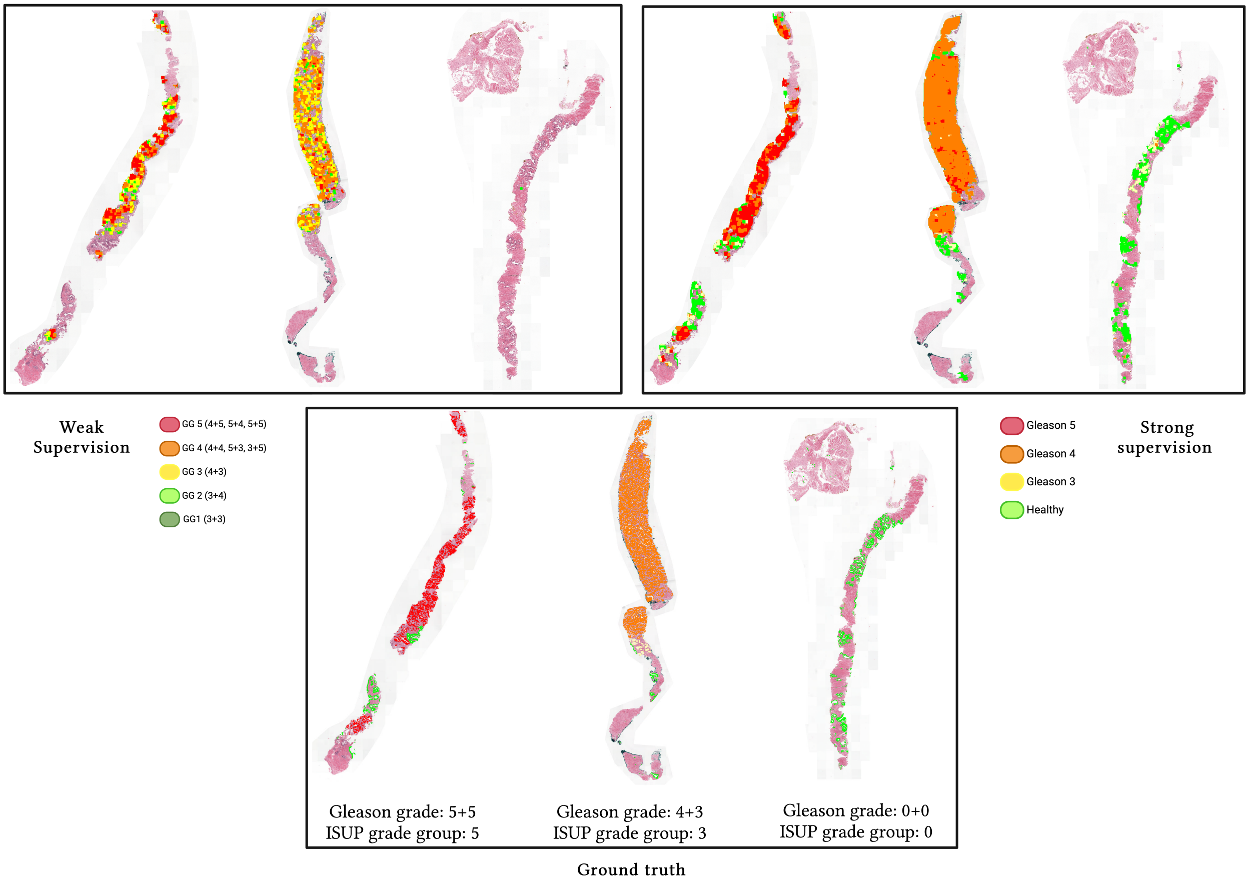

**Figure 4-3**: **Visual Heatmap Comparison Across Supervision Levels.** This figure illustrates our model-generated heatmaps under different supervision settings, compared with pathologist-provided ground truth annotations. The top left shows weakly supervised heatmaps identifying ISUP grade groups, reflecting diagnostic relevance akin to Gleason patterns in both fully-supervised and ground truth examples. The top right features detailed heatmaps from the fully supervised model, trained on Gleason-pattern-labeled patches, closely mirroring the ground truth (bottom). The detail level of these heatmaps is modifiable through patch overlap adjustments during inference.

task known for its difficulty[71].

Overall, WiSDoM ability to generate informative heatmaps in both full and weak supervision scenarios enhances its utility as an interpretative tool, providing crucial insights for clinicians and aiding in the nuanced task of prostate cancer grading. While the heatmaps offer valuable practical insights, it is important to exercise caution and not interpret them as exact segmentation masks with pixel-level precision. However, despite this limitation, the straightforward and intuitive nature of these visualizations offers researchers significant insights into the morphological patterns that underpin the model's predictions. This understanding is crucial for both the validation of the model's decision-making process and for exploring the morphological basis of its predictive outcomes.

Despite the visual interpretability provided by heatmaps, the 'black box' nature of deep learning models is not addressed by them, even though we can now visualize the predictions made by the model. The exploration of interpretability techniques for deep learning models in medical imaging extends beyond the scope of heatmaps, encompassing a range of diverse methods. Each technique offers a unique perspective in making these complex models more transparent and clinically relevant. The methods include: concept learning models [72], prototype-based models[53, 73, 74], counterfactual explanations[75, 76], internal network representations[76], among others[60], these techniques collectively contribute to reduce the 'black-box' nature of the decision-making processes of deep learning models, enhancing their interpretability and suitability for clinical application.

WiSDoM distinguishes itself from other deep learning models in terms of interpretability. During its training phase, the model's learnable parameters, comprising a set of samples $x'$, and importance weights $p'$, are initialized using examples from the training set. These examples are first encoded into a latent space (see Fig. **4-4**) using a deep learning backbone. The parameters, representing a joint probability distribution of samples and labels, are then fine-tuned during training to optimize performance, whether for classification or ordinal regression tasks.

We refer to these parameters as 'prototypes', which offer a look into the model's internal representation of each class, including Gleason patterns or ISUP grade groups. This feature enables a practical comparison between the model's internal perception of these patterns and their established clinical interpretations. The prototypes, existing in the latent space, are exemplified by selecting the nearest samples from the training set.

We conducted a study with three resident pathologists to assess the relevance and accuracy of 36 prototypes derived from WiSDoM. For this evaluation, the prototypes were shown to the pathologists without any accompanying labels, facilitating a blind assessment of Gleason grading in context with the corresponding whole-slide images. This approach allowed for an unbiased evaluation of the prototypes. The results of this study indicated a substantial agreement ($\kappa = 0.88$) between the prototype labels and the Gleason patterns determined by the pathologists.

Figures **4-4**, **4-5**, and **4-6** demonstrate that the prototypes sampled from WiSDoM not only accurately represent but also closely mimic the morphology of the targeted regions of interest. This resemblance is crucial, as it validates the model's ability to mirror clinical observations.

Moreover, these prototypes serve as an additional interpretability tool. They enable the model to convey to clinicians how certain regions of interest correlate with specific prototypes learned by the model in its internal representation. This feature of WiSDoM is instrumental in bridging the gap between automated predictions and clinical diagnostics, providing clinicians with understandable and relatable visual representations that align with

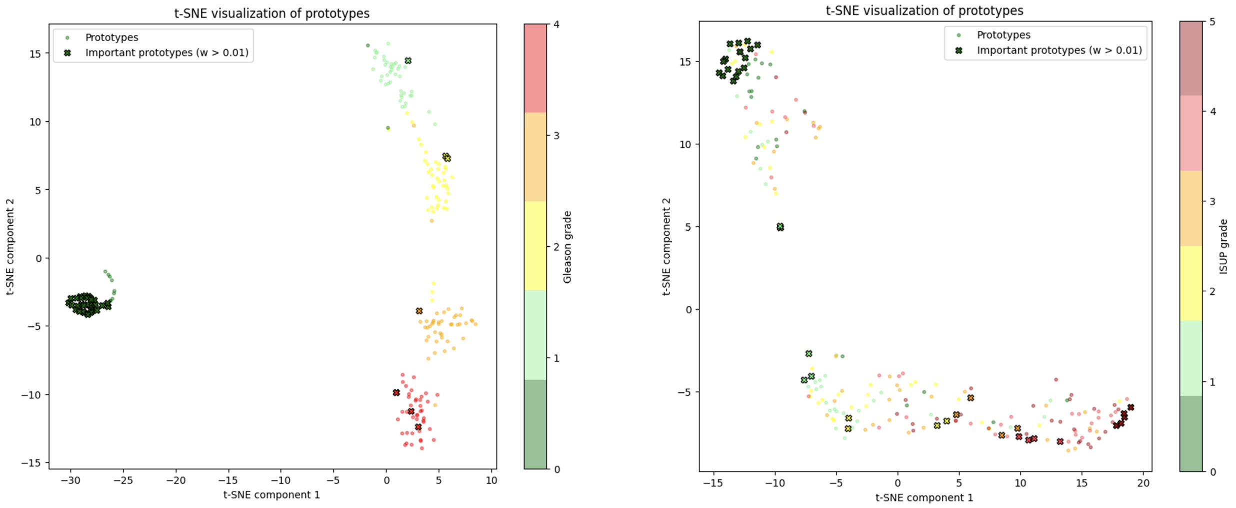

**Figure 4-4**: **Learned prototypes feature space.** t-distributed Stochastic Neighbor Embedding (t-SNE) plot of the learned prototypes inside WiSDoM for different supervision scenarios, in the fully supervised model (left), the prototypes perfectly discriminate the latent space in Gleason grades. In the weakly-supervised model (right), not all prototypes are discriminant of the latent space, however, prototypes with weights or importance over 0.01 (marked with x) can efficiently differentiate ISUP grades in the latent space.

their expertise.

Post-training, we observe that the weight $p'$ of some learnable prototypes, are weighted to zero as shown in figure **4-4**. This suggests that WiSDoM requires only a subset of the initialized prototypes to differentiate between ISUP grade groups and Gleason grades.

The final aspect of interpretability offered by WiSDoM is uncertainty quantification. This concept has been extensively studied in the field of ordinal regression to develop more interpretable models, particularly when reliability is crucial for end-users, such as in clinical environments [77]. Quantifying a model's uncertainty can mitigate the risks associated with relying solely on its predictions. This is relevant in medical contexts where incorrect diagnoses can lead to significant patient harm. Our approach utilizes a probabilistic model framework, enabling the generation of prediction outputs as actual probability distributions across various degrees. This is achieved without enforcing a distribution using softmax or similar activation functions, thereby allowing the interpretation of variance as a direct measure of uncertainty. Furthermore, unlike traditional probabilistic methods, WiSDoM can be trained via gradient descent, facilitating its seamless integration with standard deep learning architectures. As seen in figure **4-7**, an uncertainty heatmap can be obtained from WiSDoM, clearly highlighting regions of the whole slide where the model is unsure of its prediction.

WiSDoM offers a comprehensive approach to interpretability, encompassing several aspects relevant for clinical applications. This includes the generation of prediction heatmaps and

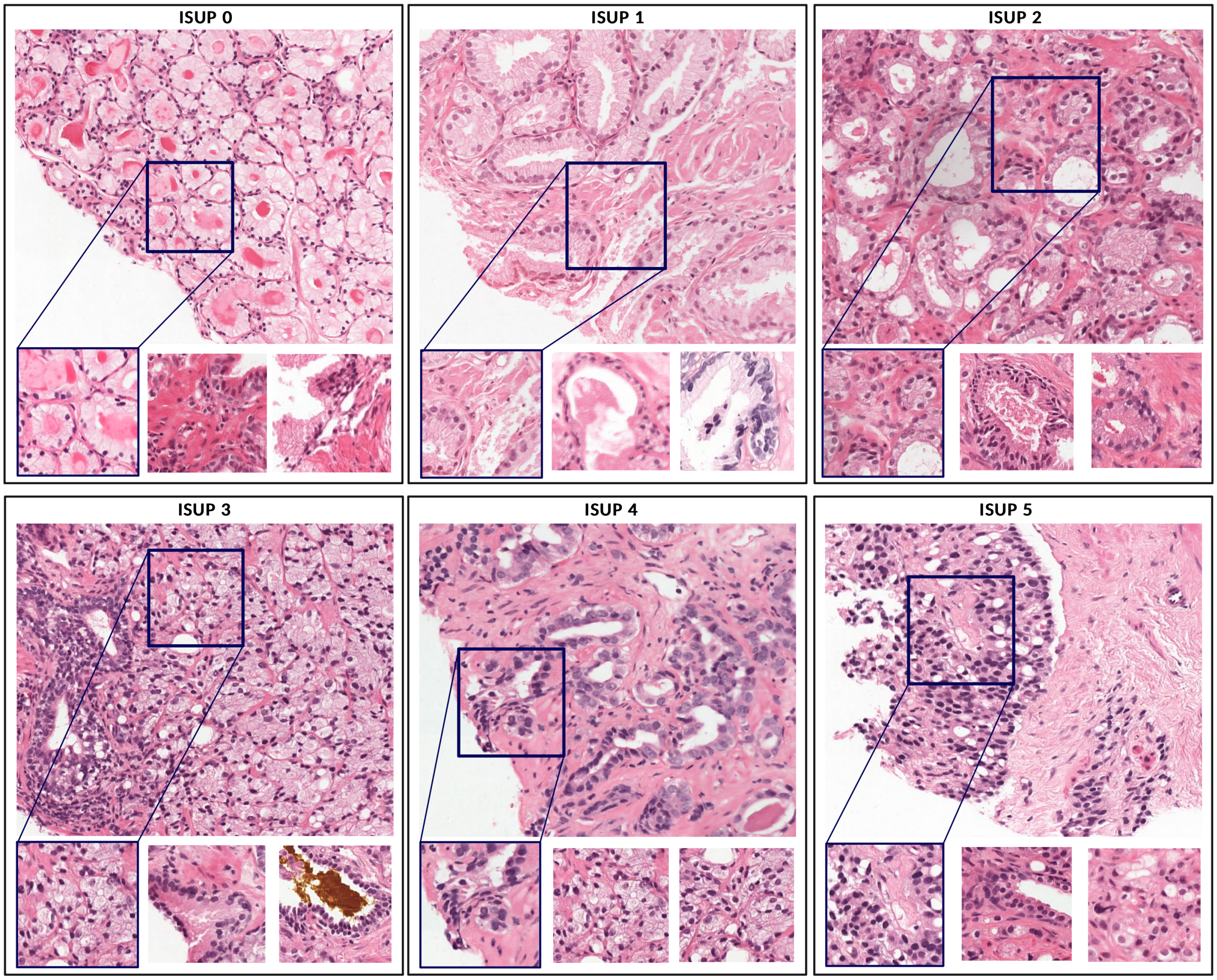

**Figure 4-5**: **Learned prototypes.** After training, prototypes are sampled from WiSDoM. These learned prototypes enhance model explainability. The figure illustrates the top three patches closest to the learned prototypes for each ISUP grade group. Additionally, the closest prototype for each grade group is displayed in the context of the whole slide. As observed, the prototypes appropriately validate that the internal representation of ISUP grade groups is effectively encapsulated by the morphological patterns inherent in the Gleason grades constituting the grade group.

prototype-based explanations, which offer visual and intuitive insights into the model's decision-making process. Additionally, WiSDoM incorporates uncertainty quantification, providing not only variance values for each prediction but also heatmaps that visually indicate areas where the model lacks confidence. These features collectively enhance the utility of WiSDoM in clinical environments, where the interpretability of computer-aided diagnosis tools is essential for gaining trust and ensuring reliability.

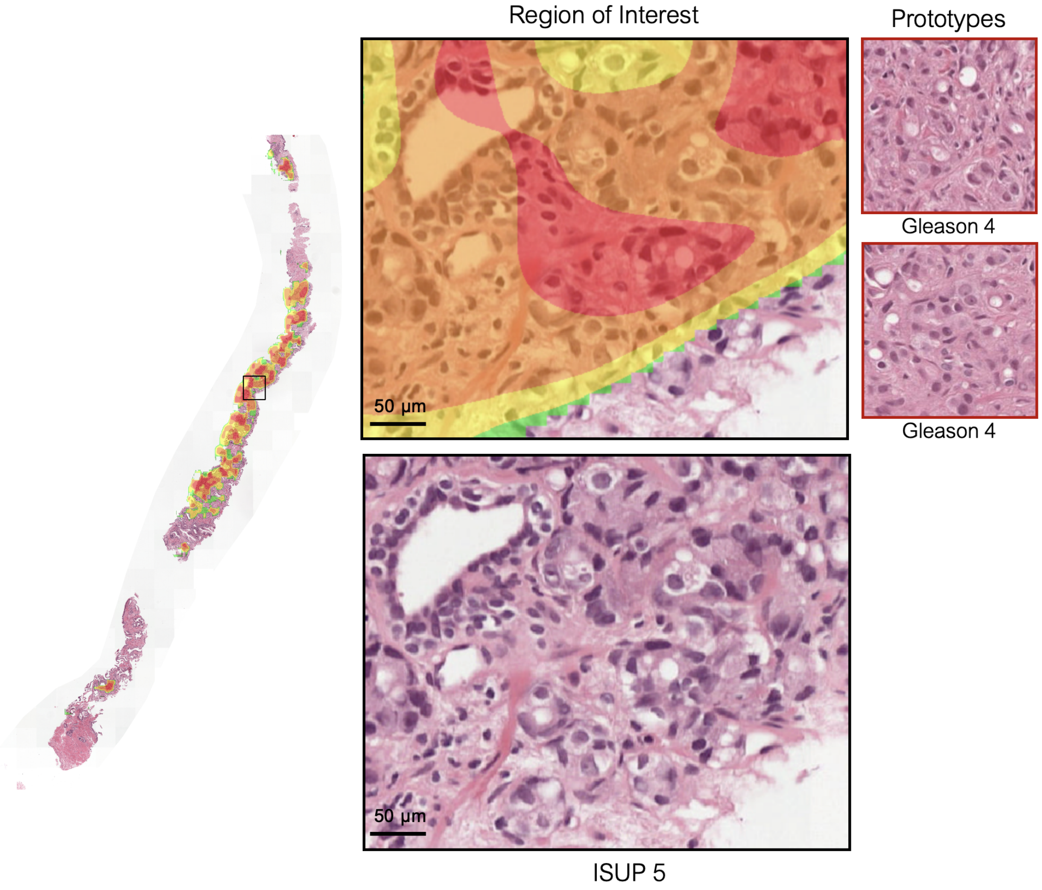

**Figure 4-6**: **Prototype-based model explainability.** Given a region of particular interest highlighted by the model, example prototypes can be sampled from the learned representation of WiSDoM with their corresponding Gleason grade, providing a visual yet clinically relevant understanding of the model's decision-making.

## 4.4 Discussion

In this study, we demonstrated the versatility and efficacy of WiSDoM in addressing various challenges in computational pathology. Our findings indicate that WiSDoM performs comparably to expert pathologists in Gleason grading tasks. Notably, WiSDoM achieves this level of accuracy both with comprehensive tissue annotations and with only slide-level labels, the latter being particularly advantageous due to the labor-intensive nature of manual annotations in whole-slide images.

A significant advantage of WiSDoM lies in its interpretability tools, which are applicable in both research and clinical settings. The model generates visual heatmaps that highlight diagnostically important regions in whole-slide images. These heatmaps are valuable for

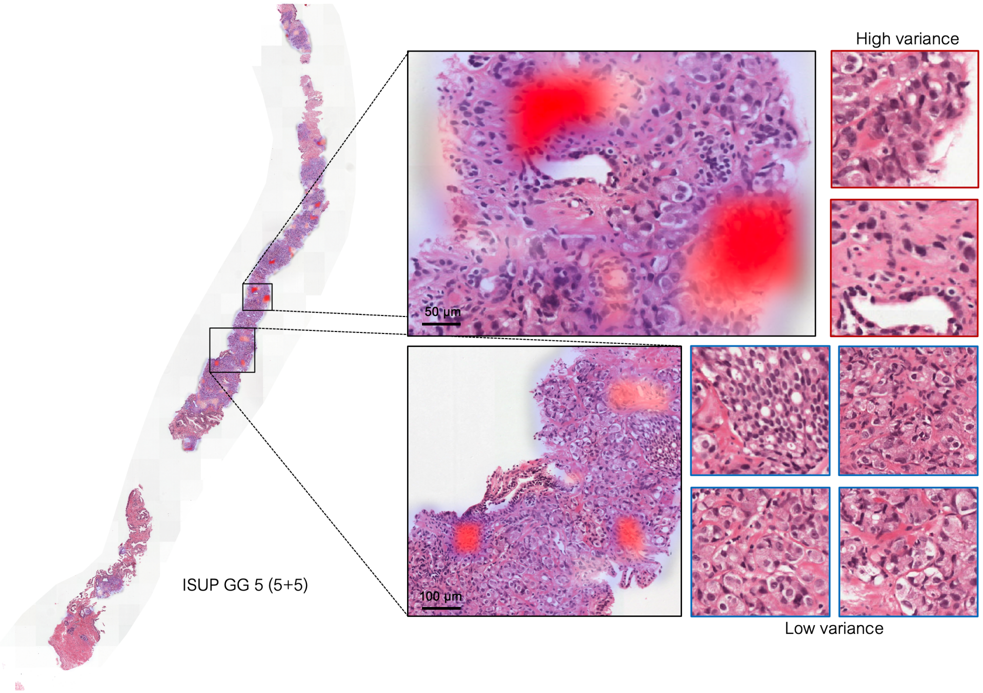

**Figure 4-7**: **Model uncertainty visualization.** A representative slide is shown with regions of high variance highlighted in red. The whole-slide heatmap was generated by obtaining the variance values for the prediction over patches tiled at 80% overlap, with zoomed-in regions on the right. Patches with a red border indicate regions where the model's uncertainty of the prediction was high, while blue borders indicate high confidence in the prediction.

identifying critical areas for both overall slide grading and finer Gleason grading, particularly when trained with detailed tissue annotations. However, we recognize that visual interpretability alone does not fully demystify the 'black box' aspect of deep learning models. To address this, WiSDoM incorporates prototype-based explainability. This feature elucidates the internal representations learned by the model, identifying the closest internal representations to diagnostically relevant regions. Such transparency is crucial to fostering trust in AI-assisted diagnostics among pathologists.

Further enhancing this trust, WiSDoM quantifies uncertainty at both the slide and region-of-interest levels. This capability allows the model to withhold predictions in cases of high uncertainty, thereby boosting confidence in its diagnostic suggestions. The model can also be configured to operate within predefined confidence thresholds, highlighting areas of diagnostic ambiguity that may be particularly informative for pathologists.

Despite these advancements, certain challenges persist and warrant further investigation. Our observations suggest that while a relatively small number of patches is sufficient for slide-level diagnosis, the requisite volume of data (in terms of the number of slides or 'bags') remains substantial for optimal model performance. Future research should focus on evaluating WiSDoM's performance across different organs and surgical resections. Our training on core-needle biopsies, which contain less tissue, raises questions about the model's efficacy in larger resection samples, potentially necessitating an increased number of patches and, consequently, higher computational resources.

Our study provides valuable insights into the development of weakly-supervised, interpretable deep learning models for clinical applications. We anticipate that these findings will facilitate the clinical adoption of such models, thereby enhancing diagnostic processes in pathology.

### 4.4.1 Code Availability

Patches from Whole Slide Images (WSIs) were generated locally using HistoPrep[78]. Network training was conducted on NVIDIA A5500 GPUs on-site and NVIDIA A100 GPUs on Google Colab Pro[1]. Our pipeline, implemented in Python (3.11)[2], utilizes OpenSlide[3], Pillow[4], and TensorFlow v2[5].

The code is publicly available here

[1] https://colab.research.google.com

[2] https://www.python.org

[3] https://openslide.org

[4] https://pillow.readthedocs.io

[5] https://www.tensorflow.org

# 5 Conclusions and future work

This thesis contributes to the field of computational pathology, particularly in the context of prostate cancer grading. We have successfully combined deep neural networks with probabilistic models, leveraging Kernel Density Matrices into a weakly supervised, interpretable computational pathology setting. This novel approach has demonstrated its efficacy in modeling uncertainty, providing additional interpretability, and achieving state-of-the-art prostate tissue differentiation, quantification, and grading performance.

Adhering to the same model philosophy, we are capable of quantifying Gleason patterns at the patch level, offering a tool that aligns naturally with pathologists' methodologies. This is achieved by grading based on the quantification of the area of the two most prevalent Gleason patterns present in the slides, albeit requiring tissue annotation masks for training. Simultaneously, our innovative application of weakly supervised MIL and KDM allows us to manage the inherent uncertainties of the training samples, even without explicit knowledge of patch labels. This technique, which learns from collections or 'bags' of patches from WSI, holds significant relevance in the field of pathology, especially in the context of ISUP group grading. This is particularly the case when comprehensive whole-slide tissue annotations are rare and costly.

Moreover, incorporating a local-global attention mechanism improved the model's performance by attributing weights to each instance within a bag. This not only enhances the model's agreement with expert pathologists but also provides an additional layer of interpretability by identifying regions of interest (ROIs), achieved without the need for expert tissue annotations during model training.

WiSDoM excels in three key areas. First, it provides quantification at the patch level of the extent of the Gleason patterns in a given slide. Second, it highlights ROIs for pathologists in both a fully supervised manner and a weakly supervised way with local-global attention. Third, it enables the sampling of prototypes of each tissue pattern from the joint probability distribution estimated during the training of the KDM $\rho_{\mathbf{x}',\mathbf{y}'}$. These prototypes serve as an explainable representation of the model's internal understanding of the various tissue patterns present in the prostate.

In conclusion, this work has made significant strides in enhancing the interpretability of deep learning models in a clinical setting. The ability to understand the reasoning behind a model's predictions is as crucial as the accuracy of the predictions themselves. WiSDoM provides valuable insights into its learning process and understanding of the data, making it a valuable contribution to the field of computational pathology.

# Bibliography

[1] Bulten et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature Medicine*, 28(1):154–163, Jan 2022. ISSN 1546-170X. doi: 10.1038/s41591-021-01620-2. URL `https://doi.org/10.1038/s41591-021-01620-2`.

[2] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[3] Prashanth Rawla. Epidemiology of prostate cancer. *World journal of oncology*, 10(2): 63, 2019.

[4] Donald F Gleason and George T Mellinger. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. *The Journal of urology*, 111(1):58–64, 1974.

[5] Mark S Litwin and Hung-Jui Tan. The diagnosis and treatment of prostate cancer: a review. *Jama*, 317(24):2532–2542, 2017.

[6] Jonathan I Epstein, Lars Egevad, Mahul B Amin, Brett Delahunt, John R Srigley, and Peter A Humphrey. The 2014 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma. *The American journal of surgical pathology*, 40(2):244–252, 2016.

[7] Tayyar A Ozkan, Ahmet T Eruyar, Oguz O Cebeci, Omur Memik, Levent Ozcan, and Ibrahim Kuskonmaz. Interobserver variability in gleason histological grading of prostate cancer. *Scandinavian journal of urology*, 50(6):420–424, 2016.

[8] Patricia Raciti, Jillian Sue, Rodrigo Ceballos, Ran Godrich, Jeremy D Kunz, Supriya Kapur, Victor Reuter, Leo Grady, Christopher Kanan, David S Klimstra, et al. Novel artificial intelligence system increases the detection of prostate cancer in whole slide images of core needle biopsies. *Modern Pathology*, 33(10):2058–2066, 2020.

[9] Wouter Bulten, Hans Pinckaers, Hester van Boven, Robert Vink, Thomas de Bel, Bram van Ginneken, Jeroen van der Laak, Christina Hulsbergen-van de Kaa, and Geert Litjens. Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study. *The Lancet Oncology*, 21(2):233–241, 2020.

[10] Lars Egevad, T Granfors, L Karlberg, A Bergh, and Per Stattin. Prognostic value of the gleason score in prostate cancer. *BJU international*, 89(6):538–542, 2002.

[11] Matthew R. Cooperberg, Jeanette M. Broering, and Peter R. Carroll. Time trends and local variation in primary treatment of localized prostate cancer. *Journal of Clinical Oncology*, 28(7):1117–1123, 2010. doi: 10.1200/JCO.2009.26.0133. URL `https://doi.org/10.1200/JCO.2009.26.0133`. PMID: 20124165.

[12] Jonathan I Epstein. An update of the gleason grading system. *The Journal of urology*, 183(2):433–440, 2010.

[13] Lars Egevad, Amar S Ahmad, Ferran Algaba, Daniel M Berney, Liliane Boccon-Gibod, Eva Compérat, Andrew J Evans, David Griffiths, Rainer Grobholz, Glen Kristiansen, Cord Langner, Antonio Lopez-Beltran, Rodolfo Montironi, Sue Moss, Pedro Oliveira, Ben Vainer, Murali Varma, and Philippe Camparo. Standardization of gleason grading among 337 european pathologists. *Histopathology*, 62(2):247–256, 2013. doi: https://doi.org/10.1111/his.12008. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/his.12008`.

[14] Kaustav Bera, Kurt A Schalper, David L Rimm, Vamsidhar Velcheti, and Anant Madabhushi. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nature reviews Clinical oncology*, 16(11):703–715, 2019.

[15] Anant Madabhushi, Michael D Feldman, and Patrick Leo. Deep-learning approaches for gleason grading of prostate biopsies. *The Lancet Oncology*, 21(2):187–189, 2020.

[16] Eirini Arvaniti et al. Automated gleason grading of prostate cancer tissue microarrays via deep learning. *Scientific Reports*, 8(1):12054, Aug 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-30535-1. URL `https://doi.org/10.1038/s41598-018-30535-1`.

[17] Ming Y Lu, Drew FK 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*, 5(6):555–570, 2021.

[18] Fabio A. González, Raúl Ramos-Pollán, and Joseph A. Gallego-Mejia. Kernel density matrices for probabilistic deep learning, 2023.

[19] Mohamed Slaoui and Laurence Fiette. Histopathology procedures: from tissue sampling to histopathological evaluation. *Drug Safety Evaluation: Methods and Protocols*, pages 69–82, 2011.

[20] Geert Litjens, Clara I. Sánchez, Nadya Timofeeva, Meyke Hermsen, Iris Nagtegaal, Iringo Kovacs, Christina Hulsbergen van de Kaa, Peter Bult, Bram van Ginneken, and Jeroen van der Laak. Deep learning as a tool for increased accuracy and efficiency of

histopathological diagnosis. *Scientific Reports*, 6(1):26286, May 2016. ISSN 2045-2322. doi: 10.1038/srep26286. URL `https://doi.org/10.1038/srep26286`.

[21] William C Allsbrook, Jr, Kathy A Mangold, Maribeth H Johnson, Roger B Lane, Cynthia G Lane, and Jonathan I Epstein. Interobserver reproducibility of gleason grading of prostatic carcinoma: General pathologist. *Hum. Pathol.*, 32(1):81–88, January 2001.

[22] William C Allsbrook, Jr, Kathy A Mangold, Maribeth H Johnson, Roger B Lane, Cynthia G Lane, Mahul B Amin, David G Bostwick, Peter A Humphrey, Edward C Jones, Victor E Reuter, Wael Sakr, Isabell A Sesterhenn, Patricia Troncoso, Thomas M Wheeler, and Jonathan I Epstein. Interobserver reproducibility of gleason grading of prostatic carcinoma: Urologic pathologists. *Hum. Pathol.*, 32(1):74–80, January 2001.

[23] Karolina Cyll, Elin Ersvær, Ljiljana Vlatkovic, Manohar Pradhan, Wanja Kildal, Marte Avranden Kjær, Andreas Kleppe, Tarjei S. Hveem, Birgitte Carlsen, Silje Gill, Sven Löffeler, Erik Skaaheim Haug, Håkon Wæhre, Prasanna Sooriakumaran, and Håvard E. Danielsen. Tumour heterogeneity poses a significant challenge to cancer biomarker research. *British Journal of Cancer*, 117(3):367–375, Jul 2017. ISSN 1532-1827. doi: 10.1038/bjc.2017.171. URL `https://doi.org/10.1038/bjc.2017.171`.

[24] Arpit Aggarwal, Sirvan Khalighi, Deepak Babu, Haojia Li, Sepideh Azarianpour-Esfahani, Germán Corredor, Pingfu Fu, Mojgan Mokhtari, Tilak Pathak, Elizabeth Thayer, Susan Modesitt, Haider Mahdi, Stefanie Avril, and Anant Madabhushi. Computational pathology identifies immune-mediated collagen disruption to predict clinical outcomes in gynecologic malignancies. *Communications Medicine*, 4(1):2, Jan 2024. ISSN 2730-664X. doi: 10.1038/s43856-023-00428-0. URL `https://doi.org/10.1038/s43856-023-00428-0`.

[25] Cristian Barrera, Germán Corredor, Vidya Sankar Viswanathan, Ruiwen Ding, Paula Toro, Pingfu Fu, Christina Buzzy, Cheng Lu, Priya Velu, Philipp Zens, et al. Deep computational image analysis of immune cell niches reveals treatment-specific outcome associations in lung cancer. *NPJ precision oncology*, 7(1):52, 2023.

[26] Jeroen Van der Laak, Geert Litjens, and Francesco Ciompi. Deep learning in histopathology: the path to the clinic. *Nature medicine*, 27(5):775–784, 2021.

[27] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[28] George Lee, Robert W Veltri, Guangjing Zhu, Sahirzeeshan Ali, Jonathan I Epstein, and Anant Madabhushi. Nuclear shape and architecture in benign fields predict biochemi-

cal recurrence in prostate cancer patients following radical prostatectomy: preliminary findings. *European urology focus*, 3(4-5):457–466, 2017.

[29] Cheng Lu, David Romo-Bucheli, Xiangxue Wang, Andrew Janowczyk, Shridar Ganesan, Hannah Gilmore, David Rimm, and Anant Madabhushi. Nuclear shape and orientation features from h&e images predict survival in early-stage estrogen receptor-positive breast cancers. *Laboratory investigation*, 98(11):1438–1448, 2018.

[30] Germán Corredor, Xiangxue Wang, Yu Zhou, Cheng Lu, Pingfu Fu, Konstantinos Syrigos, David L Rimm, Michael Yang, Eduardo Romero, Kurt A Schalper, et al. Spatial architecture and arrangement of tumor-infiltrating lymphocytes for predicting likelihood of recurrence in early-stage non–small cell lung cancer. *Clinical cancer research*, 25(5):1526–1534, 2019.

[31] Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics*, 7(1):29, 2016. ISSN 2153-3539. doi: https://doi.org/10.4103/2153-3539.186902. URL `https://www.sciencedirect.com/science/article/pii/S2153353922005478`.

[32] Maschenka C.A. Balkenhol, David Tellez, Willem Vreuls, Pieter C. Clahsen, Hans Pinckaers, Francesco Ciompi, Peter Bult, and Jeroen A.W.M. van der Laak. Deep learning assisted mitotic counting for breast cancer. *Laboratory Investigation*, 99(11): 1596–1606, 2019. ISSN 0023-6837. doi: https://doi.org/10.1038/s41374-019-0275-0. URL `https://www.sciencedirect.com/science/article/pii/S002368372200561X`.

[33] Angel Cruz-Roa, Hannah Gilmore, Ajay Basavanhally, Michael Feldman, Shridar Ganesan, Natalie NC Shih, John Tomaszewski, Fabio A González, and Anant Madabhushi. Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent. *Scientific reports*, 7(1):46450, 2017.

[34] Haibo Wang, Angel Cruz-Roa, Ajay Basavanhally, Hannah Gilmore, Natalie Shih, Mike Feldman, John Tomaszewski, Fabio Gonzalez, and Anant Madabhushi. Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. *Journal of Medical Imaging*, 1(3):034003–034003, 2014.

[35] Yousef Al-Kofahi, Wiem Lassoued, William Lee, and Badrinath Roysam. Improved automatic detection and segmentation of cell nuclei in histopathology images. *IEEE Transactions on Biomedical Engineering*, 57(4):841–852, 2009.

[36] Yun Liu, Krishna Gadepalli, Mohammad Norouzi, George E Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan, Aleksei Timofeev, Philip Q Nelson, Greg S Corrado, et al. Detecting cancer metastases on gigapixel pathology images. *arXiv preprint arXiv:1703.02442*, 2017.

[37] George Lee, Rachel Sparks, Sahirzeeshan Ali, Natalie NC Shih, Michael D Feldman, Elaine Spangler, Timothy Rebbeck, John E Tomaszewski, and Anant Madabhushi. Co-occurring gland angularity in localized subgraphs: predicting biochemical recurrence in intermediate-risk prostate cancer patients. *PloS one*, 9(5):e97954, 2014.

[38] Xiangxue Wang, Andrew Janowczyk, Yu Zhou, Rajat Thawani, Pingfu Fu, Kurt Schalper, Vamsidhar Velcheti, and Anant Madabhushi. Prediction of recurrence in early stage non-small cell lung cancer using computer extracted nuclear features from digital h&e images. *Scientific reports*, 7(1):13543, 2017.

[39] Frederick M Howard, James Dolezal, Sara Kochanny, Galina Khramtsova, Jasmine Vickery, Andrew Srisuwananukorn, Anna Woodard, Nan Chen, Rita Nanda, Charles M Perou, et al. Integration of clinical features and deep learning on pathology for the prediction of breast cancer recurrence assays and risk of recurrence. *NPJ Breast Cancer*, 9 (1):25, 2023.

[40] Kimmo Kartasalo, Wouter Bulten, Brett Delahunt, Po-Hsuan Cameron Chen, Hans Pinckaers, Henrik Olsson, Xiaoyi Ji, Nita Mulliqi, Hemamali Samaratunga, Toyonori Tsuzuki, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer in biopsies—current status and next steps. *European Urology Focus*, 7(4):687–691, 2021.

[41] Rose S. George et al. Artificial intelligence in prostate cancer: Definitions, current research, and future directions. *Urologic Oncology: Seminars and Original Investigations*, 40(6):262–270, 2022. ISSN 1078-1439. doi: https://doi.org/10.1016/j.urolonc.2022.03.003. URL `https://www.sciencedirect.com/science/article/pii/S1078143922000771`.

[42] Marit Lucas et al. Deep learning for automatic gleason pattern classification for grade group determination of prostate biopsies. *Virchows Archiv*, 475(1):77–83, Jul 2019. ISSN 1432-2307. doi: 10.1007/s00428-019-02577-x. URL `https://doi.org/10.1007/s00428-019-02577-x`.

[43] Peter Ström et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *Lancet Oncol*, 21(2):222–232, January 2020.

[44] Gabriele Campanella et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, Aug 2019. ISSN 1546-170X. doi: 10.1038/s41591-019-0508-1. URL `https://doi.org/10.1038/s41591-019-0508-1`.

[45] Gabriele Campanella, Vitor Werneck Krauss Silva, and Thomas J Fuchs. Terabyte-scale deep multiple instance learning for classification and localization in pathology. *arXiv preprint arXiv:1805.06983*, 2018.

[46] Hans Pinckaers, Wouter Bulten, Jeroen van der Laak, and Geert Litjens. Detection of prostate cancer in Whole-Slide images through End-to-End training with Image-Level labels. *IEEE Trans Med Imaging*, 40(7):1817–1826, June 2021.

[47] Kunal Nagpal et al. Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. *npj Digital Medicine*, 2(1):48, Jun 2019. ISSN 2398-6352. doi: 10.1038/s41746-019-0112-2. URL `https://doi.org/10.1038/s41746-019-0112-2`.

[48] John N Weinstein et al. The cancer genome atlas Pan-Cancer analysis project. *Nat. Genet.*, 45(10):1113–1120, October 2013.

[49] Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob van de Loo, Rob Vogels, Quirine F Manson, Nikolas Stathonikos, Alexi Baidoshvili, Paul van Diest, Carla Wauters, Marcory van Dijk, and Jeroen van der Laak. 1399 H
amp;E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience*, 7(6), 05 2018. ISSN 2047-217X. doi: 10.1093/gigascience/giy065. URL `https://doi.org/10.1093/gigascience/giy065`. giy065.

[50] Péter Bándi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, Quanzheng Li, Farhad Ghazvinian Zanjani, Svitlana Zinger, Keisuke Fukuta, Daisuke Komura, Vlado Ovtcharov, Shenghua Cheng, Shaoqun Zeng, Jeppe Thagaard, Anders B. Dahl, Huangjing Lin, Hao Chen, Ludwig Jacobsson, Martin Hedlund, Melih Çetin, Eren Halıcı, Hunter Jackson, Richard Chen, Fabian Both, Jörg Franke, Heidi Küsters-Vandevelde, Willem Vreuls, Peter Bult, Bram van Ginneken, Jeroen van der Laak, and Geert Litjens. From detection of individual metastases to classification of lymph node status at the patient level: The camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2019. doi: 10.1109/TMI.2018.2867350.

[51] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. A general-purpose self-supervised model for computational pathology. *arXiv preprint arXiv:2308.15474*, 2023.

[52] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. Explainable deep learning models in medical image analysis. *Journal of imaging*, 6(6):52, 2020.

[53] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: Deep learning for interpretable image recognition. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, ed-

itors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/adf7ee2dcf142b0e11888e72b43fcb75-Paper.pdf`.

[54] Shancheng Jiang, Huichuan Li, and Zhi Jin. A visually interpretable deep learning framework for histopathological image-based skin cancer diagnosis. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1483–1494, 2021.

[55] Jie Hao, Sai Chandra Kosaraju, Nelson Zange Tsaku, Dae Hyun Song, and Mingon Kang. Page-net: interpretable and integrative deep learning for survival analysis using histopathological images and genomic data. In *Pacific Symposium on Biocomputing 2020*, pages 355–366. World Scientific, 2019.

[56] Guangli Li, Chuanxiu Li, Guangting Wu, Donghong Ji, and Hongbin Zhang. Multi-view attention-guided multiple instance detection network for interpretable breast cancer histopathological image diagnosis. *IEEE Access*, 9:79671–79684, 2021.

[57] Soufiane Belharbi, Jérôme Rony, Jose Dolz, Ismail Ben Ayed, Luke McCaffrey, and Eric Granger. Deep interpretable classification and weakly-supervised segmentation of histology images via max-min uncertainty. *IEEE Transactions on Medical Imaging*, 41 (3):702–714, 2021.

[58] Angel Alfonso Cruz-Roa, John Edison Arevalo Ovalle, Anant Madabhushi, and Fabio Augusto González Osorio. A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II 16*, pages 403–410. Springer, 2013.

[59] Gang Xu, Zhigang Song, Zhuo Sun, Calvin Ku, Zhe Yang, Cancheng Liu, Shuhao Wang, Jianpeng Ma, and Wei Xu. Camel: A weakly supervised learning framework for histopathology image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[60] Zohaib Salahuddin, Henry C. Woodruff, Avishek Chatterjee, and Philippe Lambin. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in Biology and Medicine*, 140:105111, 2022. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2021.105111. URL `https://www.sciencedirect.com/science/article/pii/S0010482521009057`.

[61] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[62] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.

[63] Syed Ashar Javed, Dinkar Juyal, Harshith Padigela, Amaro Taylor-Weiner, Limin Yu, and Aaditya Prakash. Additive mil: intrinsically interpretable multiple instance learning for pathology. *Advances in Neural Information Processing Systems*, 35:20689–20702, 2022.

[64] Santiago Toledo-Cortés, Diego H. Useche, Henning Müller, and Fabio A. González. Grading diabetic retinopathy and prostate cancer diagnostic images with deep quantum ordinal regression. *Computers in Biology and Medicine*, 145:105472, 2022. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2022.105472. URL `https://www.sciencedirect.com/science/article/pii/S0010482522002645`.

[65] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, 2010.

[66] Fabio A. González, Raúl Ramos-Pollán, and Joseph A. Gallego-Mejia. Quantum kernel mixtures for probabilistic deep learning, 2023.

[67] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *CoRR*, abs/2201.03545, 2022. URL `https://arxiv.org/abs/2201.03545`.

[68] Mingxing Tan and Quoc V. Le. Efficientnetv2: Smaller models and faster training, 2021.

[69] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *CoRR*, abs/2106.02034, 2021. URL `https://arxiv.org/abs/2106.02034`.

[70] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[71] Hemamali Samaratunga, Lars Egevad, John Yaxley, Joanna Perry-Keene, Ian Le Fevre, James Kench, Admire Matsika, David Bostwick, Kenneth Iczkowski, and Brett Delahunt. Gleason score 3+3=6 prostatic adenocarcinoma is not benign and the current debate is unhelpful to clinicians and patients. *Pathology*, 2023. ISSN 0031-3025. doi: https://doi.org/10.1016/j.pathol.2023.10.005. URL `https://www.sciencedirect.com/science/article/pii/S0031302523002945`.

[72] Shiwen Shen, Simon X Han, Denise R Aberle, Alex A Bui, and William Hsu. An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. *Expert Systems with Applications*, 128:84–95, 2019.

ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2019.01.048. URL `https://www.sciencedirect.com/science/article/pii/S0957417419300545`.

[73] Eunji Kim, Siwon Kim, Minji Seo, and Sungroh Yoon. Xprotonet: Diagnosis in chest radiography with global and local explanations. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15714–15723, 2021. doi: 10.1109/CVPR46437.2021.01546.

[74] Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

[75] Cher Bass, Mariana da Silva, Carole Sudre, Logan ZJ Williams, Petru-Daniel Tudosiu, Fidel Alfaro-Almagro, Sean P Fitzgibbon, Matthew F Glasser, Stephen M Smith, and Emma C Robinson. Icam-reg: Interpretable classification and regression with feature attribution for mapping neurological phenotypes in individual scans. *arXiv preprint arXiv:2103.02561*, 2021.

[76] Christian F. Baumgartner, Lisa M. Koch, Kerem Can Tezcan, Jia Xi Ang, and Ender Konukoglu. Visual feature attribution using wasserstein gans. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8309–8319, 2018. doi: 10.1109/CVPR.2018.00867.

[77] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. Explainable deep learning models in medical image analysis. *J. Imaging*, 6(6):52, June 2020.

[78] Joona Pohjonen and Valeria Ariotta. Histoprep: Preprocessing large medical images for machine learning made easy! https://github.com/jopo666/HistoPrep, 2022.