Probabilistic Deep Learning for Histopathological Images: Overcoming the Labeling Bottleneck of Computer-Aided Diagnosis.

Presented to obtain the degree of

Doctor of Philosophy

by

Arne Schmidt

supervised by

Rafael Molina Soriano and Pablo Morales-Álvarez

UNIVERSIDAD
DE GRANADA

Doctoral Programme in Information and Communication Technologies

Editor: Universidad de Granada. Tesis Doctorales
Autor: Arne Schimdt
ISBN: 978-84-1195-244-6
URI: https://hdl.handle.net/10481/90746

Und sie laufen! Naß und nässer
wirds im Saal und auf den Stufen.
Welch entsetzliches Gewässer!
Herr und Meister! hör mich rufen! –
Ach, da kommt der Meister!
Herr, die Not ist groß!
Die ich rief, die Geister
werd ich nun nicht los.

Der Zauberlehrling
(Johann Wolfgang von Goethe)

Agradecimientos/Acknowledgements/Danksagungen

Me gustaría dar las gracias, en primer lugar, a mis supervisores Rafael Molina y Pablo Morales por todos los consejos, revisiones, orientaciones, seminarios e ideas durante estos tres años. Estoy muy agradecido por todo lo que he podido aprender en este tiempo y estoy seguro que vamos a seguir teniendo debates inspiradores en el futuro. Aparte del trabajo, me voy con bonitos recuerdos como las pausas de café o las salidas a la Herradura y al Sacromonte. Esta amable acogida hizo que me sintiera a gusto en Granada desde el principio e incluso me hace pasar por alto el hecho de que aún no sepáis pronunciar mi nombre. Broma aparte, muchas gracias por todo.

A continuación, me gustaría mencionar a mis colegas del CITIC. Ha sido un placer trabajar con vosotros y compartir todos estos momentos dentro y fuera de la oficina. ¡Viva Spicy Gauss!

Gracias también a todos los amigos y amigas que he conocido en Granada y Valencia. Fue increíble compartir este camino con vosotros y me llevo muchos recuerdos bonitos. Especialmente quiero agradecer a Gera por todo el apoyo y por formar esta parte importante de mi vida. Aunque en este contexto más bien debería agradecer a mi doctorado - por permitirme conocerla a ella.

I would also like to thank all the great people that I got to know in the Clarify project. Throughout this journey, I crossed paths with many kind-hearted, ambitious, and intelligent individuals. Engaging in discussions and collaborations with them has been an absolute pleasure. A special thanks to everyone hosting me during my secondments - at the UiS, NU, UPV, Tyris, and INCLIVA.

Zu guter Letzt möchte ich meiner Familie danken und all meinen alten Freunden aus Deutschland. Wo man herkommt, wie man aufwächst und alle Menschen, die einem auf dem Lebensweg begleiten, haben einen grossen Einfluss darauf, wer man ist und was man erreicht. Vielen Dank dafür, dass ich immer auf euch zählen kann.

Abstract

In the medical domain, there is an increasing need for artificial intelligence models that can improve the reliability, reproducibility, and efficiency of diagnostic processes. However, acquiring large labeled datasets for training these models poses a significant challenge in comparison to other domains, resulting in a bottleneck in the development of computer-aided diagnostic systems. To address this problem, this thesis investigates various learning paradigms that enable training with limited or imperfect annotations for histopathological images: multiple instance learning, active learning, and crowdsourcing. Notably, these paradigms involve uncertainties arising from missing or imperfect information which must be taken into account.

This thesis introduces novel probabilistic deep learning models that effectively address these uncertainties in a principled way by leveraging probability theory. They offer improved performance and provide probabilistic outputs, enabling the estimation of the confidence level associated with model predictions. The proposed models are based on Gaussian processes, Bayesian neural networks, and probabilistic generative models, tailored to each labeling paradigm and corresponding uncertainties. We establish the theoretical foundations of these models and demonstrate their practical usefulness through extensive experiments conducted on various publicly available datasets. Our findings demonstrate promising performance in histopathological image analysis, offering reliable clinical decision support even in scenarios with limited data availability. By contributing to the advancement of computer-aided diagnosis systems, the proposed models can enhance the quality of diagnostic processes, which ultimately allows an improved treatment of patients.

Resumen

En el ámbito médico, existe una necesidad creciente de modelos de inteligencia artificial que puedan mejorar la fiabilidad, reproducibilidad y eficiencia de los procesos de diagnóstico. Sin embargo, la adquisición de grandes conjuntos de datos etiquetados para el entrenamiento de estos modelos plantea un reto importante en comparación con otros dominios, lo que supone un desafío en el desarrollo de sistemas de diagnóstico asistido por ordenador. Para abordar este problema, esta tesis investiga varios paradigmas de aprendizaje que permiten el entrenamiento con anotaciones limitadas o imperfectas para imágenes histopatológicas: aprendizaje con múltiples instancias, aprendizaje activo y crowdsourcing. En particular, estos paradigmas implican incertidumbres derivadas de la falta de información o de información imperfecta que deben tenerse en cuenta.

Esta tesis introduce nuevos modelos de aprendizaje profundo probabilístico que abordan eficazmente estas incertidumbres basadas en principios de la teoría de la probabilidad. Ofrecen un rendimiento mejorado y proporcionan salidas probabilísticas, lo que permite estimar el nivel de confianza asociado a las predicciones del modelo. Los modelos propuestos se basan en procesos gaussianos, redes neuronales bayesianas y modelos generativos probabilísticos, adaptados a cada paradigma de etiquetado y a las incertidumbres correspondientes. Establecemos los fundamentos teóricos de estos modelos y demostramos su utilidad práctica mediante amplios experimentos realizados en bases de datos públicas. Nuestros experimentos demuestran un rendimiento prometedor en el análisis de imágenes histopatológicas, ofreciendo un apoyo fiable a la toma de decisiones clínicas incluso en escenarios con disponibilidad limitada de datos. Al contribuir al avance de los sistemas de diagnóstico asistido por ordenador, los modelos propuestos pueden mejorar la calidad de los procesos de diagnóstico y mejorar el tratamiento de los pacientes.

Contents

Resumen extendido en castellano

Introducción

La Inteligencia Artificial (IA) ha atraído mucha atención en los últimos años debido a su enorme impacto en la tecnología y la sociedad. Aparte de los riesgos potenciales, que exigen debates y normativas exhaustivas, existe un inmenso potencial para encontrar soluciones que hacen nuestra vida más segura, fácil y sana. En el ámbito de la visión por ordenador, los enfoques basados en la IA han progresado gracias al rápido avance de la investigación, permitiendo la detección fiable de peatones en los coches autoconducidos o la detección automatizada de enfermedades en los cultivos agrícolas, inventos técnicamente imposibles hace tan solo unos años [1]. Este éxito se basa en las grandes capacidades de reconocimiento de patrones de los modelos de aprendizaje profundo, que son iguales o incluso mejores que la percepción humana en algunas aplicaciones [1; 2; 3].

En el ámbito médico, los procesos de diagnóstico suelen incluir el reconocimiento de patrones por parte de expertos humanos. En histopatología, esto significa que los patólogos interpretan patrones celulares para diagnosticar patologías, por ejemplo, si un determinado tumor es benigno o maligno. Sin embargo, este proceso está sujeto a una importante variabilidad intraobservador e interobservador, lo que conduce a diagnósticos erróneos y a consecuencias potencialmente graves [4; 5; 6]. De hecho, hay estudios que estiman que los errores médicos son una de las principales causas de muerte en los países desarrollados y que los diagnósticos erróneos son uno de los principales problemas de las prácticas clínicas actuales [7; 8]. Además, la situación en los países en desarrollo, con escasez de expertos médicos, es aún más grave [9]. Esta urgente necesidad de mejorar las prácticas actuales motiva la investigación científica que se presenta en esta tesis.

El diagnóstico asistido por ordenador con IA ha surgido como un enfoque prometedor para hacer que los procesos de diagnóstico sean más precisos, reproducibles y eficientes. El objetivo es ayudar a los expertos médicos con predicciones de IA para garantizar el tratamiento correcto. Las mismas arquitecturas de aprendizaje profundo que reconocen rostros o señales de tráfico en otras aplicaciones con gran éxito pueden entrenarse para reconocer hemorragias en tomografías cerebrales o tejido canceroso en imágenes histopatológicas. Sin embargo, hay una diferencia importante en comparación con otras áreas: Los grandes conjuntos de datos etiquetados son muy difíciles de obtener en el ámbito médico porque requieren conocimientos especiales. Dado que los métodos habituales de IA supervisada suelen contar con miles o incluso millones de imágenes etiquetadas para alcanzar el rendimiento deseado, esto supone un gran reto [10].

En este contexto, esta tesis explora métodos novedosos para superar este desafío del etiquetado en el diagnóstico asistido por ordenador con imágenes histopatológicas. Desarrollamos nuevos modelos que pueden aprovechar etiquetas limitadas e imperfectas para

mejorar la precisión y robustez del diagnóstico asistido por ordenador, minimizando al mismo tiempo la necesidad de anotación manual. Concretamente, los métodos propuestos en esta tesis pueden clasificarse en tres paradigmas de aprendizaje diferentes:

- **Aprendizaje con múltiples instancias** (multiple instance learning) describe un escenario en el que varias instancias se agrupan en bolsas y sólo las etiquetas de las bolsas se utilizan para el entrenamiento. Por lo tanto, el etiquetado de las instancias no es necesario. En el aprendizaje clásico de instancias múltiples, la bolsa es positiva si hay al menos una instancia positiva en ella. De lo contrario, la bolsa es negativa. En histopatología, la biopsia completa de un paciente forma la bolsa y sus parches de imagen las instancias. Si un parche de imagen muestra patrones cancerosos (=positivo), se obtiene un diagnóstico global positivo de cáncer para la biopsia. Si todo el tejido es sano, el diagnóstico global es negativo. En el aprendizaje con múltiples instancias, las etiquetas de instancia que faltan introducen incertidumbre en el entrenamiento y la predicción. La información sobre qué parches de la imagen son cancerosos y, por tanto, dan lugar a un diagnóstico global positivo de cáncer, no está disponible en los datos. Por lo tanto, la estimación de la información a nivel de parche puede modelarse con métodos probabilísticos, como se muestra en los capítulos 2 y 3. En el capítulo 4 se presenta otro enfoque, que combina el aprendizaje semisupervisado y el aprendizaje con múltiples instancias.

- **Aprendizaje activo** reduce la demanda de etiquetado, etiquetando sólo los parches más informativos de los datos, elegidos por el propio modelo entrenado. En primer lugar, el modelo se entrena únicamente con un pequeño conjunto de datos etiquetados. A continuación, accedede a un conjunto de datos sin etiquetar y elige un subconjunto de imágenes para que las etiquete un especialista, como un patólogo en nuestra aplicación. Las imágenes con mayor valor informativo se etiquetan y se añaden al conjunto de entrenamiento. A continuación, se vuelve a entrenar el modelo. Este paso de adquisición se repite de forma iterativa, de modo que el rendimiento del modelo aumenta a medida que se agregan más y más datos etiquetados. En el campo de la patología, el aprendizaje activo es un método eficaz para recopilar anotaciones de tejidos locales. Sin embargo, existen varios retos específicos de imágenes histopatológicas. Por ejemplo, hay muchos parches de imagen repetitivos con tejido sano y artefactos como marcas de bolígrafo, tinta o regiones borrosas que no aportan información sustancial para el entrenamiento de un modelo de IA. Por lo tanto, el objetivo de nuestra investigación es centrarnos en la anotación de tejido canceroso y adquirir los parches de tejido correspondientes para la anotación.

  Existen distintas fuentes de incertidumbre que deben tenerse en cuenta para la adquisición de nuevos datos y para las predicciones finales del modelo. En primer lugar, la incertidumbre en los parámetros del modelo debida a la limitación de los datos de

entrenamiento. Esta incertidumbre se puede utilizar para medir la informatividad de las nuevas imágenes. El objetico para la adquisición es eligir imágenes de alta informatividad. Por el contrario, hay que evitar imágenes con incertidumbre de datos para no introducir ruido en el entrenamiento del modelo. Además, los datos fuera de distribución, como los parches de imágenes que contienen artefactos u otras anomalías, son otra fuente de incertidumbre. También deben evitarse durante la adquisición porque no aportan información adicional. Las incertidumbres del aprendizaje activo se abordan en nuestro trabajo del capítulo 5.

- **Crowdsourcing** ayuda a superar el desafío del etiquetado entrenando con etiquetas imperfectas de múltiples anotadores. Esto permite incluir a personas no expertas en el proceso de etiquetado, lo que puede dar lugar a un mayor número de anotadores. En el ámbito de las imágenes histopatológicas, en el que la variabilidad entre observadores y dentro de los mismos es elevada, el crowdsourcing resulta especialmente valioso. A diferencia de otros dominios con clases claramente definidas (como çoche.^{o} ”árbol.^{en} fotografías), las imágenes histopatológicas carecen a menudo de una única verdad básica, ya que incluso los expertos médicos pueden discrepar en algunos casos. En este paradigma de aprendizaje, la incertidumbre surge del comportamiento de etiquetado dependiente del anotador. Cada anotador tiene diferentes áreas de especialización y un nivel de experiencia distinto, lo que provoca variaciones en la anotación de imágenes médicas. Estas variaciones pueden capturarse mediante modelado probabilístico, como se muestra en el Capítulo 6.

Objetivos y estructura de la tesis

El objetivo principal de esta tesis es desarrollar nuevos métodos probabilísticos de aprendizaje profundo para superar el desafío del etiquetado en histopatología. En los marcos de los diferentes paradigmas de aprendizaje, surgen diferentes fuentes de incertidumbre que deben ser abordadas mediante un modelado probabilístico adecuado. A continuación, describimos los objetivos de las distintas áreas de investigación.

- Desarrollar nuevos algoritmos de aprendizaje con múltiples instancias modelando las etiquetas de instancia desconocidas como una fuente de incertidumbre. Existen diferentes enfoques para entrenar un modelo de aprendizaje profundo en este entorno. Nuestro objetivo es mejorar estos enfoques adoptando una perspectiva probabilística de los retos actuales y aplicando procesos gaussianos. Concretamente, existen dos líneas de investigación. (i) Muchos de los métodos de aprendizaje con múltiples instancias existentes se basan en mecanismos de atención que también se utilizan en las arquitecturas de 'transformers' [11]. En estos enfoques, la importancia de cada instancia para la predicción de la bolsa se estima mediante pesos de atención. Los métodos probabilísticos pueden ayudar a capturar las incertidumbres de los pesos de

atención y proporcionar predicciones que tengan en cuenta estas incertidumbres. Esta línea de investigación se exploró en el capítulo 2, utilizando un mecanismo de atención basado en un proceso gaussiano. (ii) Otra línea de investigación es el modelado probabilístico de las correlaciones de instancia. Como en las imágenes histopatológicas, regiones (parches) vecinas en una imagen están altamente correlacionados, esto debería reflejarse en el modelo. Dado que muchos de los enfoques existentes asumen que las instancias son independientes, un modelado probabilístico adecuado de las correlaciones entre instancias puede mejorar considerablemente las predicciones. En el capítulo 3 investigamos un marco probabilístico basado en procesos gaussianos, que tiene en cuenta las correlaciones de instancia.

- Abordar los retos relacionados con los datos en el aprendizaje activo con modelos probabilísticos. Las imágenes histopatológicas plantean grandes retos debido a las ambigüedades y artefactos de los datos. Los algoritmos de aprendizaje activo existentes asignan erróneamente una alta importancia a estas imágenes. Esto conduce su adquisición y etiquetado, aunque el valor de imágenes con ambigüedades y artefactos para el entrenamiento sea bajo. Las redes neuronales bayesianas que proporcionan incertidumbres en la salida pueden ayudar a evitar la adquisición de estas imágenes y centrarse en las imágenes con tejido canceroso. Por tanto, los nuevos modelos probabilísticos pueden mejorar el aprendizaje activo para imágenes histopatológicas y se presentan en el capítulo 5.

- Desarrollar nuevos algoritmos de crowdsourcing probabilístico que aborden la incertidumbre de las etiquetas de diferentes anotadores. En el contexto de la histopatología, donde la variabilidad intra e interobservador de las anotaciones es prominente, la ausencia de una ”verdad básica”definitiva plantea retos significativos tanto durante el entrenamiento del modelo como en las predicciones dentro de los marcos de aprendizaje profundo. En consecuencia, un objetivo clave de esta tesis doctoral es desarrollar nuevos algoritmos probabilísticos de crowdsourcing que aborden eficazmente la incertidumbre inherente de las etiquetas introducida por diferentes anotadores. Durante el proceso de entrenamiento, es importante tener en cuenta las incertidumbres asociadas a cada etiqueta, causadas por la evaluación subjetiva. Los modelos generativos probabilísticos permiten representar y tratar eficazmente estas incertidumbres. En el capítulo 6 proponemos un modelo de este tipo que tiene en cuenta las variaciones introducidas por múltiples anotadores y realiza predicciones que reflejan las posibles ambigüedades.

Los principales capítulos de la tesis se estructuran de la siguiente manera. El capítulo 1 contiene la introducción. Los capítulos 2-4 contienen las contribuciones científicas de los novedosos modelos de aprendizaje con múltiples instancias. Exploramos un mecanismo de atención probabilística con procesos gaussianos (capítulo 2), correlaciones de instancia

para un clasificador de procesos gaussianos (capítulo 3), y la combinación de aprendizaje con múltiples instancias y aprendizaje semisupervisado (capítulo 4). En el Capítulo 5 proponemos un algoritmo de aprendizaje activo basado en redes neuronales bayesianas para adquirir imágenes informativas de forma más focalizada. El capítulo 6 presenta un modelo generativo probabilístico para la segmentación de imágenes que aborda la variabilidad interobservador e intraobservador en las anotaciones médicas. En el capítulo 7 se presentan otros trabajos científicos relevantes con contribuciones significativas del candidato. Dos de esos artículos incluyen procesos gaussianos aplicados al problema de las tomografías computerizadas (ÇT scans"), mientras que un tercero explora la optimización restringida para mejorar el aprendizaje con múltiples instancias con WSI. Un cuarto artículo sobre el preprocesamiento de WSI está estrechamente relacionado con el Capítulo 5 porque explica artefactos como el desenfoque, la tinta o el rotulador, que son relevantes para el aprendizaje activo. Concluimos la tesis en el capítulo 8.

Conclusiones

En esta tesis doctoral hemos desarrollado métodos probabilísticos para el aprendizaje con múltiples instancias, el aprendizaje activo y el crowdsourcing que en nuestra opinión dan lugar a prometedores avances científicos y a resultados experimentales competitivos. Descubrimos que, para cada paradigma de aprendizaje, podíamos superar con éxito los retos existentes. Antes de presentar el trabajo científico detallado, queremos destacar algunas conclusiones generales de los artículos presentados.

Métodos novedosos para el aprendizaje con múltiples instancias, el aprendizaje activo y el crowdsourcing permiten un alto rendimiento con un entrenamiento eficiente de etiquetas. En los distintos escenarios de aprendizaje, logramos resultados competitivos utilizando menos etiquetas o etiquetas imperfectas para el entrenamiento del modelo. Incluso para conjuntos de datos pequeños, los modelos probabilísticos desarrollados fueron capaces de generalizar bien a datos no vistos. En comparación con los métodos supervisados (que se basan en el etiquetado exhaustivo de todos los datos), la diferencia de rendimiento con los avances de la investigación está desapareciendo, lo que podría hacer que el etiquetado exhaustivo quede obsoleto en el futuro. Esto es importante para todos los numerosos tipos de cáncer, tareas de clasificación y casos de uso especiales para los que aún no se dispone de grandes conjuntos de datos ampliamente etiquetados. Los métodos propuestos reducen considerablemente los recursos necesarios para entrenar modelos de IA para nuevas tareas. Elegir el modelo probabilístico adecuado para un problema específico puede mejorar el estado del arte. En los distintos artículos presentados, utilizamos diferentes modelos probabilísticos, adaptados a cada caso de uso.

Para el paradigma de aprendizaje con múltiples instancias, nos basamos en procesos gaussianos debido a su buena capacidad de regresión de funciones, su solidez frente al so-

breajuste y su integración matemáticamente sólida en marcos probabilísticos. En nuestro estudio del capitulo 2, los procesos gaussianos mostraron un rendimiento notable cuando se emplearon para la regresión del peso de la atención, superando a otros modelos existentes en tres experimentos distintos sobre el cáncer de próstata. En capitulo 3, los procesos gaussianos facilitaron la incorporación directa de correlaciones de instancia dentro del proceso de modelado probabilístico, como se puso de manifiesto en una investigación separada.

Las redes neuronales bayesianas son modelos eficientes para la clasificación multiclase y, por tanto, se utilizan para el aprendizaje activo con parches de imágenes en capitulo 5. Estos modelos son fáciles de entrenar, lo que constituye una ventaja importante en el aprendizaje activo, en el que el modelo tiene que converger en cada paso de adquisición con una cantidad creciente de datos. Además, ofrecen la posibilidad de estimar distintos tipos de incertidumbres (aleatoria y epistémica), lo que permite separar los parches de imagen informativos de los ambiguos. Los parches informativos se utilizan para el etiquetado, mientras que los ambiguos se evitan. El modelo propuesto fue capaz de demostrar esta ventaja en la práctica en comparación con otros métodos existentes, dando como resultado un rendimiento de vanguardia para el aprendizaje activo de parches de cáncer de próstata.

Los modelos generativos de segmentación tienen la ventaja de representar explícitamente la incertidumbre mediante una distribución de probabilidad. Los aplicamos en capitulo 6 a la segmentación de imágenes de cáncer de próstata con etiquetas de crowdsourcing. El nuevo modelo de segmentación probabilística fue capaz de capturar las incertidumbres de la variabilidad interobservador e intraobservador, lo que condujo a un mejor rendimiento y a una medida exacta de la incertidumbre en el resultado que otros métodos comparados.

Combinar métodos deterministas y probabilísticos es muy eficaz. No es necesario tener un modelo completamente probabilístico. En todos nuestros trabajos combinamos una red neuronal convolucional determinista con un modelo probabilístico para el razonamiento de alto nivel. Esto permite aprovechar la optimización eficiente y directa del extractor de características, mientras que la toma de decisiones la realiza el modelo probabilístico. Queremos insistir una vez más en que el modelo probabilístico concreto debe diseñarse en función del propósito. Para realizar la inferencia, a menudo optamos por la integración de Monte Carlo como una aproximación que es computacionalmente eficiente cuando se aplica en las últimas capas y permite una fácil combinación con los modelos deterministas y una implementación directa. Además, nos pareció muy beneficioso diseñar modelos que puedan implementarse en bibliotecas de aprendizaje profundo comunes (Tensorflow/Pytorch) para una compatibilidad total con el entrenamiento basado en GPU.

Captar la incertidumbre no sólo es beneficioso para el entrenamiento, sino que también puede aportar información valiosa para las predicciones. Los modelos probabilísticos propuestos mostraron un rendimiento mejorado, pero además tenían otra gran ventaja: la posibilidad de proporcionar una predicción probabilística que puede reflejar incertidumbres. Evaluar las incertidumbres es crucial en las aplicaciones médicas y demos-

tramos experimentalmente que las incertidumbres estimadas indicaban predicciones con un mayor riesgo de ser erróneas. Investigamos distintas fuentes de incertidumbre, como el mecanismo de atención en el aprendizaje con múltiples instancias, las ambigüedades de los datos, los parámetros del modelo o la variabilidad entre observadores, y demostramos que se reflejaban en la distribución predictiva. Esto aporta grandes ventajas para un uso seguro en la práctica clínica.

Chapter 1

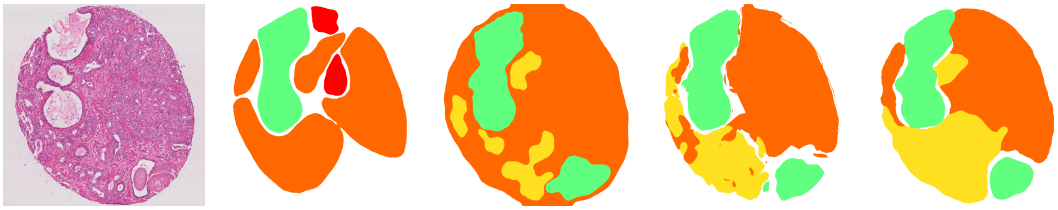
Introduction

Artificial Intelligence (AI) has attracted increasing attention in recent years due to its huge impact on technology and society. Apart from potential risks, which require thorough debates and regulations, there is an immense potential to find solutions that make our life safer, easier, and healthier. In the area of computer vision, AI-driven approaches have matured due to fast-progressing research, allowing reliable detection of pedestrians in self-driving cars or automated disease detection of crops in agriculture, inventions that have been technically impossible only a few years ago [1]. This success is based on the great pattern recognition capabilities of deep learning models which are equal or even better than human perception in some applications [2; 3; 1].

In the medical domain, diagnostic processes often include pattern recognition by human experts. In histopathology, this means that pathologists interpret cell patterns to determine pathologies, for example if a given tumor is benign or malign. However, this process is subject to significant intra- and inter-observer variability (as exemplary shown in Fig. 1.1), leading to misdiagnosis and potentially serious consequences [4; 5; 6]. In fact, studies estimated medical errors as a leading cause of death in developed countries with misdiagnosis identified as one major problem in current clinical practices [7; 8]. Moreover, the situation in developing countries with a shortage of medical experts is even more severe [9]. This serious need to improve the current practices motivates the scientific research presented in this thesis.

Computer-aided diagnosis with AI has emerged as a promising approach to make diagnostic processes more accurate, reproducible, and efficient. The goal is to assist medical experts with AI predictions (see Fig. 1.1) to ensure the correct treatment. The same (or similar) deep learning architectures that recognize faces or traffic signs in other applications with great success can be trained to recognize hemorrhage in CT brain scans or cancerous tissue in histopathological images. However, there is one major difference in comparison to other areas: Large labeled datasets are very hard to obtain in the medical domain because they require medical knowledge. As common supervised AI methods typically count on thousands or even millions of labeled images to reach the desired performance, this imposes a major challenge [10].

This thesis explores novel methods to overcome this labeling bottleneck in computer-



(a) Image (b) Junior Pathologist (c) Senior Pathologist (d) Majority Voting of 6 Pathologists (e) AI Prediction

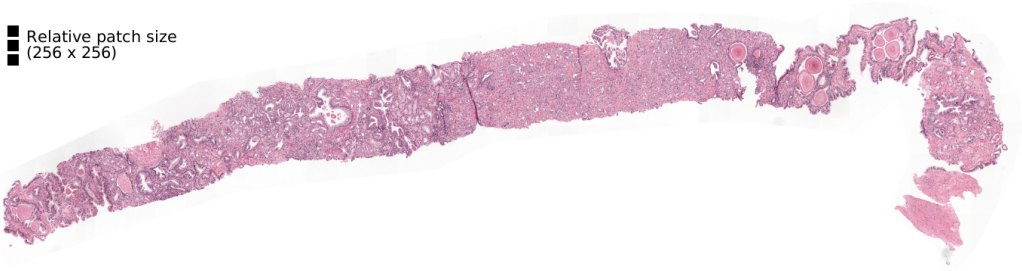
Figure 1.1: Example of the inter-observer variability in assessing prostate cancer tissue and potential improvement with AI predictions. The estimated classes are represented by colors: green for healthy tissue, yellow for Gleason Grade (GG) 3, orange for GG4, and red for GG5. The junior and senior pathologists show substantial differences in the estimated severeness of some tissue regions. The AI model is able to offer an accurate prediction (similar to the majority voting of 6 pathologists) and can help to standardize the diagnostic process.

aided diagnosis with histopathological images. We developed novel models that can leverage limited and imperfect labels to improve the accuracy and robustness of computer-aided diagnosis while minimizing the need for manual annotation. In this direction, the proposed methods of this thesis can be categorized into three different learning paradigms: (i) multiple instance learning, where instances are grouped into bags and only the bag label must be provided, (ii) active learning, where the model itself selects only the most informative samples to be iteratively labeled and (iii) crowdsourcing, where the imperfect labels of multiple (possibly non-expert) annotators can be used for training.

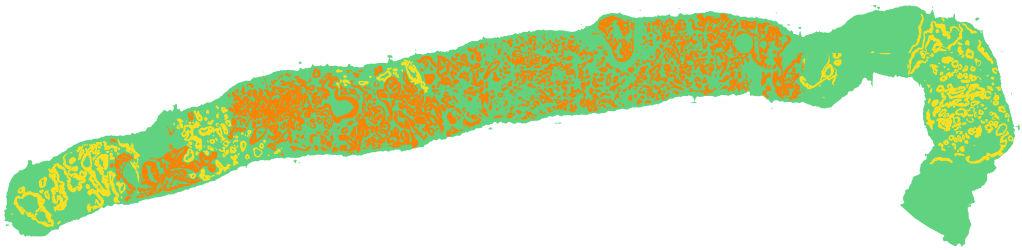
As these learning paradigms rely on limited or imperfect labels, there are different uncertainties that must be taken into account, as described in more detail in Section 1.2. To address them in a principled manner, we propose several novel probabilistic deep learning models to train and make informed decisions within these learning paradigms.

In all proposed models, the feature extraction from the images is performed by convolutional neural networks, which are (deterministic) deep learning methods. Based on the extracted features, we perform probabilistic, high-level reasoning based on Gaussian processes [12], Bayesian neural networks [13], and probabilistic generative models [14], depending on the application. We show that they provide accurate predictions and adequately capture the uncertainties.

In the following sections, we introduce the challenges and proposed solutions of this thesis. First, we introduce the background of histopathological image data in Section 1.1 and different learning paradigms in Section 1.2. These learning paradigms are the challenges we want to solve for the purpose of training with fewer or imperfect annotations. Section 1.3 outlines probabilistic models which are possible solutions to those challenges. Section 1.4 describes which probabilistic method is proposed for each learning paradigm,



(a) Whole Slide Image



(b) Annotation Mask

Figure 1.2: Example of a WSI (23,040 $\times$ 5888 pixels) for the classification of prostate cancer and its corresponding annotation mask. The relative patch size (256 $\times$ 256 pixels) is shown exemplary by three black squares in the upper left corner. The tissue was segmented into healthy (green), Gleason grade 3 (yellow) and Gleason grade 4 (orange) by a pathologist. The extensive local annotation as in this example is very time-consuming and represents a bottleneck for the training of deep learning models.

forming the objectives of the thesis. In Section 1.5 we outline the Methodology to validate the proposed methods. We present general results as well as details of the proposed models for each use-case in Section 1.6.

The main chapters of the thesis are structured as follows. Chapters 2-4 contain scientific contributions of novel multiple instance learning models. We explore a probabilistic attention mechanism with Gaussian processes (Chapter 2), instance correlations for a Gaussian process classifier (Chapter 3), and the combination of semi-supervised and multiple instance learning (Chapter 4). In Chapter 5 we propose an active learning algorithm based on Bayesian neural networks to acquire informative images in a more focused way. Chapter 6 presents a probabilistic generative model for image segmentation which addresses the inter- and intra-observer variability in medical annotations. Further relevant scientific work with significant contributions of the candidate is presented in Chapter 7. Two of those articles include Gaussian processes applied to the problem of computerized tomography (CT) scans, while a third one explores constrained optimization to improve multiple instance learning with WSIs. A fourth article about WSI preprocessing is closely related to Chapter 5 because it explains artifacts like blurr, ink, or pen marker which are relevant for active learning. We conclude the thesis in Chapter 8.

1.1 Histopathological Image Data

In the field of digital pathology, several data-specific particularities must be taken into account. Therefore, we will briefly describe the background of this data. To obtain a biopsy, human tissue is extracted, conserved, and prepared on a glass slide [15]. This slide is then scanned to obtain a digital image, the so-called Whole Slide Image (WSI). These WSIs are commonly very large with millions of pixels and several Gigabytes of size, with the exact size depending on the application and the scanner. To be able to process WSIs with deep learning methods, they are usually sliced into image patches with a lower resolution (e.g. 256x256), see Fig. 1.2a. Another image format is the Tissue Microarray (TMA), as shown in Fig. 1.1a. TMAs are extracted with a small needle and with an image size considerably lower than the size of WSIs. Therefore, they do not necessarily have to be sliced into image patches.

Labels of biopsies for deep learning can vary in levels of detail, reflecting the effort required for their acquisition. At the highest level, for WSIs (or TMAs) we have the global label, which can typically be derived directly from the patient's clinical records. Pathologists usually do not need to invest additional effort in assigning a global label. It serves as the foundation for diagnostic assessments and therefore its estimation is of high relevance. Moving to a more local level, we encounter patch labels and pixel labels. Patch labels pertain to the majority class of tissue present in a single image patch, offering information about the local tissue content. AI predictions on the patch level provide a moderate level of detail. On the most granular level, we have pixel labels, which are equal to a semantic segmentation of the complete image. To obtain them, annotation masks have to be drawn by the pathologist, as depicted in Figure 1.2b. This labeling approach offers the highest level of spatial detail, enabling precise identification and classification of tissue within the biopsy. Therefore, patch classification and segmentation are important to provide information on local patterns and regions of interest such as tumor nests.

1.2 Learning Paradigms

Having introduced the background of histopathological image data, now we move to the different learning paradigms that alleviate the burden of extensive manual labeling. For each case, a general explanation, mathematical formulation, the context of histopathological images and possible sources of uncertainties are provided.

Supervised Classification is a widely used standard method, where a model trains with one label for each image. Mathematically, we define a set of instances $\mathcal{X} = \{x_1, x_2, .., x_N\}$ and the corresponding set of labels as $\mathcal{Y} = \{y_1, y_2, .., y_N\}$. The model during training approximates the function $f : \mathcal{X} \rightarrow \mathcal{Y}$. For histopathological images, supervised classification is commonly employed to analyze image patches with labels that represent prevalent cell patterns within each image. As the datasets for supervised methods must be very large

to obtain satisfying results, we investigate alternatives that require less manual annotation for the model training. Although in the case of supervised learning, there is no additional uncertainty due to missing or imperfect annotations, there can be other general sources of uncertainty. Commonly, a general predictive uncertainty can be provided by probabilistic models, which describes the risk of a wrong prediction [13].

Multiple instance learning describes a setting where several instances are grouped into bags and only bag labels are used for training. Therefore, the labeling of the instances becomes obsolete. In mathematical terms, a bag of instances is defined as $X_b = \{x_{b1}, x_{b2}, .., x_{bN_b}\}$, with N_b describing the number of instances in each bag b [16]. In the classical multiple instance learning setting, the bag label T_b is given by

$$T_b = 0 \;\Leftrightarrow\; \forall i = 1, ..., N_b,\; y_{bi} = 0, \tag{1.1}$$

$$T_b = 1 \;\Leftrightarrow\; \exists i \in \{1, ..., N_b\} : y_{bi} = 1. \tag{1.2}$$

with the (binary) labels $y_{bi} \in \{0, 1\}$ that remain unknown.

In multiple instance learning with histopathological images, a WSI is considered a bag X_b and its diagnosis as the bag label T_b. The image patches represent the instances. As this learning paradigm only requires bag labels for training, the local annotation of the patches by expert pathologists is not necessary. In the common case, that the bag label can be derived directly from clinical records, the labeling cost can even reduce to zero. Apart from the binary classification (cancerous vs. non-cancerous), we are also interested in the cancer class of the WSI. This class can be determined based on the features of the cancerous patches. In this setting, equation (1.1) still applies for non-cancerous (negative) WSIs, while for cancerous WSIs we want to specify the class $T_b = c$, with $c \in \{c_1, c_2, .., c_K\}$ representing each one of the possible K cancer classes based on the cancerous areas.

In multiple instance learning, the missing instance labels introduce uncertainty in training and predicting. The information about which image patches are cancerous and therefore resulting in an overall cancer-positive diagnostic, is not available in the data. The estimation of patch-level information can therefore be modeled with probabilistic methods, as shown in Chapters 2 and 3.

Active learning reduces the labeling demand by labeling only the most informative image patches of the data, chosen by the trained model itself [17]. Firstly, the model is trained only on a small set of labeled data $\mathcal{D}_{train} = \{x_i, y_i\}_{i=1,..,N}$ of image patches x_i and labels y_i. Then, it can access a pool of unlabeled data $\mathcal{D}_{pool}$ and chooses a subset $\mathcal{A} \subset \mathcal{D}_{pool}$ of unlabeled images to be labeled by a specialist, such as a pathologist in the given application. The choice is made with the help of an acquisition function $a(x, \mathcal{M})$ which estimates the informativeness of each image x, where $\mathcal{M}$ represents the deep learning model. The images with the highest acquisition scores are labeled and added to $\mathcal{D}_{train}$. Then, the model is retrained with the updated training set. This acquisition step is repeated

iteratively such that the model performance increases while more and more labeled data is aggregated. The overall amount of required labels can be reduced significantly because only the most informative samples are chosen to be labeled until the desired performance is reached.

In the field of pathology, active learning is an effective method to gather local tissue annotations. However, there are several data-specific challenges. In WSIs, there are many healthy, repetitive tissue regions and artifacts like pen markers, ink, or blurred regions that do not contribute substantial information for training an AI model. Therefore, the goal of our research is to focus on the annotation of informative, cancerous tissue and acquire the corresponding tissue patches for annotation.

There are different sources of uncertainty that must be taken into account for the acquisition of new data and for the final predictions of the model. First of all, model uncertainty, also called epistemic uncertainty, is an important entity. It denotes the uncertainty arising from the lack of labeled data for specific inputs, making it a crucial factor in selecting informative images for labeling during the active learning acquisition phase. Subsequently, as more training data is incorporated, model uncertainty diminishes iteratively, leading to improved model performance.

Conversely, data uncertainty, or aleatoric uncertainty, represents inherent uncertainties present in the data due to ambiguities. Ambiguous data contributes limited information and cannot be reduced by acquiring more labeled samples. Consequently, during the active learning process, images with high data uncertainty should be avoided to prevent introducing noise into the model training.

Additionally, out-of-distribution data introduces another type of uncertainty. These instances, such as image patches containing artifacts or other anomalies, fail to provide valuable information for model training. Thus, excluding such out-of-distribution data from the acquisition step aids in maintaining the overall robustness and generalization capabilities of the active learning model. The uncertainties of active learning are addressed by our work in Chapter 5.

Crowdsourcing helps to overcome the labeling bottleneck by training with imperfect labels from multiple annotators. Mathematically, the crowdsourcing setting can be described as training a model on images $X = \{x_1, x_2, .., x_N\}$ with several annotation sets $Y^r = \{y_1^r, y_2^r, .., y_N^r\}$ provided by different raters $r = \{1, 2, .., R\}$. Some or all images can be annotated by multiple raters, such that some labels y_i^r can be empty. To assess the quality of the crowdsourcing model, a limited set of 'gold labels' is typically employed, which represents the consensus among experts. These 'gold labels' serve the purpose of validating the algorithm and, in some cases, are also used for training. Obtaining these 'gold labels' can be accomplished through discussions or iterative corrections until an agreement is reached. Alternatively, label fusion techniques such as majority voting can be applied to images with multiple expert annotations to estimate agreement.

In the domain of histopathological images, where inter- and intra-observer variability is high [5; 4; 6], crowdsourcing is particularly valuable. Unlike other domains with clearly defined classes like ”car” or ”tree,” histopathological images often lack a single ground truth, as even medical experts can disagree in some cases. Crowdsourcing models, by training with multiple, potentially contradicting annotations, can account for such ambiguities. Also for test images, possible ambiguities can be reflected. One of the key advantages of this learning paradigm is its potential to significantly reduce the labeling effort, as it enables the inclusion of non-experts in the annotation process. This spares us from laborious iterations of obtaining clean, high-quality labels for all instances that expert pathologists agree on.

Under this learning paradigm, uncertainty arises from the annotator-dependent labeling behavior. Each annotator has different areas of expertise and a different level of experience, leading to variations in the annotation of medical images. These variations can be captured by probabilistic modeling, as shown in Chapter 6.

Given these learning paradigms, we now move to the possible solutions. In the next section, we outline the different probabilistic models that were used for multiple instance learning, active learning, and crowdsourcing.

1.3 Probabilistic Methods

For critical tasks like computer-aided diagnosis, where the consequences of misclassification can be severe, considering uncertainties for AI models is crucial. Here, human reasoning can serve as an example for the designed probabilistic AI models. When a human being does estimations, he or she takes possible uncertainties into account, resulting in predictions with varying levels of confidence. When we rely on those estimations, we expect the person to communicate the degree of confidence in their assessments. This consistent communication helps to establish trust in the reliability of their judgments.

In the context of AI algorithms, uncertainties due to ambiguous or missing labels can be captured by probability theory. Probabilistic AI models allow the efficient handling of inherent uncertainties within the problem domain and the ability to generate probabilistic predictions. Unlike deterministic methods that provide single-point estimates, probabilistic models produce predictive distributions based on a sound mathematical background [12; 13; 18]. The uncertainty estimations derived from the predictive distributions enable the distinction between safe and unsafe predictions. Therefore, they are enabling probabilistic reasoning, similar to a careful human assessment. In the following, we introduce the probabilistic methods that were investigated for the different applications in this thesis.

Gaussian processes are non-parametric machine learning models that define a probability distribution over a space of functions $f : \mathcal{X} \mapsto \mathbb{R}$. They are characterized by the property that, for any arbitrary set of input points $X = (x_1, x_2, \ldots, x_N)^T$, the output $F =$

$(f(x_1), f(x_2), \ldots, f(x_N))^T$ follows a multivariate Gaussian distribution $\mathcal{N}(\mu, K)$. The mean is defined by a mean function $\mu : \mathcal{X} \mapsto \mathbb{R}$ such that $\mu = (\mu(x_1), \mu(x_2), \ldots, \mu(x_N))$. The covariance is defined by a kernel function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, such that the matrix with $(k(x_i, x_j))_{1<=i,j<=n}$ is positive semidefinite for all x_i, x_j.

In the context of machine learning, we briefly describe how Gaussian processes work for the most generic case, one-dimensional function regression, based on the work of Rasmussen and Williams [12]. First, a prior is defined, encoding prior knowledge of the problem. A common choice is a zero mean function:

$$F \sim \mathcal{N}(0, k(X, X)) \tag{1.3}$$

with $X = (x_1, x_2, \ldots, x_N)^T$ describing the N-dimensional vector of training points. Then, the observed data $Y = (y_1, y_2, \ldots, y_N)^T$ is taken into account, assuming that the output values are a result of the underlying function f and observation noise $\epsilon \sim \mathcal{N}(0, \sigma^2_{\text{obs}})$, such that $y = f(x) + \epsilon$. Based on the prior distribution and the observed data, the posterior distribution for a test point x_* can be calculated in closed form as follows:

$$\mu_* = k(x_*, X)(k(X, X) + \sigma^2_{\text{obs}} I)^{-1} Y \tag{1.4}$$

$$\sigma^2_* = k(x_*, x_*) - k(x_*, X)[k(X, X) + \sigma^2_{\text{obs}} I]^{-1} k(X, x_*). \tag{1.5}$$

As the calculation involves inverting an $N \times N$ matrix, Gaussian processes are not scalable to large datasets. Therefore, a limited number of pseudo inputs (so-called inducing points) can be used to represent the data. The location and output distribution of these inducing points are part of the optimization process to obtain the best possible approximation. These approximate models are also called Sparse Gaussian processes [19], and a common strategy for optimization is to maximize the log evidence lower bound using variational inference. Gaussian processes are a powerful tool due to their mathematical interpretability, good function approximation properties, and sound uncertainty estimation, as provided by the estimated variance (see eq. 1.5).

In this thesis, we use Gaussian processes in the context of multiple instance learning to address the uncertainty of missing instance labels. In Chapter 2 they are used to estimate the probabilistic attention weight of each instance, while in Chapter 3 they provide the probabilistic framework to introduce instance correlations. We also relied on Gaussian processes for multiple instance learning in related articles about the detection of hemorrhage in computerized tomography scans, see Sections 1.3 7.2. Although the image domain is different, these techniques could be easily transferred to the domain of histopathological images.

Bayesian neural networks are artificial neural networks with probabilistic weights W that follow a probability distribution [13]. Based on Bayesian principles, a prior distribution $p(W)$ is defined, commonly following a Gaussian distribution. Given the observed input

images X and corresponding labels Y, the posterior distribution over the weights can be computed using Bayes' theorem:

$$p(W|X,Y) = \frac{p(Y|X,W)p(W)}{p(Y|X)}. \tag{1.6}$$

However, calculating the posterior distribution analytically is usually intractable due to the complex nature of neural networks. Therefore, approximate inference methods are employed to estimate the posterior distribution. One common approach is variational inference, where a variational distribution $q(W)$ is introduced and optimized to approximate the true posterior. Similar to Sparse Gaussian processes, the optimization can be performed using the log evidence lower bound. The predictive distribution for a test input x_* can be obtained by marginalizing over the weights:

$$p(y_*|x_*,X,Y) = \int p(y|x_*,W)p(W|X,Y)dW. \tag{1.7}$$

This prediction represents a distribution of possible output values at the test points, accounting for both aleatoric uncertainty (inherent noise in the data) and epistemic uncertainty (uncertainty due to limited data). Due to the high number of network weights, the integral in eq. 1.7 is commonly approximated, for example, by Monte Carlo sampling.

There are several advantages of Bayesian neural networks. They are similar to deterministic neural networks which makes them very efficient to train and allow a straightforward classification. Additionally, they can estimate different types of uncertainties, which is an important advantage for specific applications.

In our work, Bayesian neural networks are used to estimate different uncertainties in the context of active learning in Chapter 5. The model and data uncertainty (epistemic and aleatoric uncertainty) are used during acquisition to find the most informative image patches and provide a confidence estimation during prediction.

Probabilistic generative models are models that encode uncertainty in a random latent variable [14]. Commonly, the distribution is defined as a Gaussian:

$$z \sim \mathcal{N}(\mu, \Sigma), \tag{1.8}$$

parameterized by mean μ and covariance matrix Σ. In our context, the random variable represents the different segmentation variations that are possible, given for a certain image. For an image x_*, the corresponding feature map v_* and the latent variable z for the rater r are then combined to generate a probabilistic segmentation s_*:

$$q(s_*|x_*,r,\theta) = \int f_\theta(v_*,z)q(z|r)dz. \tag{1.9}$$

where f_θ describes a convolutional segmentation head with parameters θ. Previously, such

a model has been designed as a conditional variational auto-encoder [14]. In our work, we propose to directly train the distributions of the latent variables.

Probabilistic generative segmentation models have the advantage that the uncertainty can be modeled explicitly by a random latent variable. This allows an exact mathematical representation of intra- and inter-observer variability. They were used in Chapter 6 for medical image segmentation.

1.4 Objectives

The main objective of this thesis is to develop new probabilistic deep learning methods to overcome the labeling bottleneck in histopathology. In the frameworks of the different learning paradigms presented in section 1.2, different sources of uncertainties arise that must be addressed by adequate probabilistic modeling. In the following, we describe the objectives of the different research areas.

To develop novel multiple instance learning algorithms by modeling the unknown instance labels as a source of uncertainty. There are different existing approaches to train a deep learning model in the multiple instance learning setting. We aim to improve these approaches by taking a probabilistic perspective on current challenges and apply Gaussian processes. Concretely, there are two research directions. (i) Many existing multiple instance learning methods rely on attention mechanisms which are also used in the popular transformer architectures [11]. In these approaches, the importance of each instance for the bag prediction is estimated by attention weights. Probabilistic methods can help to capture the uncertainties of the attention weights and provide predictions that take these uncertainties into account. This line of research was investigated in Chapter 2, using a Gaussian process based attention mechanism. (ii) Another line of research is the probabilistic modeling of instance correlations. As in histopathological images, neighboring patches (instances) are highly correlated, this should be reflected in the model. As many existing approaches assume that the instances are independent, adequate probabilistic modeling of the instance correlations can improve the predictions considerably. In Chapter 3 we investigate a probabilistic framework based on Gaussian processes, that takes instance correlations into account.

To address the data-related challenges in active learning with probabilistic models. Histopathological images impose major challenges for active learning algorithms due to ambiguities and artifacts in the data. Existing active learning algorithms mistakenly assign high informativeness to images with ambiguities and artifacts. This leads to the acquisition and labeling of these images, although their value for training is low. Bayesian neural networks which provide uncertainties in the output can help to avoid the acquisition of these images with ambiguities and artifacts and focus on images with cancerous tissue. Novel probabilistic models can therefore improve active learning for histopathological images and

are presented in Chapter 5.

To develop novel probabilistic crowdsourcing algorithms that address the label uncertainty of different annotators. In the context of histopathology, where intra- and inter-observer variability of annotations is prominent, the absence of a definitive "ground truth" poses significant challenges both during model training and predictions within deep learning frameworks. Consequently, one key objective of this Ph.D. thesis is to develop novel probabilistic crowdsourcing algorithms that effectively tackle the inherent label uncertainty introduced by different annotators. During the training process, it is important to consider the uncertainties associated with each label, caused by the subjective assessment. Probabilistic generative models enable the representation and handling of these label uncertainties effectively. In Chapter 6 we propose such a model which accounts for the variations introduced by multiple annotators and makes predictions that reflect possible ambiguities.

1.5 Methodology

The novel methods developed in the thesis require a sound theoretical foundation as well as extensive empirical experiments to prove their applicability in practice. The scientific guidelines for this purpose are represented by the following steps:

1. **Observation:** We first study the literature regarding multiple instance learning, active learning, and crowdsourcing as well as existing probabilistic models used in the addressed and other domains.

2. **Data collection:** We assess the performance of the proposed methods on toy datasets as well as real-world histopathological datasets that are publicly available for the sake of reproducibility of our work.

3. **Hypothesis formulation:** We select state-of-the-art models and propose new ones to improve the results, addressing the problems presented in the objectives.

4. **Experimentation:** We perform rigorous experimentation with the collected data in step two. We use the computational resources of the Visual Information Processing research group of the University of Granada. For each experiment we choose different metrical measurements, including measures that were reported in previous studies for comparison.

5. **Hypothesis contrast:** We compare, analyze and validate the results obtained in the experimentation against the state-of-the-art techniques in the literature.

6. **Hypothesis proof or refusal:** We check if the extracted conclusions agree with the hypothesis previously formulated. We report the benefits and disadvantages of the proposed methods. If the results are not satisfactory, we will go back to step three and formulate a new hypothesis.

7. **Thesis extraction:** We formalize the conclusions during the research process and justify the developed methods through experimentation. All the proposals and results are synthesized in this memory.

1.6 Results

Developing probabilistic methods for multiple instance learning, active learning, and crowdsourcing led to promising scientific advances and competitive experimental results. We found that for each learning paradigm, we could overcome existing challenges successfully. Before presenting the detailed scientific work, we want to highlight some general findings of the presented articles.

Novel methods in multiple instance learning, active learning, and crowdsourcing allow a high performance with label-efficient training. In the different learning scenarios, we were able to achieve competitive performances using fewer or imperfect labels for model training. Even for small datasets, the developed probabilistic models were able to generalize well to unseen data. In comparison to supervised methods (relying on extensive labeling of all data), the performance gap with advancing research is vanishing, which might make extensive labeling obsolete in the future. This is important for all the numerous cancer types, classification tasks and special use-cases for which no large, extensively labeled datasets are available yet. The proposed methods reduce the required resources to train AI models for new tasks considerably.

Choosing the right probabilistic model to target a specific problem can improve the state-of-the-art. In the different presented articles, we used different probabilistic models, tailored to each use-case.

For the multiple instance learning paradigm, we relied on Gaussian processes due to their good function regression capabilities, robustness to overfitting, and mathematically sound integration in probabilistic frameworks. In our study, Gaussian processes exhibited remarkable performance when employed for attention weight regression, surpassing other existing models in three distinct prostate cancer experiments. In a different work, Gaussian processes facilitated the straightforward incorporation of instance correlations within the probabilistic modeling process, as evidenced in a separate investigation.

Bayesian neural networks are efficient models for multiclass classification and therefore used for active learning with image patches. These models are easy to train, which is an important advantage in active learning, where the model has to converge in each acquisition step with a growing amount of data. Additionally, they offer the possibility to estimate different types of uncertainties (aleatoric and epistemic uncertainty) which allows to separate informative from ambiguous image patches. The informative patches are used for labeling, while the ambiguous patches were avoided. The proposed model was able to prove this advantage in practice in comparison to other existing methods, resulting in state-of-the-art

performance for active learning of prostate cancer patches.

Generative segmentation models have the advantage of explicitly representing uncertainty by a probability distribution. We applied them to the segmentation of prostate cancer images with crowdsourced labels. The novel probabilistic segmentation model was able to capture uncertainties of inter- and intra-observer variability leading to a better performance and exact uncertainty measure in the output than other compared methods.

Combining deterministic and probabilistic methods is very effective. It is not necessary to have a completely probabilistic model. We combined in all of our works a deterministic convolutional neural network with a probabilistic model for high-level reasoning. This allows leveraging the efficient and straightforward optimization of the feature extractor, while the decision-making is performed by the probabilistic model. We want to emphasize again that the concrete probabilistic model should be designed depending on the purpose. In order to perform inference, we often opted for Monte Carlo integration as an approximation that is computationally efficient when applied in the last layers and allows an easy combination with the deterministic models and straightforward implementation. Additionally, we found it highly beneficial to design models that can be implemented in common deep learning libraries (Tensorflow/Pytorch) for full support of GPU-based training.

Capturing uncertainty is not only beneficial for training - it can also provide valuable information for predictions. The proposed probabilistic models showed an improved performance but also had another great advantage: the possibility to provide a probabilistic prediction that can reflect uncertainties. Assessing uncertainties is crucial in medical applications and we showed experimentally that the estimated uncertainties indicated predictions with a higher risk of being wrong. We investigated different sources of uncertainties such as the attention mechanism in multiple instance learning, data ambiguities, model parameters or inter-observer variability and showed that they were reflected in the predictive distribution. This provides great benefit for a safe use in clinical practice.

The detailed results of our scientific work are presented in several scientific articles, ordered by chapters in this thesis. Here, we provide a general overview.

Chapter 2: In this work, we propose a multiple instance learning model with an attention mechanism based on Gaussian processes. The probabilistic attention estimation leads to an overall probabilistic output that captures the uncertainty induced by missing instance labels. The proposed model is evaluated on two toy datasets, as well as two public prostate cancer datasets with WSIs.

Chapter 3: Based on the probabilistic multiple instance learning model VGPMIL [20], we introduce instance correlations to improve the overall performance. Existing models usually assume the independence of all instances, although the instances in one bag, such as neighboring patches, are often highly correlated. The proposed method models instance

correlation explicitly in a probabilistic framework.

Chapter 4: In this work, we combine semi-supervised and multiple instance learning to classify histopathological patches with a convolutional neural network. The proposed model is able to train with bag labels and an arbitrary amount of instance labels. It reaches a classification performance close to the supervised one with a small percentage of instance labels for three public datasets of breast and prostate cancer. The proposed model has been proven to be an ideal feature extractor to obtain features for probabilistic models and was used in the work of Chapters 2 and 3.

Chapter 5: For active learning with histopathological image patches, we propose a novel Bayesian neural network. The motivation is that existing methods are distracted by ambiguous images and artifacts for which they assign a high informativeness at the acquisition step. We empirically show that they acquire these images for labeling, although they do not provide much information for the training. Our proposed model measures different types of uncertainties to focus on the truly informative images, while avoiding ambiguous images or those with artifacts.

Chapter 6: We propose a probabilistic semantic segmentation model for crowdsourcing which can incorporate labels from different annotators. To the best of our knowledge, this is the first work that explicitly models the inter- and intra- observer variability. The labeling behavior of each annotator is encoded by a random latent variable following a trainable probability distribution. We derive the theoretical background and report promising experimental findings for different experiments on public breast and prostate cancer datasets.

Chapter 7: In this chapter we list additional important scientific contributions that are relevant to this thesis. The three further articles are devoted to multiple instance learning closely related to Chapters 2 - 4. Two articles present probabilistic models for multiple instance learning but on computerized tomography scans for hemorrhage detection, described in Sections 7.1 and 7.2. Another presented article proposes a novel algorithm for multiple instance learning with WSis based on constrained optimization, see section 7.3. Section 7.4 describes the process of obtaining digital images from human tissue and the origin of artifacts.

Chapter 2

Attention Estimation with Gaussian Processes for Multiple Instance Learning

2.1 Publication details

Authors: Arne Schmidt, Pablo Morales-Álvarez, Rafael Molina

Title: Probabilistic Attention based on Gaussian Processes for Deep Multiple Instance Learning

Reference: IEEE Transactions on Neural Networks and Learning Systems, 1-14, 2023, doi: 10.1109/TNNLS.2023.3245329

Status: Published

Quality indices:

- Impact Factor (JCR 2022): 10.4
  - Rank 14/145 (D1) in Computer Science, Artificial Intelligence
  - Rank 16/275 (D1) in Engineering, Electrical and Electronic

2.2 Main contributions

- We propose a probabilistic attention mechanism that provides a probability distribution over the random attention weights.
- The model is based on Sparse Gaussian processes that learn the regression of the instance attention weights in an end-to-end fashion within a deep learning architecture.
- The uncertainty of the attention mechanism is propagated to the final output prediction and helps to assess the risk of wrong predictions.
- Extensive experiments on the public prostate cancer datasets Sicapv2 and Panda show superior performance in comparison to existing state-of-the-art models.

PROBABILISTIC ATTENTION BASED ON GAUSSIAN PROCESSES FOR DEEP MULTIPLE INSTANCE LEARNING

Arne Schmidt
Department of Computer Science and AI
University of Granada
Granada, Spain

Pablo Morales-Álvarez
Department of Statistics and Operations Research
University of Granada
Granada, Spain

Rafael Molina
Department of Computer Science and AI
University of Granada
Granada, Spain

ABSTRACT

Multiple Instance Learning (MIL) is a weakly supervised learning paradigm that is becoming increasingly popular because it requires less labeling effort than fully supervised methods. This is especially interesting for areas where the creation of large annotated datasets remains challenging, as in medicine. Although recent deep learning MIL approaches have obtained state-of-the-art results, they are fully deterministic and do not provide uncertainty estimations for the predictions. In this work, we introduce the Attention Gaussian Process (AGP) model, a novel probabilistic attention mechanism based on Gaussian Processes for deep MIL. AGP provides accurate bag-level predictions as well as instance-level explainability, and can be trained end-to-end. Moreover, its probabilistic nature guarantees robustness to overfitting on small datasets and uncertainty estimations for the predictions. The latter is especially important in medical applications, where decisions have a direct impact on the patient's health. The proposed model is validated experimentally as follows. First, its behavior is illustrated in two synthetic MIL experiments based on the well-known MNIST and CIFAR-10 datasets, respectively. Then, it is evaluated in three different real-world cancer detection experiments. AGP outperforms state-of-the-art MIL approaches, including deterministic deep learning ones. It shows a strong performance even on a small dataset with less than 100 labels and generalizes better than competing methods on an external test set. Moreover, we experimentally show that predictive uncertainty correlates with the risk of wrong predictions, and therefore it is a good indicator of reliability in practice. Our code is publicly available.

Keywords Attention Mechanism · Multiple Instance Learning · Gaussian Processes · Digital Pathology · Whole Slide Images

1 Introduction

Machine learning classification algorithms have achieved excellent results in many different applications [1, 2, 3, 4, 5]. However, these algorithms need large datasets that must be labelled by an

expert. Such labelling process often becomes the bottleneck in real-world applications. In the last years, Multiple Instance Learning (MIL) has become a very popular weakly supervised learning paradigm to alleviate this burden. In MIL, instances are grouped in bags, and only bag labels are needed to train the model [6].

MIL is especially interesting for the medical field [7, 6, 8, 9]. As an example, consider the problem of cancer detection in histopathological images, where the goal is to predict whether a given image contains cancerous tissue or not (binary classification problem). Since these images are extremely large (in the order of gigapixels), they cannot be completely fed to a classifier (typically a deep neural network). Therefore, the classical approach is to split the image in many smaller patches and train a classifier at patch level. Unfortunately, this requires that an expert pathologist labels *every patch* as cancerous or not, which is a daunting, time-consuming, expensive and error-prone task [10]. In the MIL setting, each image is considered as a bag that contains many different instances (its patches). Importantly, since MIL only requires bag labels, the workload for the pathologist is reduced to labelling each image (and not every single patch).

Different underlying classification methods have been proposed for the MIL problem. Early approaches relied on traditional methods such as support vector machines [11], expectation maximization [12] or undirected graphs [13]. In recent years, many approaches are based on deep learning models due to their flexibility and their capacity to learn complex functions [14, 15]. In particular, the current state-of-the-art is given by an attention-based deep learning model originally introduced in [16]. The idea of attention-based MIL is to predict an attention weight for each instance, which determines its influence on the final bag prediction.

The attention mechanism has several advantages: it can be trained end-to-end with deep learning architectures because it is differentiable, it provides accurate bag-level predictions, and it provides explainability at instance-level (by looking at the instances with higher attention). Due to its success, it has been adapted and extended several times [17, 18, 19, 20, 21]. However, all these attention-based MIL approaches are based on *deterministic* transformations (usually one or two fully connected layers to calculate the attention weights). This has several drawbacks, such as the lack of uncertainty estimation and the overfitting to small datasets. These limitations can be addressed by introducing a probabilistic model as a backbone for the MIL attention module. Moreover, such a sound probabilistic treatment leads to better predictive performance, see e.g. [22, 23, 24, 25].

In this work, we introduce a novel probabilistic attention mechanism based on Gaussian processes for MIL, which will be referred to as AGP. It leverages a Gaussian process (GP) to obtain the attention weight for each instance. GPs are powerful Bayesian models that can describe flexible functions and provide accurate uncertainty estimation due to their probabilistic nature (their most relevant properties will be reviewed in Section 2). Moreover, AGP uses variational inference to ensure a probabilistic treatment of the estimated parameters. We experimentally evaluate AGP on different datasets, including an illustrative MNIST-based MIL problem, CIFAR-10, and three real-world prostate cancer classification tasks. Specifically, we show that: 1) AGP outperforms state-of-the-art and related MIL methods, 2) the estimated uncertainty can be used to identify which model predictions should be disregarded or double-checked, 3) higher attention is assigned to the most relevant instances of each bag (e.g. cancerous patches in images), and 4) AGP generalizes better than competitors to different datasets (and the estimated uncertainty reflects this extrapolation).

In machine learning literature, some related approaches have used GPs in the context of MIL, such as GPMIL [26] or VGPMIL [27]. These methods rely on a sparse GP for the instance classification, followed by an (approximated) maximum function for the final bag prediction. In contrast to our approach, both GPMIL and VGPMIL focus on instance-level predictions which are later combined for bag predictions, they are limited due to the simplicity of the instance aggregation (approximated max-aggregation), and cannot perform end-to-end training with deep learning models. In [9] we proposed the combination of a deep learning attention model and VGPMIL, but as a two-stage approach: first, the attention mechanism serves to train the feature extractor; in a second step, the VGPMIL model is applied to the extracted features. Although this approach showed promising results, it did not overcome all the aforementioned limitations of VGPMIL. In particular, the attention

mechanism is discarded after the first training phase and not used at all for the final predictions, so the model cannot fully leverage the advantages of combining an attention mechanism and GPs. So three major challenges remain unsolved: (i) end-to-end training, (ii) uncertainty estimation, and (iii) multiclass classification. Our proposed AGP model provides all three features, as described below.

For a different scenario, time series prediction, the combination of GP and attention has recently shown promising results [28]. For this problem, the GP replaced the final regression layer while the attention weights were calculated deterministically, which is a major difference to our work. Another existing approach combines an attention mechanism with GPs for channel attention [29]. Here, the GPs model correlations in activation maps of convolutional neural networks (CNNs) to estimate the channel attention weights. The channel attention helps improve the CNNs performance for visual tasks. However, to the best of our knowledge, our approach is the first one that estimates the attention weights with GPs in the context of deep MIL. Moreover, the novel method fully leverages the strengths of GPs for the attention estimation, including the strong function regression capabilities and uncertainty estimation. Indeed, the proposed AGP model does not only provide accurate predictions, outperforming existing state-of-the-art methods, but also provides an estimation of the uncertainty introduced by the attention. At the same time, our model inherits all positive properties of deterministic attention modules, such as explainability on instance level (as later discussed in section 4.2) and end-to-end training with deep learning feature extractors.

The paper is organized as follows. Section 2 describes the probabilistic model and inference used by AGP, preceded by the notation and background on the attention mechanism and GPs. Section 3.1 includes an illustrative and visual MIL experiment using MNIST. In section 4, we carry out three experiments on prostate cancer classification. We not only report a strong performance of the AGP model, but also analyze the probabilistic predictions for these real-world datasets. Finally, section 5 summarizes the main conclusions.

2 Methodology

In this section we present the theoretical framework for AGP. First, we describe some required background, the MIL notation (section 2.1), the attention mechanism (section 2.2), and the basics on (sparse) Gaussian Processes (section 2.3). Then, section 2.4 focuses on the description of AGP, including the probabilistic modelling, the variational inference, and how to make predictions.

2.1 Multiple Instance Learning (MIL) for Cancer Classification

In the classical MIL setting we assume that instances $\chi \in \mathbb{R}^D$ are grouped into bags $\mathcal{X}_b = \{\chi_{b1}, \chi_{b2}, .., \chi_{bN_b}\}$, where the number of instances N_b in each bag b can vary. Notice that this notation is not the most standard in MIL, where X_b is usually used for bags and x_{bi} for instances. However, we will use these letters for the input of the GP in Section 2.3. Therefore, to avoid confusion, we have chosen to use $\mathcal{X}_b$ and χ_{bi} for bags and instances, respectively. Each instance has a (binary) label $y_{bi} \in \{0, 1\}$ that remains unknown. The bag label T_b is known, and it is given by:

$$T_b = 0 \;\Leftrightarrow\; \forall i = 1, ..., N_b,\; y_{bi} = 0, \tag{1}$$

$$T_b = 1 \;\Leftrightarrow\; \exists i \in \{1, ..., N_b\} : y_{bi} = 1. \tag{2}$$

In cancer classification with histopathological images, the MIL setting considers a Whole Slide Image (WSI) as a bag $\mathcal{X}_b$ and its diagnosis as the bag label T_b. Since the complete WSI is too big to be processed by a common convolutional neural network, it is sliced into patches that form the instances. As MIL only requires bag labels for training, the local annotation of the patches by expert pathologists is not necessary. This provides huge benefits in terms of time and cost of the labeling process. Apart from the binary classification (cancerous vs. non-cancerous), we are also interested in the cancer class of the WSI. This class can be determined based on the features of the cancerous (positive) patches. In this setting, equation (1) still applies for non-cancerous (negative)

WSIs, while for cancerous WSIs we want to specify the class $T_b = c$, with $c \in \{c_1, c_2, .., c_K\}$ representing each one of the possible K cancer classes based on the cancerous areas.

2.2 Attention Mechanism

As mentioned in the introduction, AGP leverages a probabilistic GP-based attention mechanism to aggregate the information of the different instances in a bag. This is inspired by the deterministic attention introduced in [16]. Specifically, the model proposed in [16] consists of three main components: the feature extractor f_{fe}, the attention mechanism, and the final classification layer f_{cl}. The feature extractor is given by a convolutional neural network followed by some fully connected layers. It is used to process each instance, resulting in one feature vector $h_{bi} = f_{fe}(\chi_{bi})$ per instance, with $h_{bi} \in \mathbb{R}^P$. Then, the attention mechanism estimates a *deterministic* attention weight per instance a_{bi} based on these features. Specifically, the used mapping is:

$$a_{bi} = \frac{\exp\{w^\top \tanh(V h_{bi})\}}{\sum_j \exp\{w^\top \tanh(V h_{bj})\}}, \tag{3}$$

where the vector $w \in \mathbb{R}^{L\times 1}$ and the matrix $V \in \mathbb{R}^{L\times P}$ are optimized during training. Finally, the classification is done using the average of the extracted features, weighted by the attention values:

$$\hat{T}_b = f_{cl}\left(\textstyle\sum_i h_{bi} a_{bi}\right). \tag{4}$$

Our goal is now to replace the deterministic attention mechanism given in equation (3) by a GP model. The GP is a probabilistic method that is able to give a better estimation of the attention weights. Moreover, it allows for capturing the uncertainty introduced by the attention mechanism.

2.3 (Sparse) Gaussian Processes

Gaussian Processes are stochastic processes where the output distribution is assumed to be multivariate Gaussian [30]. They can be used to estimate an objective function f in a probabilistic way: the GP defines a prior distribution over functions (whose properties depend on the type of kernel used), and the posterior is computed given such prior and the observed data [31]. The major drawback of GPs is that the computation of the posterior is not scalable, because it involves inverting a matrix of size $N \times N$, with N the number of datapoints [32]. This is clearly relevant for our MIL scenario, where there exist typically plenty of instances.

To overcome this limitation, different types of Sparse Gaussian Processes (SGPs) have been introduced in the last years [33, 34, 35, 36]. Here we will follow the approach in [34], since it allows for training in batches. The idea behind SGP is to define the GP posterior distribution on a set of M inducing point locations $Z = \{z_i\}_{i=1}^M$, instead of doing it on the N real instances $X = \{x_i\}_{i=1}^N$. The amount of inducing points is taken $M \ll N$, and their location must be representative for the training distribution (in fact, they can be optimized during training, as we will do in AGP).

The formulation of SGP is as follows. Let U be the output of the GP at the inducing point locations Z, and F the output at the datapoints X. The SGP model is given by

$$p(U|Z) = \mathcal{N}(U|0, K_{ZZ}), \tag{5}$$

$$p(F|U, Z, X) = \mathcal{N}(F|K_{XZ}K_{ZZ}^{-1}U, \hat{K}), \tag{6}$$

where $K_{AB} := k(A, B)$ is the matrix obtained by applying the GP kernel function on A and B. Moreover, we have

$$\hat{K} = K_{XX} - K_{XZ}K_{ZZ}^{-1}K_{ZX}. \tag{7}$$

To perform inference, a posterior Gaussian distribution $q(U) = \mathcal{N}(U|\mu_u, \Sigma_u)$ is used on the inducing points. Therefore, the parameters to be estimated during training are μ_u, Σ_u, the kernel parameters, and the inducing points locations.

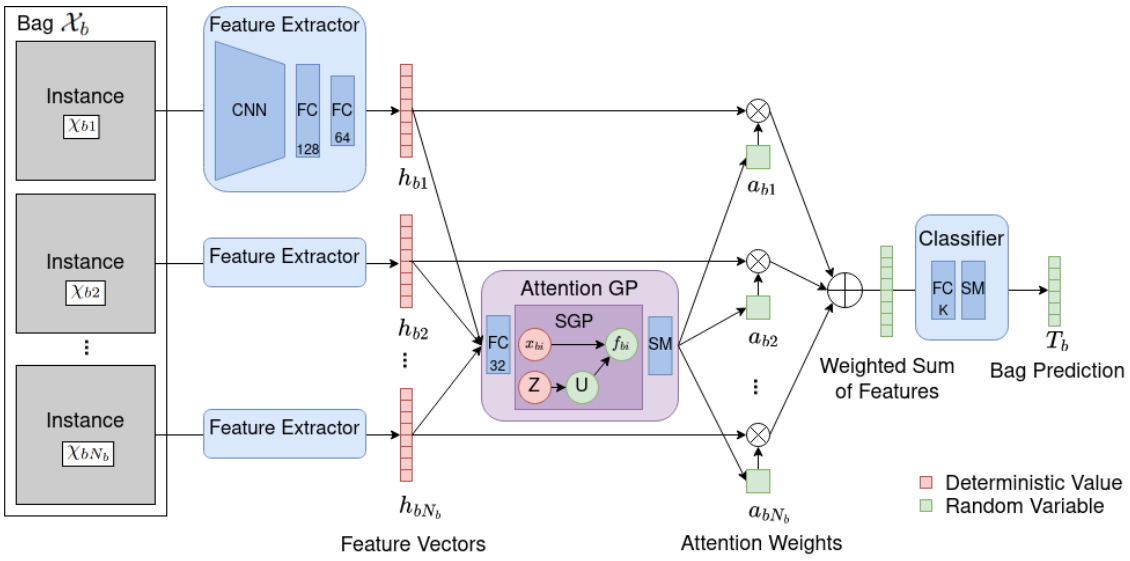


Figure 1: The AGP model architecture. The feature extractor consists of a Convolutional Neural Network (CNN) and two Fully Connected Layers (FC). The attention module incorporates another FC layer, the Sparse Gaussian Process (SGP), and a softmax (SM) function. The final classification is performed by another fully connected layer (where the dimensionality depends on the number of classes) and a softmax activation. While the feature vectors are deterministic values, the use of a (sparse) GP makes the attention weights and the final prediction random variables.

Given a test point, the prediction of the SGP model is a random variable (and not a single deterministic value). The mean of the random variable provides the value to regress, while the standard deviation provides the uncertainty. Finally, in this section we have written X for the input to the GP, as is standard in the GP literature. However, we want to stress that in our case the input to the GP will be given by the features extracted in some previous step, and not the raw input itself.

2.4 Probabilistic attention based on Gaussian Process (AGP)

Probabilistic modelling. The AGP model is depicted in Figure 1. It is a combination of a deterministic convolutional network that serves as a feature extractor f_{fe}, an SGP to estimate the attention, and a deterministic fully connected layer for the final classification f_{cl}. Next, we describe the different components using Figure 1 as reference.

Remember that $\mathcal{X}_b = \{\chi_{b1}, .., \chi_{bN_b}\}$ describe the instances in one bag b. First, the feature extractor f_{fe} is applied. It consists of a CNN and two fully connected layers with ReLu activation and 128 and 64 units, respectively. The choice of the CNN backbone depends on the task. For cancer classification, we will use EfficientNetB5 as the CNN backbone [37]. The output of the feature extractor are high-level feature vectors $H_b = \{h_{b1}, .., h_{bN_b}\}$ where each $h_{bi}, i = 1, .., N_b$ has 64 dimensions:

$$H_b = f_{fe}(\mathcal{X}_b). \tag{8}$$

We focus now on the attention module. Here, we first apply to each instance the same fully connected layer with sigmoid activation and 32 units. This further reduces the dimensionality, resulting in the SGP input feature vectors $X_b = \{x_{b1}, .., x_{bN_b}\}$ of 32 dimensions each. This alleviates the optimization of the inducing point locations, which are defined in the input space. Moreover, the sigmoid function guarantees values between 0 and 1, which facilitates the initialization of the inducing point locations. In summary, each vector h_{bi} with 64 components is transformed into a vector x_{bi} with 32 components. With these feature vectors, the SGP model described in Section 2.3 regresses the values $F_b = \{f_{b1}, .., f_{bN_b}\}$, which are normalized through a softmax (SM) layer to

calculate the attention weights $A_b = \{a_{b1}, .., a_{bN_b}\}$:

$$a_{bi} = \frac{\exp\{f_{bi}\}}{\sum_j \exp\{f_{bj}\}}. \tag{9}$$

Importantly, note that the attention weights are random variables, since they are computed from the SGP output. During inference, we will use Monte Carlo sampling to approximate their distribution.

Finally, the classifier f_{cl} is used to obtain the bag label T_b. The bag label T_b follows a categorical distribution with K classes, $T_b \sim \text{Cat}(p_1, \ldots, p_K)$, $\sum_k p_k = 1$. These probabilities are computed by applying the classifier f_{cl} over the average of the feature vectors H_b weighted by the attention weights A_b:

$$(p_1, \ldots, p_K) = f_{cl}\left(\textstyle\sum_i h_{bi} a_{bi}\right). \tag{10}$$

As shown in Figure 1, the classifier f_{cl} consists of one fully connected layer with one unit per class and a softmax activation function. Again, since the attention weights A_b are random variables, the probabilities $p_1, \ldots, p_K$ are random variables whose distribution will be estimated through Monte Carlo sampling. This probabilistic nature will allow for computing uncertainty estimation in the predictions.

Once we have described how AGP processes one bag $\mathcal{X}_b$, the joint full probabilistic model is

$$p(T, F, U) = p(T|F)p(F|U)p(U). \tag{11}$$

Here, we have written $T = \{T_1, \ldots, T_B\}$ for the collection of all the bag labels, and analogously for F. The inducing points U and their locations Z are global for all the bags because the feature space is the same for all instances of all bags and the SGP should be able to generalize to unseen bags. Notice that, to lighten the notation, we are not writing explicitly the dependency on all the variables. For instance, $p(U) = p(U|Z)$ depends on the inducing point locations Z; $p(F|U) = p(F|U, Z, X)$ also depends on Z and the SGP input X; and $p(T|F) = p(T|F, X)$ depends on X and depends on F only through A (recall eqs. (9)-(10)). Also, we are not writing explicitly the dependency on other parameters such as all the neural network weights (which are collectively denoted as W) and the SGP kernel parameters (which are denoted as θ).

Variational inference. To perform inference in AGP, we need to obtain the posterior distribution $p(F, U|T)$ and the learnable parameters W, θ, Z. Since eq. (11) is not analytically tractable, we resort to variational inference [38]. Namely, variational inference considers a parametric posterior distribution and finds the parameters that minimize the distance to the true posterior in the Kullback-Leibler divergence sense (by maximizing the log evidence lower bound, ELBO). In our case, we consider the distribution $q(F, U) = p(F|U)q(U)$, where $p(F|U)$ equals the (prior) conditional distribution in eq. (6) and $q(U) = \mathcal{N}(U|\mu_u, \Sigma_u)$ is a (multivariate) Gaussian with mean vector μ_u and covariance matrix Σ_u, both to be estimated during training (variational parameters).

With this choice, the ELBO to be maximized is

$$\log p(T) \geq \text{ELBO} = \mathbb{E}_{q(F)} \log p(T|F) - \text{KL}(q(U)||p(U)), \tag{12}$$

where $q(F) = \int p(F|U)q(U)dU$ is a Gaussian distribution since both $p(F|U)$ and $q(U)$ are Gaussian, recall eqs. (5)-(6). Specifically, we have $q(F) = \mathcal{N}(F|K_{XZ}K_{ZZ}^{-1}\mu_u, K_{XX} - K_{XZ}K_{ZZ}^{-1}(K_{ZZ} - \Sigma_u)K_{ZZ}^{-1}K_{ZX})$. The KL divergence $\text{KL}(q(U)||p(U))$ can be calculated in closed-form, since both distributions are Gaussian too. In practice, this term acts as a regularizer for the SGP model, since it encourages the posterior on the inducing points to stay close to the prior. To calculate the other term (log-likelihood), we have

$$\mathbb{E}_{q(F)} \log p(T|F) = \textstyle\sum_b \mathbb{E}_{q(F_b)} \log p(T_b|F_b), \tag{13}$$

since we naturally assume that bag labels are independent. Although the terms $\mathbb{E}_{q(F_b)} \log p(T_b|F_b)$ cannot be obtained in closed-form, they can be approximated by Monte Carlo integration:

$$\mathbb{E}_{q(F_b)} \log p(T_b|F_b) \approx \frac{1}{S} \textstyle\sum_s \log p_{T_b}^{(s)}. \tag{14}$$

Algorithm 1 AGP training procedure

Input: Instances $\{\chi_{bi}\}_{i=1,..,N_b}$ (e.g. image patches) for each bag $b = 1, \ldots, B$; bag labels $\{T_b\}$; number of epochs E.
Output: Optimal model parameters $Z, \mu_u, \Sigma_u, \theta, W$.
  for $e = 1$ to E (all epochs) **do**
    for $b = 1$ to B (all bags) **do**
      Predict features $H_b \leftarrow f_{fe}(\mathcal{X}_b)$.
      Apply fully connected layer in the attention module, i.e. $X_b \leftarrow f_{FC}(H_b)$.
      Calculate SGP output $p(F_b|U, Z, X_b)$ (eq. 6).
      Draw S Monte-Carlo samples $\tilde{F}_b^s \sim p(F_b|U, Z, X_b)$.
      Calculate log likelihood (LL) term following eq. (14).
      Calculate KL term in eq. (12) in closed-form.
      Calculate loss as $\mathcal{L} = -\mathrm{LL} + \mathrm{KL}$.
      Update $Z, \mu_u, \Sigma_u, \theta, W$ with $\nabla\mathcal{L}$ using Adam.
    end for
  end for
  return Optimal model parameters $Z, \mu_u, \Sigma_u, \theta, W$.

Here, the subindex T_b indicates that we take the class probability that corresponds to the (observed) bag label T_b (recall from eq. (10) that there exists one p_k for each class). The S samples $\{p_{T_b}^{(s)}\}_s$ are obtained by sampling $F_b^{(s)}$ from the Gaussian $q(F_b)$ with the reparametrization trick [39] and propagating through the rest of the network (that is, applying eqs. (9)-(10)). Notice that maximizing the log-likelihood term is equivalent to minimizing the standard cross-entropy between the estimated class probabilities and the ground truth vector (in a one-hot encoding), as shown in [22].

In summary, AGP training consists in maximizing the ELBO in eq. (12) with respect to the variational parameters (μ_u and Σ_u), the neural network parameters W, the SGP kernel parameters θ and the inducing point locations Z. To do so, we use stochastic optimization with the Adam algorithm and mini-batches [40]. In the experiments, each mini-batch is given by all the instances of one bag. Notice that the proposed model and inference allow for end-to-end training through the ELBO maximization. Algorithm 1 summarizes the training process.

Predictions. After training is completed, we are interested in predicting the class label T_b^* for a previously unseen bag $\mathcal{X}_b^*$. The prediction of the AGP model is given by a K-class categorical distribution with class scores $p_1, \ldots, p_K$ which are random variables. The mean of such random variables, $\overline{p}_k$, represents the predicted probabilities per class (and the predicted class is the one with highest probability, i.e. $T_b^* = \arg\max_k \overline{p}_k$). Additionally, the standard deviation of each class probability provides its degree of uncertainty. The total uncertainty for the bag prediction is defined as the mean of the standard deviations for each class. Notice that this approximation follows popular existing literature [34, 22].

To calculate the predictive random variables p_k, we have

$$p(T_b^*) = \int p(T_b^*|F_b)q(F_b)dF_b. \tag{15}$$

Similarly to eq. (13), this cannot be obtained in closed-form, since it requires integrating out the neural network f_{cl}. Following the same idea as there, we approximate the predictive distribution by Monte Carlo sampling. Namely, we take S samples from $q(F_b)$ and propagate them through the rest of the network (eqs. (9)-(10)) to obtain S samples $p_k^{(s)}$ for each class $k = 1, \ldots, K$. The mean and standard deviation for each p_k are computed empirically based on these samples. Analogously, we have S samples $a_{bi}^{(s)}$ for the predictive distribution over the attention weights for each instance inside the bag. We will see that these values provide explainability on which instances are the most relevant to obtain the bag prediction. Using just $S = 20$ samples works well in practice.

Implementation.

Algorithm 1 provides an overview of the implementation. It summarizes the main steps to train the model. In practice, the model is implemented with the deep learning libraries tensorflow (version 2.3.0) and its extension tensorflow-probability (version 0.11.1). The class *tensorflow_probability.layers.VariationalGaussianProcess()* is specially useful for the implementation of the SGP. It allows for an efficient, parallel execution of the algorithm on the GPU, as well as end-to-end training. Another benefit is the easy integration in other existing deep learning projects. The complete code for AGP is publicly available. [1]

Another key component for the implementation is the reparametrization trick to sample from $q(F_b)$, recall eq. (14). The (Gaussian) output distribution of the SGP is 'reparametrized' into one deterministic part and one probabilistic part to perform backpropagation through the probabilistic layer. Namely, the random vector $F_b \sim \mathcal{N}(\mu, \Sigma)$ can be split into $(\mu + L\epsilon) \sim \mathcal{N}(\mu, \Sigma)$, where L is the Cholesky factor of Σ and $\epsilon \sim \mathcal{N}(0, I)$. For MC sampling, a random sample $\hat{\epsilon} \sim \mathcal{N}(0, I)$ is drawn to obtain a sample from F_b, i.e. $\hat{F}_b = \mu + L\hat{\epsilon}$. This allows the backpropagation of the gradient through μ and L, while the random variable ϵ is independent from the model parameters. The reparametrization trick is already implemented in the above mentioned tensorflow-probability library. For further details we refer the interested reader to the original work [39].

3 Synthetic Experiments

In this section we evaluate our method on two synthetic MIL problems. First, Section 3.1 shows a visual experiment based on MNIST that helps better understand the proposed method. Then, Section 3.2 provides a more sophisticated multi-class problem where we compare our method against a wide range of state-of-the-art baselines.

3.1 An illustrative example: MNIST bags

The goal of this section is to illustrate the behavior of AGP in a simple and intuitive example. Specifically, we analyze two aspects of AGP: 1) its predictive performance. We will see that bag-level predictions are correct and high attention weights are given to positive instances. 2) The information provided by the estimated uncertainty (i.e. the standard deviation of the predictions). We will see that high uncertainty is assigned to bags that are difficult to classify.

The well-known MNIST dataset [41] contains 60000 training and 10000 test images. To define a MIL problem, we randomly group these images into bags of 9 instances each. We define the digit "0" as the positive class, and the rest of digits as negative class. Thus, a bag is positive if at least one of the nine digits in that bag is a "0". Otherwise, the bag is negative. We choose "0" because it can be mistaken with "6" or "9". Similar procedures to define a MIL problem on MNIST have been used in previous work [16]. The resulting MIL dataset has 6667 bags for training (2558 negative, 4109 positive), which are made up of 60.000 instances (54077 negative, 5923 positive). For testing it has 1112 bags (404 negative, 708 positive), which are made up of 10.000 instances (9020 negative, 980 positive). For all splits, around 40% of the bags have one positive instance, 17% two, 4% three and 1% four and the rest are negative bags.

As the given problem is less complex than the cancer classification task, we simplify the feature extractor. Namely, it contains one convolutional layer (4 filters, 3x3 convolutions) and one fully connected layer (64 units). The rest of the model remains as described in Figure 1. We train it end-to-end with cross-entropy and the Adam optimizer with a learning rate of 0.0001, for 5 epochs.

Regarding the predictive performance, AGP achieves 98.02% test accuracy at bag level. This is slightly better than when using deterministic attention (i.e. A-Det, which obtains 97.82%). Figures 2a and 2b show the predictions obtained by AGP for a negative and a positive bag, respectively. Notice that AGP learns to discriminate between positive and negative instances, assigning a high

[1] `https://github.com/arneschmidt/attention_gp`

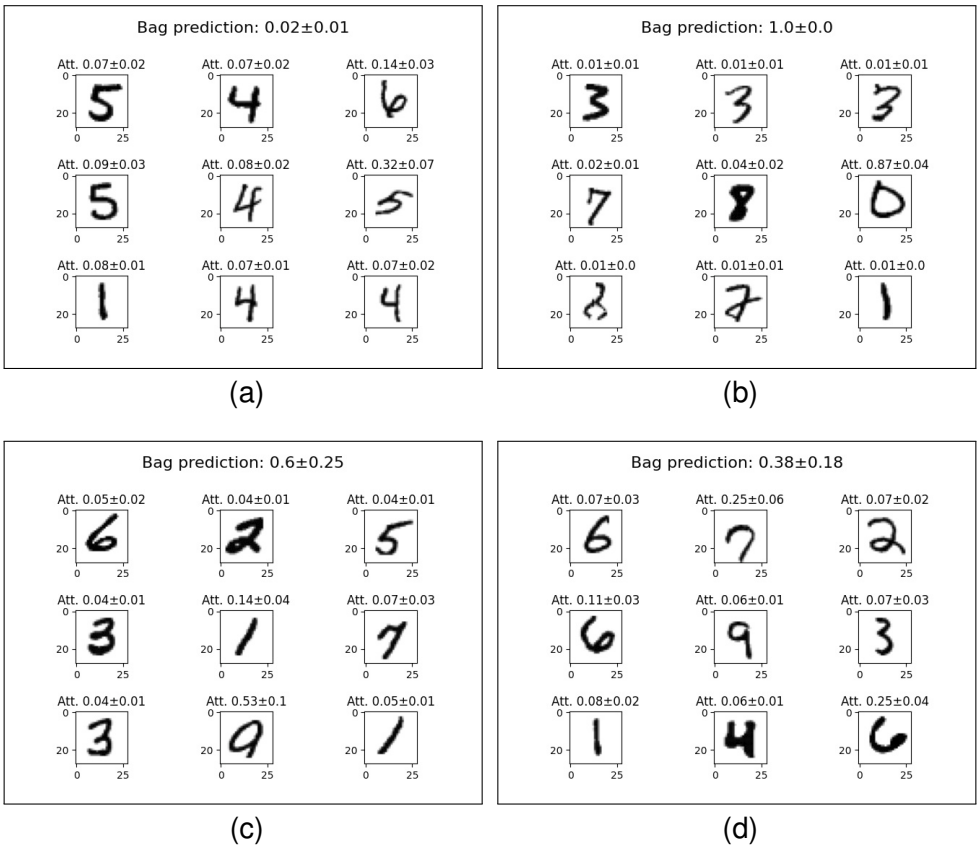


Figure 2: AGP classification of MNIST bags. In this dataset, the number "0" represents a positive instance, and all other digits are considered negative instances. We show the AGP bag prediction (probability to be positive) and corresponding attention weights for each instance. As these estimations are random variables in the AGP model, we report the mean and standard deviation. The top two figures show confident predictions for a negative (a) and a positive (b) bag. The bottom figures, (c) and (d), show two unconfident predictions with high standard deviations. Both bags are negative, but the model misclassifies (c) due to an ambiguous digit.

attention weight to the digit "0" in the positive bag. Also, notice that the standard deviation for both the attention weights and the bag prediction is low (i.e. the algorithm is confident on the decision).

Next, we analyze the role of the uncertainty by visualizing the prediction for ambiguous bags. In Figures 2c and 2d we see two examples of predictions with high standard deviations. These high standard deviations originate in ambiguous instances that lead to an uncertain final bag prediction. In Figure 2c, there is a '9' which is visually similar to a '0' because one line is (almost) missing. The model assigns a high attention but also a high standard deviation to this instance. The final bag prediction is false positive, but the high standard deviation of the bag prediction indicates a high uncertainty. In Figure 2d, two digits are ambiguous (those corresponding to the items $(1, 2)$ and $(3, 3)$ of the 3×3 matrix of digits). They are assigned a higher attention but again a slightly higher standard deviation than the other instances. The final negative bag prediction is correct, but the high standard deviation reflects a high uncertainty. Finally, notice that this qualitative observation on the uncertainty can also be confirmed statistically: while correctly classified bags have an average standard deviation of 0.006, the average standard deviation of incorrectly classified bags is more than ten times higher (0.065). Therefore, a high standard deviation indicates a high risk of a wrong prediction.

3.2 Evaluation on CIFAR-10

In this experiment with the CIFAR-10 dataset [42] we want to compare different deterministic and probabilistic approaches for a more difficult multi-class MIL problem.

The CIFAR-10 dataset consists of 32x32 images containing 10 different classes (airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks). The dataset contains 10.000 test images and we split the remaining images into 50.000 for training and 10.000 for validation. Originally, all labels of the images are known, but we create a multi-class MIL problem for our use-case. As in the MNIST experiment, we use bags of nine instances each. In this case, we select two positive classes while all other classes are negative, as explained next in more detail. Each bag has either the label 'airplane', 'car' or 'negative'. Negative bags contain only negative instances (i.e. birds, cats, deer, dogs, frogs, horses, ships, and/or trucks). Bags labelled as 'airplane' contain at least one image of airplane, while the remaining instances are negative. Similarly, each bag with the label 'car' contains at least one image of cars (and the rest of instances are negative). We choose to have an equal distribution of each class on the bag-level for training (1481 bags per class, 4443 bags in total), validation (370 bags per class, 1110 bags in total) and test (370 bags per class, 1110 bags in total). As the instances are drawn randomly, the exact amount per class vary in this setup. On average, 5.67% of the instances are from the 'airplane' class, 5.67% from the 'car' class, and the rest are negative instances.

We use the model architecture described in section 2 and depicted in Figure 1. For the feature extraction, we choose a CNN backbone that consists of 3x3 convolutions with relu activation and max pooling with a stride of 2x2. The exact layers are: two convolutional layers with 32 filters, max pooling, two convolutional layers with 64 filters, max pooling, two convolutional layers with 128 filters and max pooling. The fully connected layers have 128 and 64 units, respectively. The whole architecture is trained end-to-end. We use the Adam optimizer with a learning rate of 0.0001, cross-entropy, and 15 training epochs.

We compare our method against three state-of-the-art deterministic baselines that only differ in the MIL aggregation mechanism. In all the cases, we use the same feature extractor architecture, hyperparameters and iterations. The compared methods are:

- **Mean Aggregation (Mean-Agg)**. Instead of using an attention module, we aggregate the extracted features from each instance by taking their mean. This mean vector is then used for the final classification.
- **Attention Deterministic (A-Det)**. The attention module as proposed in [16] is used. Their attention weights and the final prediction are deterministic values.
- **Attention Deterministic Gated (A-Det-Gated)**. The advanced attention module proposed in [16] as an extension to A-Det is used. The gating mechanism was introduced to allow the algorithm to efficiently learn more complex relationships between instances.
- **Attention Gaussian Process (AGP)**. The probabilistic model proposed in this work, as described in section 2.

As shown in Table 1, AGP outperforms all the baselines, including the state-of-the-art attention mechanism A-Det-Gated. This suggests that our probabilistic attention is able to accurately assign attention weights to different instances. Indeed, this can be explained by the good regression capabilities of GPs, as known from previous studies [33, 35, 36]. To illustrate the learning process, we plot a training/validation curve of the AGP model in Figure 3. As seen in the plot, the model is robust to overfitting as the validation accuracy remains stable once it converges.

Finally, as a first approach towards future work, we also investigated other options to implement a probabilistic attention mechanism, based on Bayesian Neural Networks (BNNs) instead of the SGP. Leaving the rest of the architecture as shown in Figure 1, we first exchange the AGP attention mechanism by two fully connected Bayesian layers with weights following a Gaussian distribution [43] (32 and 1 units for the layers, respectively). In the same experiment setup, this model achieved

Method	Type	Acc. mean	Acc. S.E.	F1 mean	F1 S.E.
Mean-Agg	Det.	0.732	0.008	0.730	0.009
A-Det	Det.	0.735	0.003	0.735	0.003
A-Det-Gated	Det.	0.730	0.005	0.730	0.005
AGP	Prob.	**0.749**	0.008	**0.750**	0.008

Table 1: Results for Cifar-10 experiments with bags of 9 images and three classes. The experiment was repeated in 10 independent runs, we report the mean and standard error.

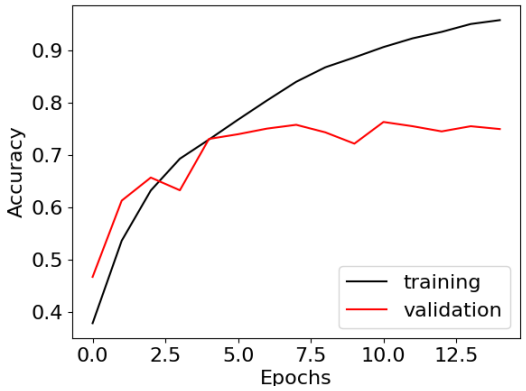


Figure 3: Training and validation accuracy for the AGP model in the CIFAR-10 experiment. Although the amount of labeled bags is low, the model is robust to overfitting as the validation accuracy remains stable.

0.642 accuracy and 0.644 F1-score, quite far from AGP. Similarly, we tested a model based on MC dropout [22] with an attention mechanism composed by one fully connected layer with 32 units, a Bayesian dropout layer, and a fully connected layer with 1 unit. The dropout probability was set to 0.5. This model obtained better results, achieving 0.74 accuracy and 0.739 F1-score. However, this is still lower than AGP (0.749 accuracy, 0.750 F1-score), which will be the focus in terms of probabilistic methods in the rest of this paper.

4 Experiments on prostate cancer classification

In this section, we evaluate AGP on the real-world problem of cancer classification. This is a very timely problem, since the development of computer aided diagnosis tools is attracting plenty of attention due to the large workload that pathologists are experiencing in the last years [8, 44]. For easier reproducibility, we use publicly available datasets. We will focus on prostate cancer, although our model is agnostic to the cancer type and can be applied to other cancer classification tasks.

In the rest of this section, we present the datasets used (SICAPv2 and PANDA), the implementation details, and the baselines used for comparison. Then, Section 4.1 focuses on SICAPv2 data, Section 4.2 focuses on PANDA data, and Section 4.3 evaluates the ability to extrapolate from one dataset to the other. The goal of these experiments is not only to show the strong performance of AGP on real-world data, but also to highlight the usefulness of the probabilistic output to estimate the predictive reliability.

Datasets. We use two publicly available datasets: SICAPv2 and PANDA. The extracted biopsies (WSIs) are classified by pathologists based on the appearance and quantity of cancerous tissue. There are two different scales: the Gleason Score and the ISUP grade. For further background on

both scales, we refer the interested reader to [45]. In our experiments, we use the Gleason Score for SICAPv2 and the ISUP grade for PANDA. The reason for this is twofold. First, to compare our results with previous literature (for which we need to use the same grading scale as them). Second, to show that the proposed method is robust to the grading scale used (obtaining good results in both scenarios).

SICAPv2[2] consists of 155 biopsies (WSIs). The class distribution of the assigned Gleason Score (GS) is the following: Non-Cancerous: 36, GS6: 14, GS7: 45, GS8: 18, GS9: 35, GS10: 7. The dataset is already split into four cross-validation folds, which contain between 86 and 97 WSIs for training and a separate set for testing. The publishers of the data distributed the biopsies so that the class proportions are reflected in each of the train and test splits. For more details, see [46]. The PANDA dataset[3] consists of 10616 WSIs and was presented at the MICCAI 2020 conference as a challenge. The total number of WSIs for each ISUP grade is the following: Non-Cancerous: 2892, G1: 2666, G2: 1343, G3: 1242, G4: 1249, G5: 1224. As the test set of PANDA is not publicly available, we use the train/validation/test split proposed in [47], which has 8469, 353 and 1794 WSIs, respectively. Again, each split follows the overall class proportions.

For both datasets, a 10x magnification is used and the WSIs are split into 512x512 patches with a 50% overlap. The patch-level annotations of the datasets are discarded for our experiments, since our model only requires bag labels for training.

Implementation Details. The AGP architecture used for these experiments is depicted in Figure 1, and explained in Section 2.4. Here we provide the rest of the details for full reproducibility. We use 64 inducing points with 32 dimensions each, whose locations are initialized with random values between 0.3 and 0.7 (because this is typically the range of values initially obtained by the previous sigmoid layer). For Monte-Carlo integration, we draw 20 samples for training and for testing. We use a class balanced loss with cross-entropy and the Adam optimization algorithm. We set the learning rate to 0.001 for the first 10 epochs, and use learning rate decay afterwards with the factor $e^{-0.1}$ per epoch. The total number training epochs is set to 100. Finally, as it is computationally unfeasible to train the feature extractor on all patches of a (huge) WSI at once, we first extract the high-level features of each patch. We train the CNN and the first fully connected layer with the method proposed in [48], using only WSI labels. The obtained 128 dimensional feature vectors per patch are then used to train the last fully connected layer of the feature extractor and the rest of the model.

Baselines. In all the experiments, we compare mean aggregation as a baseline and three state-of-the-art MIL approaches trained with the same feature vectors, hyperparameters and iterations. The only variation is the attention mechanism. Additionally, in each experiment we compare with other related approaches that have used the same data. For details about the approaches (Mean-Agg, A-Det, A-Det-Gated) please recall the bullet points in Section 3.2.

Evaluation metric. As common in prostate cancer classification tasks [49, 47, 50], we report the performance in terms of quadratic Cohen's kappa, which measures the agreement between the labels provided by pathologists and the model's predictions. A kappa value of 0 means no agreement (random predictions) and a kappa value of 1 means complete agreement. In all cases, we show the mean and the standard error of the results over several independent runs.

4.1 SICAPv2 Results

The SICAPv2 experiment is used to test our model on a very small dataset, where there is a high risk of overfitting. Recall that the training set for each cross-validation fold has less than 100 WSIs, and correspondingly there are less than 100 labels for the MIL models.

As a first part of the SICAPv2 experiment, we perform an ablation study to show the effect of different hyperparameters that are important in the proposed attention module. We identify three

[2]Available at: https://data.mendeley.com/datasets/9xxm58dvs3/1

[3]Available at: https://www.kaggle.com/c/prostate-cancer-grade-assessment

Method	κ mean	κ S.E.
AGP-feat-dim-32	0.832	0.004
AGP-feat-dim-64	**0.847**	**0.001**
AGP-feat-dim-128	0.835	0.001
AGP-feat-dim-256	0.844	0.001
AGP-ind-points-16	0.832	0.004
AGP-ind-points-32	0.840	0.002
AGP-ind-points-64	**0.847**	**0.001**
AGP-ind-points-128	0.689	0.007
AGP-relu	0.675	0.008
AGP-sigmoid	**0.847**	**0.001**
AGP-tanh	0.830	0.007

Table 2: Ablation studies with the SICAPv2 dataset. We study the effect of three important components: the feature vector dimension, the number of inducing points for the SGP, and the activation function used inside the attention module. Bold letters indicate the configuration used in the final setup.

hyperparameters that are especially interesting for this analysis: the dimension of the feature vectors h, the number of inducing points of the SGP, and the activation function inside the attention module (the one before the SGP). We separately vary these hyperparameters while all the other values are set to default (as described in Section 4, *Implementation Details* paragraph).

As shown in Table 2, the AGP model is robust against variations of the feature vector dimensionality, but the model with 64 feature dimensions slightly outperforms the others. For the number of inducing points we see a robust performance for less inducing points (16-32-64), but a high number (128) led to instabilities in model training. Indeed, we had to reduce the learning rate to 0.0001 (from 0.001) to obtain convergence, but we still observe a remarkable performance drop. We believe that, as fully connected layers have reduced the complexity and dimensionality of the features, a relatively small amount of inducing points allows precise predictions. Finally, the same convergence problems appear if a ReLu activation function is used before the SGP (we again used a lower learning rate of 0.0001 in this case to get convergence). The tanh function, which is more similar to the sigmoid function, shows a robust performance. Interestingly, these results suggest that limiting the input range to the SGP is clearly beneficial, as the sigmoid and tanh functions output values in the range $(0, 1)$ and $(-1, 1)$, respectively. The relu function has outputs in the range $[0, \infty)$, which makes it harder to find adequate inducing point locations for the SGP.

After the ablation study, the best performing model is compared to other state-of-the-art methods. Notice that the best performing configuration is precisely the one that was described in Section 4, *Implementation Details* paragraph. The results in Table 3 show that the AGP model outperforms all other MIL approaches, including existing attention based methods A-Det and A-Det-Gated, which can be considered state of the art for MIL. Also, the Cohen's quadratic kappa value of 0.847 is a remarkable one for such a small dataset. Furthermore, we see that AGP outperforms the existing supervised methods Silva-Rodríguez et al. [49] and Arvaniti et al. [50], which use all patch-level annotations to train the feature extractor (reported by [49] for this dataset). This can be partly explained by the stronger focus on patch-level predictions instead of bag-level predictions of these approaches (for bag-level predictions, they implement a simple aggregation method). Interestingly, notice that the AGP model provides an accurate WSI diagnosis without local annotations.

We also report the confusion matrices for all the compared models, see Figure 4. We see that AGP is strong in distinguishing cancerous from non-cancerous WSIs: there is only one false positive and one false negative in AGP predictions. The other two approaches that show a comparable

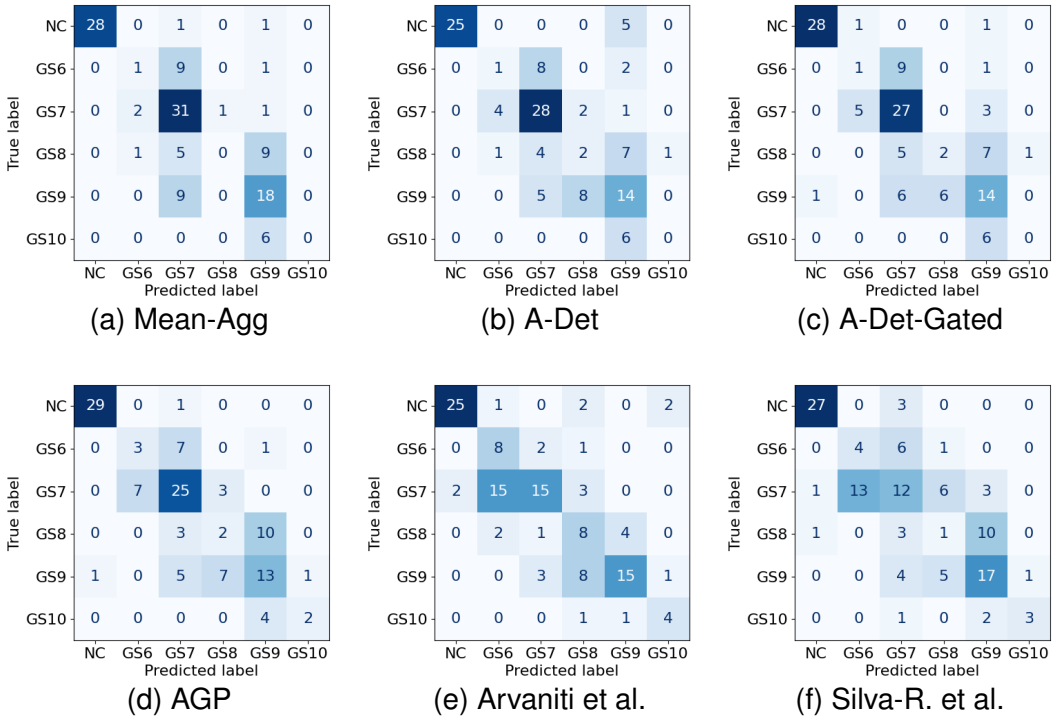


Figure 4: Confusion matrices for the 4-fold cross-validation of SICAPv2 with the classes "non-cancerous" (NC) and Gleason Score 6 to 10 (GS6-GS10).

Method	Learning	κ mean	κ S.E.
Mean-Agg	MIL	0.800	0.041
A-Det	MIL	0.770	0.008
A-Det-Gated	MIL	0.814	0.007
AGP	MIL	**0.847**	0.001
Arvaniti et al. [50] [49]	Supervised	0.769	N.A.
Silva-Rodríguez et al. [49]	Supervised	0.818	N.A.

Table 3: Results for SICAPv2 dataset. We report the mean and standard error of Cohen's quadratic kappa (κ) for 4 independent runs, with a four-fold cross-validation in each run. The last two methods do not report the standard error.

performance in the binary (cancerous vs non-cancerous) classification task, Mean-Agg and A-Det-Gated, show a major systematic error: they misclassified all the WSIs with Gleason Score 10.

The training process of the proposed AGP model (with previously extracted features) takes less than 2 minutes for the SICAPv2 dataset, and is negligible in comparison to the training of the feature extractor ($\sim$ 7 hours in this case [48]). The test time is on average 2.1 seconds for a complete WSI. This computing time mainly corresponds to the feature extraction ($\sim$ 7 seconds [48]), while the attention mechanism and final classification together can be executed in less than 0.1 seconds per WSI. The runtime bottleneck is therefore the feature extractor, and not the proposed probabilistic attention module. This efficiency is an important benefit for the clinical practice, as well as other areas where speed in prediction is paramount.

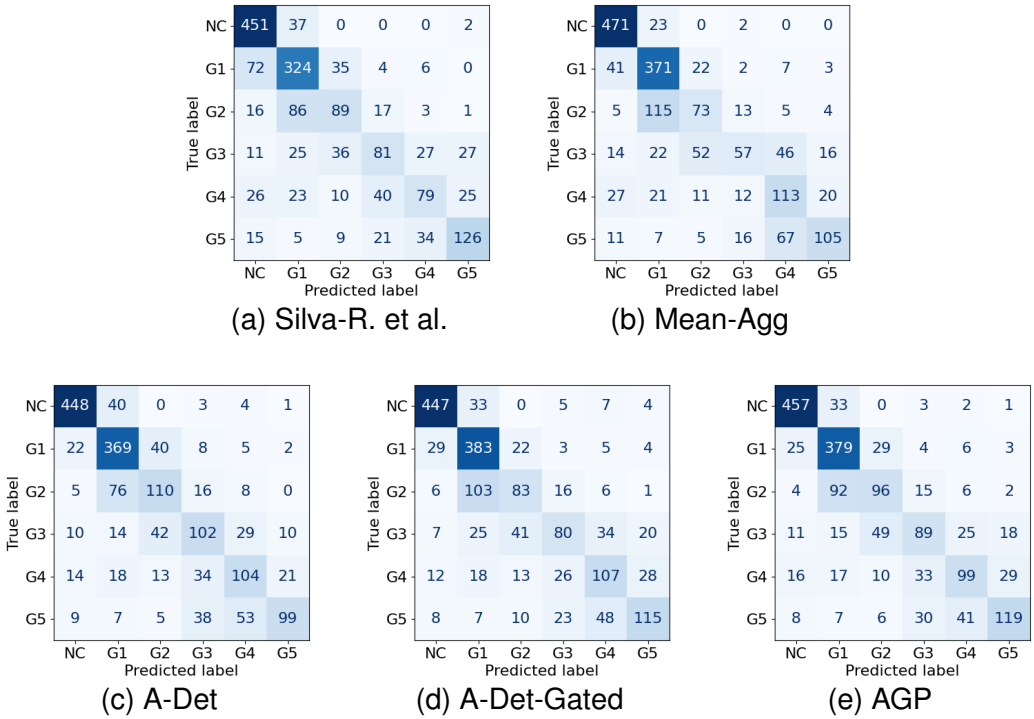


Figure 5: Confusion matrices for the PANDA test set. The six classes are "non-cancerous" (NC) and ISUP grades 1 to 5 (G1-G5).

Method	Learning	κ mean	κ S.E.
Mean-Agg	MIL	0.803	0.003
A-Det	MIL	0.811	0.004
A-Det-Gated	MIL	**0.816**	0.004
AGP	MIL	**0.817**	0.003
Silva-Rodríguez et al. [47]	MIL	0.793	N.A.

Table 4: Results for PANDA dataset. We report the mean and standard error of Cohen's quadratic kappa (κ) for 4 independent runs. The last method does not report the standard error.

4.2 PANDA Results

In this experiment, we show that AGP also outperforms other approaches in a large real-world problem. Moreover, by visually inspecting the predictions, we check that the AGP attention mechanism allows for identifying cancerous regions. Finally, we analyze the relevance of the probabilistic predictions provided by AGP, which can be used to detect wrong predictions. Also, as explained before, the different grading scale used here (ISUP scale) shows the robustness of AGP.

In Table 4 we see that the AGP performance is superior to the other MIL approaches. We have included Silva-Rodríguez et al. [49], a recent MIL method that was evaluated on the same dataset. Although the difference in performance is small in some cases (e.g. against A-Det-Gated), we will see that the probabilistic nature of AGP provides additional benefits (such as the degree of uncertainty on the predictions).

The confusion matrices, see Figure 5, show again that the AGP model is very strong at differentiating cancerous from non-cancerous WSIs. Although the models of Silva-Rodriguez et al. [49] and Mean-Agg have less false positives (see the first row of the matrices), these models suffer from

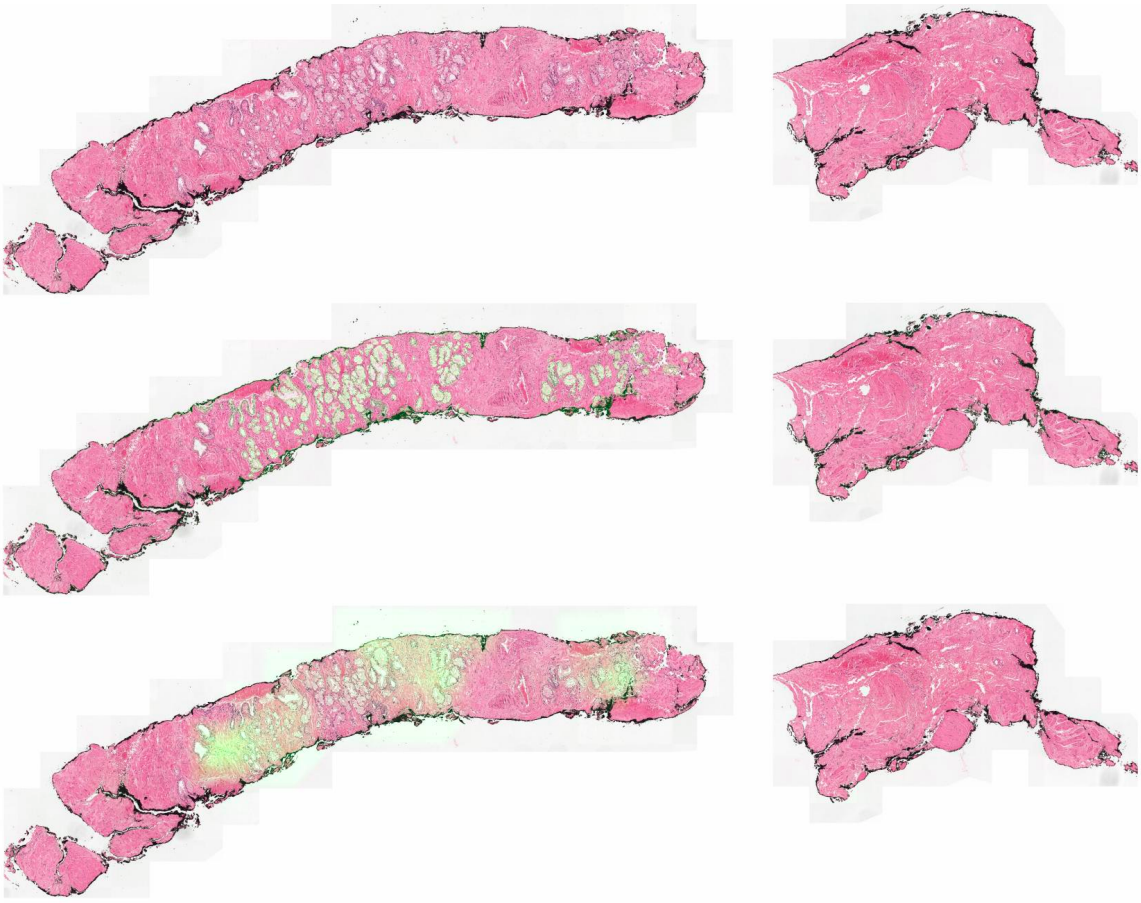


Figure 6: Attention weights for a test WSI of the PANDA dataset. The top image shows the original WSI, the middle image shows the cancerous areas marked by an expert pathologist (in green) and the bottom image shows the areas of high attention as predicted by the model (in green). The predicted attention weights were normalized and interpolated from the patch coordinates by linear interpolation. As can be seen in the image, the model successfully assigns high attention weights to the discriminative areas of the WSI. This improves explainability of the models prediction and helps the pathologist to find suspicious areas.

more false negatives (see first column of the matrices). This can also be confirmed by the binary F1 score (cancerous vs. non-cancerous): AGP outperforms all other approaches with a binary F1 score of 0.960 (Silva-R: 0.927, Mean-Agg: 0.950, A-Det: 0.958, A-Det-Gated: 0.956).

Next, we illustrate how the AGP attention mechanism provides an explainable prediction at instance level. The top row in Figure 6 shows one test WSI example (it has two pieces of tissue, a big one on the left and a small one on the right). In the second row, the cancerous areas are colored in green (this example has been manually segmented by an expert pathologist for this evaluation; recall that AGP only uses bag-level labels). The third row shows the areas with high attention weights predicted by AGP highlighted in green. For this figure, the attention weights were predicted by the model for each patch of the image (remember that the patches are of 512x512 resolution with 50% overlap). To obtain the heatmap for the complete WSI, linear interpolation was performed between the grid of patches. We find that the parts with high attention correspond to areas of the WSI that are most affected by cancer. Other parts that are non-cancerous, such as the whole piece on the right, are not assigned high attention weights. This means that AGP works as expected and the final prediction is based on discriminative areas. The attention helps the pathologist verify the prediction

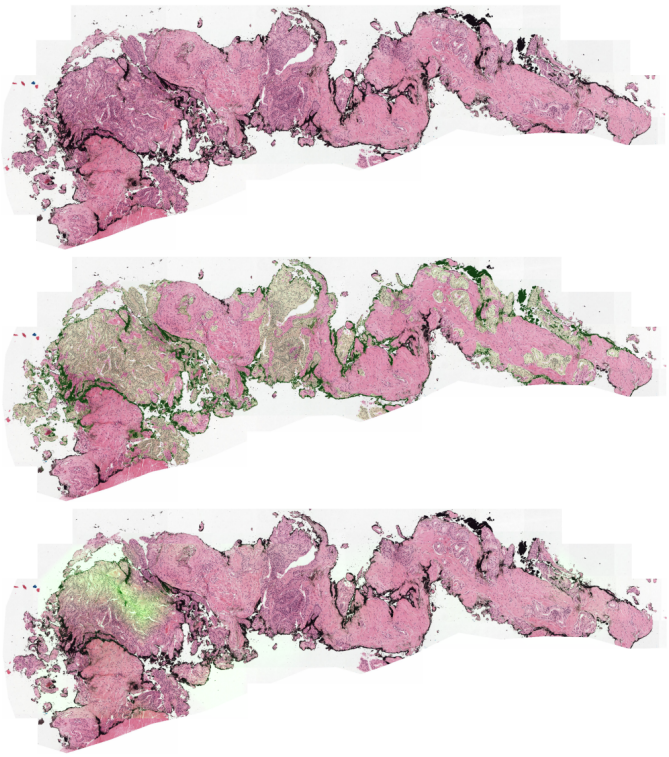


Figure 7: Example of an inaccurate assignment of attention weights. The top image shows the original WSI, the middle image shows the cancerous areas marked by an expert pathologist (in green), and the bottom image shows the areas of high attention as predicted by the model (in green). Although in most cases the areas of high attention correspond to the tumorous areas, recall Figure 6, we observed some inaccurate cases as the presented in this image. Here, the correct class (ISUP grade 4) is predicted, but the attention mechanism does not capture all the cancerous tissue parts.

and further inspect the affected tissue. Moreover, it might even point out cancerous regions that the pathologist may have missed.

Although the attention mechanism works well for most WSIs (as supported by the superior predictive performance), we observed that for some WSIs the attention does not capture all the important areas. In Figure 7 we show the same plots as in Figure 6 for a failure case of the attention mechanism (top: WSI, middle: annotation, bottom: attention estimation). In this case, not all the cancerous areas are assigned a high attention weight. However, the attention is still useful here as a source of explainability. It highlights all the areas which the classification is based on. Notice also that the highlighted areas are indeed tumorous, and the correct class (ISUP grade 4) is predicted.

Finally, we focus on the probabilistic bag predictions that provide not only a class score, given by the mean, but also the predictive uncertainty, given by the standard deviation. Figure 8 shows a histogram over the predictive uncertainty (i.e. the standard deviation) for all the AGP bag predictions. The green (resp. red) bars indicate the number of correctly (resp. incorrectly) predicted bags whose standard deviation falls in a certain range. It is clearly visible that correct predictions tend to have a lower standard deviation than the incorrectly classified bags. In other words: a high standard deviation correlates with a high risk of a wrong classification. This suggests that the standard deviation provides a useful measure of the predictive reliability. In fact, if we only take the reliable predictions with a std. below 0.02, the Cohen's kappa value for the PANDA test set rises to 0.864 for the AGP model (from 0.817 in Table 4). Therefore, in practice, the uncertainty estimation can help the pathologists decide when the models' prediction should be disregarded or double-checked.

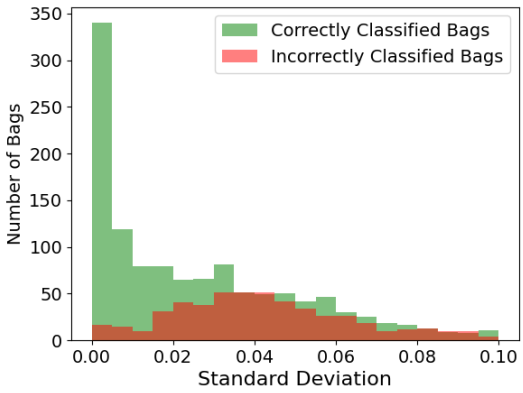


Figure 8: Distribution of the predicted standard deviations. Bags (images) with a low standard deviation are likely to be classified correctly, while a high standard deviation indicates a high risk of a wrong prediction. The standard deviations are divided into bins of width 0.005, and the y-axis shows the number of bags with the corresponding std.

Method	Learning	κ mean	κ S.E.
Mean-Agg	MIL	0.911	0.007
A-Det	MIL	0.903	0.004
A-Det-Gated	MIL	0.910	0.007
AGP	MIL	**0.920**	0.001
Silva-Rodríguez et al. [47]	MIL	0.885	N.A.

Table 5: Models trained on the PANDA dataset, tested on SICAPv2 with ISUP grading scale. The mean and standard error of Cohen's quadratic kappa (κ) are reported for four independent test runs. The last method does not report the standard error.

4.3 External Validation with PANDA and SICAPv2

The third experiment tests the generalization capability for the models evaluated in Section 4.2. Similar to [47], we take the models trained on the PANDA dataset, and use all images of SICAPv2 as an external test set (since the training is done under ISUP grading, SICAPv2 uses ISUP grading here too; therefore, results are not comparable to those obtained in Section 4.1). Table 5 shows the results. Again, AGP outperforms the rest of approaches, and achieves a remarkable Cohen's quadratic kappa value of 0.92.

Similar to Section 4.2, it is worth analyzing the standard deviation of the predictions. In addition to distinguishing between correctly and incorrectly classified images, there is now an additional dimension to consider: whether the test images follow the same distribution as the training ones or not (i.e. whether we are using PANDA or SICAPv2 as test set, respectively).

Table 6 shows the average predicted standard deviation of correct and wrong predictions for each of the test sets. The main findings are twofold: (1) the output standard deviation is higher for wrong predictions, and (2) the output standard deviation is higher for data originating from a different data distribution. This shows that the output uncertainty is not only helpful to identify model failures for in-distribution data, but also accounts for uncertainty added by a data shift. [51] This can help to determine images that might be out of scope for the model and should be handled with caution.

	Average uncertainty	
Test set	Correct predictions	Wrong predictions
PANDA	0.029	0.046
SICAPv2	0.045	0.051

Table 6: The average standard deviation of correct and wrong bag predictions for AGP trained on the PANDA dataset. The first (resp. second) row shows the result when using PANDA (resp. SICAPv2) as test set. The values, which reflect the uncertainty in the prediction, get higher for wrong predictions and when testing on a different test set.

5 Conclusions

We have proposed AGP, a novel probabilistic attention mechanism based on GPs for deep multiple instance learning (MIL). We have evaluated AGP in a wide range of experiments, including real-world cancer detection tasks. The novel attention module is capable of accurately assigning attention weights to the instances, and outperforms state-of-the-art deterministic attention modules. Furthermore, it provides important advantages due to its probabilistic nature. For instance, it addresses the problem of reliability in safety critical environments such as medicine: the probabilistic output of our model can be used to estimate the uncertainty on each prediction. For future research, we plan to explore the use of deep GPs (instead of GPs) to further improve the performance. Also, the promising results of AGP encourage the application of GP-based attention to other recent methods such as transformer networks, self-attention or channel attention. Moreover, alternative probabilistic attention mechanisms based on other Bayesian approaches (instead of GPs) can be explored.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. [Online]. Available: https://doi.org/10.1038/nature14539

[2] Z. Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[3] M. Mahmud, M. S. Kaiser, A. Hussain, and S. Vassanelli, "Applications of deep learning and reinforcement learning to biological data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2063–2079, 2018.

[4] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21, 2021.

[5] P. Morales-Álvarez, P. Ruiz, S. Coughlin, R. Molina, and A. K. Katsaggelos, "Scalable variational gaussian processes for crowdsourcing: Glitch detection in ligo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1534–1551, 2022.

[6] G. Quellec, G. Cazuguel, B. Cochener, and M. Lamard, "Multiple-instance learning for medical image and video analysis," *IEEE Reviews in Biomedical Engineering*, vol. 10, pp. 213–234, 2017.

[7] J. Melendez, B. van Ginneken, P. Maduskar, R. H. H. M. Philipsen, K. Reither, M. Breuninger, I. M. O. Adetifa, R. Maane, H. Ayles, and C. I. Sánchez, "A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest x-rays," *IEEE Transactions on Medical Imaging*, vol. 34, no. 1, pp. 179–192, 2015.

[8] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Medicine*, vol. 25, no. 8, pp. 1301–1309, 2019. [Online]. Available: http://www.nature.com/articles/s41591-019-0508-1

[9] Y. Wu, A. Schmidt, E. Hernández-Sánchez, R. Molina, and A. K. Katsaggelos, "Combining attention-based multiple instance learning and gaussian processes for CT hemorrhage detection," in *Medical Image Computing and Computer Assisted Intervention – MICCAI*, 2021, pp. 582–591.

[10] O. Ciga, T. Xu, S. Nofech-Mozes, S. Noy, F.-I. Lu, and A. L. Martel, "Overcoming the limitations of patch-based learning to detect cancer in whole slide images," *Scientific Reports*, vol. 11, no. 1, pp. 1–10, 2021.

[11] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in Neural Information Processing Systems*, vol. 15, 2003. [Online]. Available: https://proceedings.neurips.cc/paper/2002/file/3e6260b81898beacda3d16db379ed329-Paper.pdf

[12] Q. Zhang and S. Goldman, "EM-DD: An improved multiple-instance learning technique," in *Conference on Neural Information Processing Systems - NIPS*, vol. 14, 2002. [Online]. Available: https://proceedings.neurips.cc/paper/2001/file/e4dd5528f7596dcdf871aa55cfccc53c-Paper.pdf

[13] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li, "Multi-instance learning by treating instances as non-i.i.d. samples," in *International Conference on Machine Learning - ICML*, 2009, pp. 1–8. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1553374.1553534

[14] Y. Yan, X. Wang, X. Guo, J. Fang, W. Liu, and J. Huang, "Deep multi-instance learning with dynamic pooling," in *Asian Conference on Machine Learning - ACML*, ser. Proceedings of Machine Learning Research, vol. 95, 2018, pp. 662–677. [Online]. Available: https://proceedings.mlr.press/v95/yan18a.html

[15] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern Recognition*, vol. 74, pp. 15–24, 2018.

[16] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *International Conference on Machine Learning - ICML*, 2018, pp. 2127–2136.

[17] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," in *Conference on Computer Vision and Pattern Recognition - CVPR*, 2021, pp. 14 313–14 323. [Online]. Available: https://ieeexplore.ieee.org/document/9578683/

[18] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, 2021. [Online]. Available: https://doi.org/10.1038/s41551-020-00682-w

[19] Q. Kong, C. Yu, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, "Weakly labelled AudioSet tagging with attention neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1791–1802, 2019. [Online]. Available: http://arxiv.org/abs/1903.00765

[20] K. Chen, L. Yao, D. Zhang, X. Wang, X. Chang, and F. Nie, "A semisupervised recurrent convolutional attention model for human activity recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 5, pp. 1747–1756, 2020.

[21] D. Zhang, L. Yao, K. Chen, S. Wang, X. Chang, and Y. Liu, "Making sense of spatio-temporal preserving representations for eeg-based human intention recognition," *IEEE Transactions on Cybernetics*, vol. 50, no. 7, pp. 3033–3044, 2020.

[22] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning - ICML*, vol. 48, 2016, pp. 1050–1059.

[23] V. B. Alex Kendall and R. Cipolla, "Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," in *British Machine Vision Conference - BMVC*, 2017, pp. 57.1–57.12. [Online]. Available: https://dx.doi.org/10.5244/C.31.57

[24] P. Ruiz, P. Morales-Álvarez, R. Molina, and A. K. Katsaggelos, "Learning from crowds with variational gaussian processes," *Pattern Recognition*, vol. 88, pp. 298–311, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320318304060

[25] P. Morales-Álvarez, D. Hernández-Lobato, R. Molina, and J. M. Hernández-Lobato, "Activation-level uncertainty in deep neural networks," in *International Conference on Learning Representations - ICLR*, 2020.

[26] M. Kim and F. De la Torre, "Multiple instance learning via gaussian processes," *Data Mining and Knowledge Discovery*, vol. 28, no. 4, pp. 1078–1106, 2014. [Online]. Available: https://doi.org/10.1007/s10618-013-0333-y

[27] M. Haussmann, F. A. Hamprecht, and M. Kandemir, "Variational bayesian multiple instance learning with gaussian processes," in *Conference on Computer Vision and Pattern Recognition - CVPR*, 2017, pp. 810–819. [Online]. Available: https://ieeexplore.ieee.org/document/8099576/

[28] K. Chen and C.-G. Lee, "Attentive gaussian processes for probabilistic time-series generation," *ArXiv*, vol. abs/2102.05208, 2021.

[29] J. Xie, Z. Ma, D. Chang, G. Zhang, and J. Guo, "GPCA: A probabilistic framework for gaussian process embedded channel attention," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2021.

[30] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*. MIT press, 2006.

[31] P. Morales-Álvarez, A. Pérez-Suay, R. Molina, and G. Camps-Valls, "Remote sensing image classification with large-scale gaussian processes," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 1103–1114, 2017.

[32] D. H. Svendsen, P. Morales-Álvarez, A. B. Ruescas, R. Molina, and G. Camps-Valls, "Deep gaussian processes for biogeophysical parameter retrieval and model inversion," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 68–81, 2020.

[33] E. Snelson and Z. Ghahramani, "Sparse gaussian processes using pseudo-inputs," in *Advances in Neural Information Processing Systems*, vol. 18, 2006. [Online]. Available: https://proceedings.neurips.cc/paper/2005/file/4491777b1aa8b5b32c2e8666dbe1a495-Paper.pdf

[34] J. Hensman, A. Matthews, and Z. Ghahramani, "Scalable variational gaussian process classification," in *International Conference on Artificial Intelligence and Statistics - AIStat*, vol. 38, 2015, pp. 351–360. [Online]. Available: https://proceedings.mlr.press/v38/hensman15.html

[35] H. Liu, Y.-S. Ong, X. Shen, and J. Cai, "When gaussian process meets big data: A review of scalable GPs," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 11, pp. 4405–4423, 2020.

[36] P. Morales-Álvarez, P. Ruiz, R. Santos-Rodríguez, R. Molina, and A. K. Katsaggelos, "Scalable and efficient learning from crowds with gaussian processes," *Information Fusion*, vol. 52, pp. 110–127, 2019.

[37] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning - ICML*, 2019.

[38] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt, "Advances in variational inference," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 2008–2026, 2018.

[39] D. P. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," in *International Conference on Neural Information Processing Systems - NIPS*, 2015, p. 2575–2583.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations - ICLR*, 2015.

[41] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[42] A. Krizhevsky, V. Nair, and G. Hinton, "Cifar-10 (canadian institute for advanced research)." [Online]. Available: http://www.cs.toronto.edu/~kriz/cifar.html

[43] M. Vladimirova, J. J. Verbeek, P. Mesejo, and J. Arbel, "Understanding priors in bayesian neural networks at the unit level," in *ICML*, 2019.

[44] M. López-Pérez, M. Amgad, P. Morales-Álvarez, P. Ruiz, L. A. Cooper, R. Molina, and A. K. Katsaggelos, "Learning from crowds in digital pathology using scalable variational gaussian processes," *Scientific reports*, vol. 11, no. 1, pp. 1–9, 2021.

[45] J. I. Epstein, L. Egevad, M. B. Amin, B. Delahunt, J. R. Srigley, and P. A. Humphrey, "The 2014 international society of urological pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma: Definition of grading patterns and proposal for a new grading system," *American Journal of Surgical Pathology*, vol. 40, no. 2, pp. 244–252, 2016. [Online]. Available: https://journals.lww.com/00000478-201602000-00010

[46] J. Silva-rodríguez, A. Colomer, M. A. Sales, R. Molina, and V. Naranjo, "Going deeper through the Gleason scoring scale : An automatic end-to-end system for histology prostate grading and cribriform pattern detection," *Computer Methods and Programs in Biomedicine*, vol. 195, 2020.

[47] J. Silva-Rodríguez, A. Colomer, J. Dolz, and V. Naranjo, "Self-learning for weakly supervised gleason grading of local patterns," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 8, pp. 3094–3104, 2021. [Online]. Available: http://arxiv.org/abs/2105.10420

[48] A. Schmidt, J. Silva-Rodríguez, R. Molina, and V. Naranjo, "Efficient cancer classification by coupling semi supervised and multiple instance learning," *IEEE Access*, vol. 10, pp. 9763–9773, January 2022.

[49] J. Silva-Rodríguez, A. Colomer, M. A. Sales, R. Molina, and V. Naranjo, "Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection," *Computer Methods and Programs in Biomedicine*, vol. 195, p. 105637, 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S016926072031470X

[50] E. Arvaniti, K. S. Fricker, M. Moret, N. Rupp, T. Hermanns, C. Fankhauser, N. Wey, P. J. Wild, J. H. Rüschoff, and M. Claassen, "Automated gleason grading of prostate cancer tissue microarrays via deep learning," *Scientific Reports*, vol. 8, no. 1, p. 12054, 2018. [Online]. Available: http://www.nature.com/articles/s41598-018-30535-1

[51] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," *Advances in neural information processing systems - NeurIPS*, vol. 32, 2019.

Chapter 3

Gaussian Process Model with Instance Correlations for Multiple Instance Learning

3.1 Publication details

Authors: Arne Schmidt, Pablo Morales-Álvarez, José Miguel Hernández-Lobato, Rafael Molina

Title: Introducing Instance Label Correlation in Multiple Instance Learning. Application to Cancer Detection on Histopathological Images

Reference: Pattern Recognition (Accepted for publication)

Status: Accepted

Quality indices:

- Impact Factor (JCR 2022): 8.0
  - Rank 25/145 (Q1) in Computer Science, Artificial Intelligence
  - Rank 30/276 (Q1) in Engineering, Electrical and Electronical

3.2 Main contributions

- Instance correlations are introduced within a sparse Gaussian process model. The instance correlations correspond to the similarities of neighboring image patches - if a patch is surrounded by cancerous patches, it is most likely to be cancerous, too.
- The proposed model is inspired by the Ising model which describes magnetic fields in statistical physics.
- In extensive experiments we show an improved performance and the effective avoidance of false positive predictions in comparison to models that do not take the instance correlations into account.

INTRODUCING INSTANCE LABEL CORRELATION IN MULTIPLE INSTANCE LEARNING. APPLICATION TO CANCER DETECTION ON HISTOPATHOLOGICAL IMAGES

Pablo Morales-Álvarez
Department of Statistics and Operations Research
University of Granada
Granada, Spain

Arne Schmidt
Department of Computer Science and AI
University of Granada
Granada, Spain

José Miguel Hernández-Lobato
Department of Engineering
University of Cambridge
Cambridge, United Kingdom

Rafael Molina
Department of Computer Science and AI
University of Granada
Granada, Spain

ABSTRACT

In the last years, the weakly supervised paradigm of multiple instance learning (MIL) has become very popular in many different areas. A paradigmatic example is computational pathology, where the lack of patch-level labels for whole-slide images prevents the application of supervised models. Probabilistic MIL methods based on Gaussian Processes (GPs) have obtained promising results due to their excellent uncertainty estimation capabilities. However, these are general-purpose MIL methods that do not take into account one important fact: in (histopathological) images, the labels of neighboring patches are expected to be correlated. In this work, we extend a state-of-the-art GP-based MIL method, which is called VGPMIL-PR, to exploit such correlation. To do so, we develop a novel coupling term inspired by the statistical physics Ising model. We use variational inference to estimate all the model parameters. Interestingly, the VGPMIL-PR formulation is recovered when the weight that regulates the strength of the Ising term vanishes. The performance of the proposed method is assessed in two real-world problems of prostate cancer detection. We show that our model achieves better results than other state-of-the-art probabilistic MIL methods. We also provide different visualizations and analysis to gain insights into the influence of the novel Ising term. These insights are expected to facilitate the application of the proposed model to other research areas.

Keywords Multiple Instance Learning · Gaussian Processes · Ising model Variational Inference · Whole Slide Images · Histopathology

1 Introduction

Multiple instance learning (MIL) has caught great attention in fields where there is a challenging lack of labelled data. Although it has been applied in many different areas [1], we will focus on the case of computational pathology. In the last years, thanks to the increasing digitalization of whole-slide images (WSIs), the field of computational pathology is developing computer-aided diagnosis systems based on machine learning for cancer detection [2, 3]. The goal of computational

pathology is to provide a fast and reliable diagnosis for the most prototypical cases, letting the pathologists focus on the most challenging ones. Ultimately, this will enable a much wider access to early cancer diagnosis [4].

In order to make accurate predictions, machine learning classification methods need to be trained using a labelled set of instances [5]. In the case of computational pathology, these instances are typically patches from the WSIs (and not the complete images themselves) [6, 7, 8]. The reason for this is twofold: i) it is useful to have predictions at patch level in order to know where exactly in the image the cancer is located, and ii) WSIs are extremely large and cannot be directly fed to a classifier. As a consequence, notice that expert pathologists must label *every single patch* in the training data as cancerous or not (we will consider the binary problem cancer/no-cancer throughout this work). Given the large number of patches and the limited availability of pathologists, this becomes a daunting task in real practice [9].

To address this problem, different weakly supervised learning paradigms have been proposed in recent years. Here we focus on MIL, which has become very popular in the medical domain [10, 8]. The idea in MIL is that instances are grouped in bags, and only bag labels are needed for training. In the case of WSIs, all the patches coming from the same image are considered a bag. Therefore, the labelling workload on pathologists decreases enormously: from labelling every single patch, to only labelling the complete WSI as cancerous or not.

Different machine learning algorithms have been developed to learn under the MIL setting. Notice that dealing with uncertainty is essential in MIL models, since instance-level labels are unknown. To deal with uncertainties, different probabilistic methods have been developed, such as Dirichlet Process Mixture Models [11], Markov chain [12], Monte-Carlo chain [13, 14] and Gaussian Processes (GPs) [15, 16]. In particular, GPs have attracted plenty of attention in the last years, due to their expressive power and their capacity to handle uncertainty in a principled manner. Moreover, we are interested in this type of probabilistic models, since they will allow for introducing correlations in a theoretically sound way.

Among GP-based MIL methods, we will focus on the two most successful ones: VGPMIL and VGPMIL-PR. VGPMIL [17] was proposed in 2017 to overcome the limitations of two earlier formulations [15, 16] (namely, the use of the inefficient Laplace approximation and the impossibility to obtain instance-level predictions, respectively). In short, VGPMIL relies on variational inference and allows for closed-form updates of its parameters. However, the use of the logistic function implies that VGPMIL needs to resort to a theoretical approximation during inference (namely, the Jaakola bound [17, Eq. (10)]). As shown in [18], such approximation hurts predictive performance in practice. As an alternative, the authors of [18] propose the utilization of the probit function, which removes the need for the aforementioned approximation. This method, which will be referred to as VGPMIL-PR, is considered the current state of the art among probabilistic MIL approaches.

Methods such as VGPMIL and VGPMIL-PR are general-purpose MIL models that can be used in any MIL problem (that is, whenever the label is known only at bag level, see different use-cases in [17, 19]). However, the underlying MIL assumption that the labels of the instances in a bag are independent of each other is unrealistic in many real problems. For example, in the particular case of WSI images (and in many image-related MIL problems), the labels of neighboring patches are expected to be correlated [20]. We hypothesize that the predictive performance of MIL methods can be enhanced by incorporating this type of prior knowledge into the model.

In this work, we introduce a novel GP-based MIL algorithm that takes into account the correlation between the labels of neighboring patches, and we apply it to the real-world problem of prostate cancer detection on histopathological images. We model the correlation through a coupling term inspired by the Ising model [5, Section 19.4.1], an statistical physics method that has found several applications in computer vision [20, 21]. Our GP-MIL modeling builds on VGPMIL-PR, so our method will be referred to as VGPMIL-PR-I (Ising). In VGPMIL-PR-I, a hyperparameter λ regulates the influence of the Ising-inspired terms. Variational inference is used to estimate the model parameters, and the update formulas of VGPMIL-PR are recovered when $\lambda \to 0$ (that is, when

the influence of the coupling term vanishes). In the experimental section, we show that VGPMIL-PR-I outperforms the state-of-the-art GP-based MIL approaches VGPMIL and VGPMIL-PR when predicting at both instance and bag levels, while keeping an analogous computational cost. Moreover, to gain insights into the influence of the new coupling term, we analyze the role of λ, and provide several visualizations for the predictions.

The rest of the paper is organized as follows. Section 2 presents the probabilistic model and inference for the novel VGPMIL-PR-I. Closely related methods such as VGPMIL and VGPMIL-PR are also discussed in this section. Section 3 focuses on the empirical evaluation of the model, including the data description, the experimental framework, and the discussion of results. Section 4 provides the main conclusions and some future outlook.

2 Probabilistic model and inference

In this section we present the theoretical description for VGPMIL-PR-I. Specifically, Section 2.1 explains the problem formulation and the main notation. Section 2.2 explains the closely-related methods VGPMIL and VGPMIL-PR, which are at the base of our formulation. Section 2.3 introduces the novel coupling term that accounts for patch label correlation, which is used to define VGPMIL-PR-I. Section 2.4 shows how to perform variational inference to estimate the parameters in VGPMIL-PR-I. Section 2.5 explains the procedure to make predictions at both instance and bag levels.

2.1 Notation and problem formulation

Our notation follows the state-of-the-art work [22]. The training data is given by a set of bags $\mathbf{X} = \{\mathbf{X}_b\}_{b\in\mathcal{B}}$ and their corresponding labels $\mathbf{y} = \{y_b\}_{b\in\mathcal{B}}$. We deal with a binary problem, i.e. $y_b \in \{0, 1\}$. Each bag $\mathbf{X}_b = \{\mathbf{x}_i\}_{i\in b}$ contains $|b|$ instances, i.e. $b = \{i_1, \ldots, i_{|b|}\} \subseteq [N]$ (N is the total amount of instances). Notice that different bags may have different amounts of instances. Each instance $\mathbf{x}_i$ is given by a vector in $\mathbb{R}^D$. In the MIL setting, one assumes that each instance has its (unknown) label $h_i \in \{0, 1\}$. We write $\mathbf{h}_b$ for the labels of all the instances belonging to bag b. The MIL labelling assumption dictates that a bag is considered positive (class 1) if at least one of its instances is positive. Mathematically, this is

$$\mathrm{p}(y_b|\mathbf{h}_b) = 1[y_b = \max_{i\in b} h_i], \tag{1}$$

where $1[\cdot]$ is the indicator function (i.e. it equals one when its argument is true and zero otherwise). Finally, we will collectively denote $\mathbf{h} = \{\mathbf{h}_b\}_{b\in\mathcal{B}}$.

In the case of WSIs, each $\mathbf{X}_b$ is an image, which is composed of its patches $\{\mathbf{x}_i\}_{i\in b}$. Each patch has an unknown label h_i (0 for non-cancerous and 1 for cancerous), and we only have access to the bag label y_b (whether the image is cancerous or not, i.e. whether it contains at least one patch that is cancerous).

The goal in MIL is to train a model based only on bag labels $\{y_b\}_{b\in\mathcal{B}}$. And such model must be able to predict at both instance and bag levels. That is, given a previously unseen instance $\mathbf{x}^\star \in \mathbb{R}^D$, we are interested in the probability $\mathrm{p}(h^\star = 1)$. Likewise, given a previously unseen complete bag $\mathbf{X}^\star$, we are interested in $\mathrm{p}(y^\star = 1)$.

2.2 Background: VGPMIL and VGPMIL-PR

As mentioned in the introduction, our model is inspired by two closely related methods: VGPMIL and VGPMIL-PR. To understand our contribution, it is essential to fully understand their formulations, which we explain next.

2.2.1 VGPMIL formulation

VGPMIL was introduced in [17]. The idea is to consider a sparse GP classification model [23] to describe the relationship between instance features $\mathbf{X}$ and their (unknown) labels $\mathbf{h}$. Then, an additional bag likelihood must be considered to model the (observed) bag labels $\mathbf{y}$ given the instance labels $\mathbf{h}$. Both components are described next.

The sparse GP classification model. Instances $\mathbf{x}_i$ are associated latent variables $f_i \in \mathbb{R}$ which are modelled through a GP, $f \sim \mathcal{GP}(0, \kappa)$. We write κ for the GP kernel, which encodes the properties of the considered functions. Then, the instance labels h_i are defined from f_i through a classification likelihood ν:

$$\mathrm{p}(h_i|f_i) = \nu(f_i)^{h_i}(1-\nu(f_i))^{1-h_i}. \tag{2}$$

Specifically, VGPMIL uses the logistic function $\nu(x) = (1+e^{-x})^{-1}$. Intuitively, a large (resp. low) value of f_i implies that the class is likely to be one (resp. zero). Moreover, since standard GPs scale poorly with the number of training instances N, VGPMIL makes use of sparse GPs [23], which summarize the training data through $M \ll N$ inducing points. These inducing points $\mathbf{u} = \{u_1, \ldots, u_M\}$ represent the value of the GP at some inducing points locations $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_M\}$ (just like $\mathbf{f} = \{f_1, \ldots, f_N\}$ are the GP values at $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$). Therefore, the distributions of $\mathbf{u}$ and $\mathbf{f}|\mathbf{u}$ are:

$$\mathrm{p}(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K_{ZZ}}), \tag{3}$$

$$\mathrm{p}(\mathbf{f}|\mathbf{u}) = \mathcal{N}(\mathbf{f}|\mathbf{K_{XZ}}\mathbf{K_{ZZ}^{-1}}\mathbf{u}, \mathcal{K}). \tag{4}$$

The expression of $\mathcal{K}$ is given by the particular sparse GP approach used in VGPMIL, which is FITC [23], so we have $\mathcal{K} = \mathrm{diag}(\mathbf{K_{XX}} - \mathbf{K_{XZ}}\mathbf{K_{ZZ}^{-1}}\mathbf{K_{ZX}})$. As is standard in GP literature, we are writing $\mathbf{K_{XX}}$ for the $N \times N$ covariance matrix $\mathbf{K_{XX}} = (\kappa(\mathbf{x}_i, \mathbf{x}_j))_{1 \leq i,j \leq N}$. The definitions for $\mathbf{K_{XZ}}$ and $\mathbf{K_{ZZ}}$ are analogous.

The bag likelihood. VGPMIL introduces the following parameterization to model the bag labels from the instance labels:

$$\mathrm{p}(y_b|\mathbf{h}_b) = \frac{H^{G_b}}{H+1}, \tag{5}$$

where $G_b := 1[y_b = \max_{i \in b} h_i]$ and H is a large and fixed value (in their examples, they use $H = 100$). Eq. (5) approximates the MIL assumption introduced in eq. (1): if some instance label h_i is one, then the bag label y_b is one with very high probability (namely, with probability $\frac{H}{H+1}$). Otherwise (that is, if all instance labels h_i are zero), the bag label is one with very low probability (namely, $\frac{1}{H+1}$).

In summary, VGPMIL is given by eqs. (3), (4), (2) and (5). For additional details, the interested reader is referred to the original work [17].

2.2.2 VGPMIL-PR formulation

VGPMIL-PR was recently proposed in [18] as an improvement over VGPMIL. Namely, the logistic function used by VGPMIL in eq. (2) is not conjugate with the Gaussian distribution coming from the GP, recall eqs. (3)–(4). This means that, in order to achieve mathematical tractability, VGPMIL needs to resort to the Jaakola bound [17, Eq. (10)]. However, the use of this bound introduces an *approximation* in the training objective. As shown in [18], such approximation damages the predictive performance in practice. Consequently, the authors of [18] introduce an alternative formulation based on the probit function, VGPMIL-PR. They show that, via a variable augmentation approach, VGPMIL-PR allows for directly optimizing the training objective (without approximations).

More specifically, VGPMIL-PR uses the probit function $\nu(x) = \int_{-\infty}^{x} \mathcal{N}(t|0,1)\mathrm{d}t$ in eq. (2). Also, the bag likelihood is given by eq. (1) (instead of eq. (5)). Then, to circumvent the need for approximations, VGPMIL-PR leverages a variable augmentation approach [24]. Namely, for each instance we introduce a new variable $m_i \in \mathbb{R}$ between f_i and h_i, which is defined as $m_i \sim \mathcal{N}(f_i, 1)$.

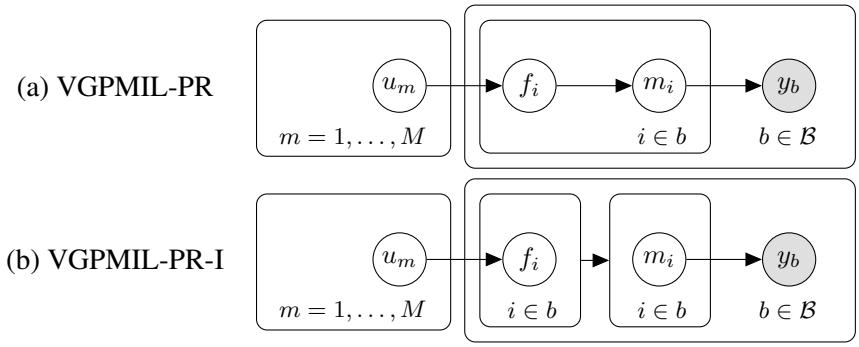


Figure 1: Probabilistic graphical model for VGPMIL-PR (a) and VGPMIL-PR-I (b). Gray nodes are observed variables, and white ones are latent variables to be estimated. The main difference is that VGPMIL-PR-I introduces correlation between instances in the same bag. Therefore, the distribution of $\mathbf{m}_b$ given $\mathbf{f}_b$ does not factorize across instances. The correlation is introduced through a novel term inspired by the Ising model, see Section 2.3 for details.

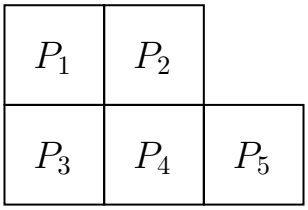


Figure 2: A simplified illustration of an image composed by five patches: $P1, \ldots, P5$.

Since we are using a probit likelihood, we have that $h_i = 1[m_i > 0]$. Analogously to the rest of variables, we write $\mathbf{m}_b = \{m_i\}_{i \in b}$ for all the m_i's inside bag b, and we use $\mathbf{m} = \{\mathbf{m}_b\}_{b \in \mathcal{B}}$ to collectively denote all the m_i's in the model. Then, by marginalizing out $\mathbf{h}$, we have:

$$\mathrm{p}(\mathbf{m}|\mathbf{f}) = \prod_b \mathrm{p}(\mathbf{m}_b|\mathbf{f}_b) = \prod_b \mathcal{N}(\mathbf{m}_b|\mathbf{f}_b, I), \tag{6}$$

$$\mathrm{p}(\mathbf{y}|\mathbf{m}) = \prod_b \mathrm{p}(y_b|\mathbf{m}_b), \tag{7}$$

where I is the identity matrix (of size $|b|$) and $\mathrm{p}(y_b = 0|\mathbf{m}_b) = \prod_{i \in b} 1[m_i < 0]$. Importantly, these augmented variables $\mathbf{m}$ will prove extremely helpful to introduce the Ising correlation in the next section.

In summary, VGPMIL-PR is given by eqs. (3), (4), (6) and (7). Figure 1(a) shows the probabilistic graphical model for VGPMIL-PR.

2.3 Correlating patch labels: VGPMIL-PR-I

As mentioned in the introduction, VGPMIL and VGPMIL-PR are general MIL approaches that can be used in any MIL problem. Indeed, there are plenty of applications where MIL methods can be used. For instance, think of a recommendation system where a reviewer has not evaluated every single item in the database, but has reviewed "groups" of them (e.g., he/she likes science-fiction movies, although he/she has not rated individual movies). Consider also a task of anomaly detection in which we do not have labels for individual transactions, but we only know whether there was some anomalous behavior in a certain period of time (which contains many different transactions).

Here we focus in the particular use-case of images, where bags are images and their instances are their patches. In this case, there exists very valuable information coming from the structure of the image itself, which can be exploited in the model. For example, it is natural to think that neighboring

patches in the same image are likely to have similar labels. The main goal of this work is to introduce such correlation into the VGPMIL-PR formulation. To do so, we are inspired by the Ising model.

The novel coupling term. The Ising model arose from statistical physics to describe the behavior of magnets. In some magnets, called ferro-magnets, neighboring spins tend to line up in the same direction, whereas in other kinds of magnets, called anti-ferromagnets, the spins are repelled from their neighbors [5]. This type of interactions based on the Ising model have been used previously in machine learning and computer vision to describe relationships between pixels of an image, see e.g. [20, 21, 5]. However, to the best of our knowledge, they have never been used in the context of MIL.

Our first idea was to consider an Ising model over the patch labels of each image, $\{h_i\}_{i\in b}$. However, inference proved very challenging in this case, due to the non-conjugacy of the Ising model and the GP-based MIL formulation. As an alternative, we considered a continuous counterpart of the Ising model over the variables $\mathbf{m}_b = \{m_i\}_{i\in b}$ introduced in VGPMIL-PR, which are directly related to the patch labels (recall from Section 2.2.2 that $h_i = 1[m_i > 0]$). Notice that such continuous version of the Ising model corresponds to the well-known Conditional Autoregression (CAR) [25]. Importantly, as we will see in the rest of this section, this alternative formulation yields a Gaussian distribution on $\mathbf{m}_b$, which can be treated analytically together with the GP-based MIL model.

Specifically, we consider the following coupling term for each image, which is defined over the augmented variables $\mathbf{m}_b$, recall Section 2.2.2:

$$\begin{aligned}\mathcal{C}(\mathbf{m}_b) &= \exp\left(-\frac{\lambda}{2}\cdot\sum_{\substack{i,j\in b\\ i<j}}\mathbf{1}[i,j \text{ are contiguous}]\cdot(m_i-m_j)^2\right) = \\ &= \exp\left(-\frac{\lambda}{2}\mathbf{m}_b^\intercal\mathbf{C}_b\mathbf{m}_b\right). \end{aligned} \tag{8}$$

Notice that $\mathcal{C}(\mathbf{m}_b)$ is always in the range $[0, 1]$, and it becomes close to zero when the value of m is very different for neighboring patches. For the second equality in eq. (8), notice that the sum only produces quadratic terms in m, so it can be written as $\mathbf{m}_b^\intercal\mathbf{C}_b\mathbf{m}_b$ for some positive semidefinite matrix $\mathbf{C}_b$.

Since m determines the label of each patch (recall from Section 2.2.2 that $h_i = 1[m_i > 0]$), the term $\mathcal{C}(\mathbf{m}_b)$ can be used to favor "smoothness" in the labels associated to the different patches. Also, the hyperparameter λ, which can be set to any non-negative value, regulates the strength of the coupling term: the larger λ, the more importance is given to differences in m. For example, when $\lambda = 0$, $\mathcal{C}(\mathbf{m}_b)$ becomes constant and it does not account for correlation between patch labels.

An example of $\mathcal{C}(\mathbf{m}_b)$. To illustrate the proposed coupling term, consider an image with five patches $P_1, \dots, P_5$ distributed as in Figure 2. In this case, the quadratic terms of $\mathbf{m}_b$ are:

$$(m_1-m_2)^2 + (m_1-m_3)^2 + (m_2-m_4)^2 + (m_3-m_4)^2 + (m_4-m_5)^2, \tag{9}$$

and therefore we have:

$$\mathbf{C}_b = \begin{pmatrix} 2 & -1 & -1 & 0 & 0\\ -1 & 2 & 0 & -1 & 0\\ -1 & 0 & 2 & -1 & 0\\ 0 & -1 & -1 & 3 & -1\\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}. \tag{10}$$

In general, it is easy to compute the matrix $\mathbf{C}_b$ for any given image. Notice that it is always a positive semidefinite matrix, and thus it is associated to a (singular) normal distribution.

The VGPMIL-PR-I formulation. To introduce the new coupling term $\mathcal{C}(\mathbf{m}_b)$ in the MIL formulation, we modify eq. (6) and define:

$$\mathrm{p}(\mathbf{m}|\mathbf{f}) \propto \prod_b \mathcal{C}(\mathbf{m}_b)\cdot\mathcal{N}(\mathbf{m}_b|\mathbf{f}_b, I). \tag{11}$$

Notice that the probability of a configuration $\mathbf{m}_b$ is proportional to the coupling term $\mathcal{C}(\mathbf{m}_b)$, which favors smoothness across labels of neighboring patches. Decisively, since both $\mathcal{C}(\mathbf{m}_b)$ and $\mathcal{N}(\mathbf{m}_b|\mathbf{f}_b, I)$ only contain (the exponential of) quadratic terms in $\mathbf{m}_b$, the new distribution can be written as a Gaussian:

$$\mathrm{p}(\mathbf{m}|\mathbf{f}) = \prod_b \mathcal{N}(\mathbf{m}_b|\boldsymbol{\Sigma}_b\mathbf{f}_b, \boldsymbol{\Sigma}_b), \tag{12}$$

with $\boldsymbol{\Sigma}_b = (\lambda\mathbf{C}_b + I)^{-1}$. Notice that this new formulation provides a generalization of VGPMIL-PR. Namely, when $\lambda \to 0$, we have that $\boldsymbol{\Sigma}_b \to I$ and we recover eq. (6).

In summary, the proposed model is given by eqs. (3), (4), (12), and (7). Notice also that, instead of FITC, in eq. (4) we leverage the more recent sparse GP approach introduced in [26]. Basically, this means that $\mathcal{K} = \mathbf{K_{XX}} - \mathbf{K_{XZ}}\mathbf{K_{ZZ}^{-1}}\mathbf{K_{ZX}}$ in eq. (4). Our method will be referred to as VGPMIL-PR-I (I denotes Ising). The probabilistic graphical model is depicted in Figure 1(b).

2.4 Variational inference

In order to make inference in the proposed model, we need to compute the posterior distribution $\mathrm{p}(\mathbf{u}, \mathbf{f}, \mathbf{m}|\mathbf{y})$. However, this is not analytically tractable due to the definition of the bag likelihood in eq. (7), which depends on the sign of the m_i's. Following [17] and [18], we leverage standard mean-field variational inference (VI) theory [27, Section 10.1.1] to calculate an approximate posterior distribution that factorizes as

$$\mathrm{q}(\mathbf{u}, \mathbf{f}, \mathbf{m}) = \mathrm{q}(\mathbf{u})\mathrm{p}(\mathbf{f}|\mathbf{u})\mathrm{q}(\mathbf{m}). \tag{13}$$

Applying the well-known mean-field VI update equation [27, Eq. (10.9)], we have that $\mathrm{q}(\mathbf{u})$ and $\mathrm{q}(\mathbf{m})$ can be iteratively computed as

$$\mathrm{q}(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\boldsymbol{\mu}^u, \boldsymbol{\Sigma}^u), \tag{14}$$

$$\mathrm{q}(\mathbf{m}) \propto \prod_b \mathrm{p}(y_b|\mathbf{m}_b)\mathcal{N}(\mathbf{m}_b|\boldsymbol{\mu}^{\mathbf{m}_b}, \boldsymbol{\Sigma}_b), \tag{15}$$

where

$$\boldsymbol{\Sigma}^u = \left(\mathbf{K_{ZZ}^{-1}} + \mathbf{K_{ZZ}^{-1}}\mathbf{K_{ZX}}\boldsymbol{\Sigma}\mathbf{K_{XZ}}\mathbf{K_{ZZ}^{-1}}\right)^{-1}, \tag{16}$$

$$\boldsymbol{\mu}^u = \boldsymbol{\Sigma}^u\mathbf{K_{ZZ}^{-1}}\mathbf{K_{ZX}}\mathbb{E}_{\mathrm{q}(\mathbf{m})}(\mathbf{m}), \tag{17}$$

and

$$\boldsymbol{\mu}^{\mathbf{m}_b} = \boldsymbol{\Sigma}_b\mathbf{K}_{b\mathbf{Z}}\mathbf{K_{ZZ}^{-1}}\boldsymbol{\mu}^u. \tag{18}$$

Here we are writing $\boldsymbol{\Sigma}$ for the $N \times N$ block-diagonal matrix that contains all the $\boldsymbol{\Sigma}_b$'s, $b \in \mathcal{B}$. Also, we are writing $\mathbf{K}_{b\mathbf{Z}}$ for the $|b| \times M$ matrix of covariances between $\mathbf{X}_b$ and $\mathbf{Z}$. Very importantly, notice that these update rules generalize those derived in [18] for VGPMIL-PR. Namely, when $\lambda \to 0$, we have that $\boldsymbol{\Sigma}_b, \boldsymbol{\Sigma} \to I$, and then eqs. (14)–(18) match eqs.(15)–(19) in [18].

All the computations involved in eqs. (14)–(18) are straightforward, except for $\mathbb{E}_{\mathrm{q}(\mathbf{m})}(\mathbf{m})$. Indeed, each $\mathrm{q}(\mathbf{m}_b)$ is a multivariate Gaussian truncated to $(-\infty, 0)^{|b|}$ (or $\mathbb{R}^{|b|} \setminus (-\infty, 0)^{|b|}$, depending on whether $y_b = 0$ or $y_b = 1$, respectively). It is well-known that the expectation of a truncated multivariate Gaussian cannot be obtained in closed-form [28]. Notice that this is not an issue for VGPMIL-PR [18], where the absence of Ising terms implies dealing with *univariate* Gaussians, whose expectations can be analytically computed.

To overcome the problem, we first tried to leverage numerical methods proposed in [29] to approximate the expectation for the multivariate truncated case. However, these methods proved computationally too expensive to be integrated within our iterative calculation of $\mathrm{q}(\mathbf{u})$ and $\mathrm{q}(\mathbf{m})$. Therefore, we decided to approximate the multivariate Gaussian $\mathcal{N}(\mathbf{m}_b|\boldsymbol{\mu}^{\mathbf{m}_b}, \boldsymbol{\Sigma}_b)$ by the factorized $\mathcal{N}(\mathbf{m}_b|\boldsymbol{\mu}^{\mathbf{m}_b}, \mathrm{diag}(\boldsymbol{\Sigma}_b))$ and utilize the expression for one-dimensional truncated Gaussians. Although such approximation reduces the influence of the Ising correlation at this specific computation,

Algorithm 1 Training procedure for VGPMIL-PR-I.

Input : Bags $\mathbf{X} = \{\mathbf{X}_b\}_{b\in\mathcal{B}}$ and bag labels $\mathbf{y} = \{y_b\}_{b\in\mathcal{B}}$.
Calculate the matrices $\mathbf{C}_b$ that account for the instance correlation inside each bag $b \in \mathcal{B}$, recall eq. (8) and the example at eq. (10).
Initialize GP kernel parameters and inducing points locations, as well as the posterior distributions $\mathrm{q}(\mathbf{u})$ and $\mathrm{q}(\mathbf{m})$. Details on initializations in the text.
foreach iteration $t = 1, \ldots, T$ **do**
  Update $\mathrm{q}(\mathbf{u})$, using eqs. (14), (16) and (17).
  Update $\mathrm{q}(\mathbf{m})$ and obtain $\mathbb{E}_{\mathrm{q}(\mathbf{m})}(\mathbf{m})$, using eqs. (15), (18), (19), and (20).
Output : Posterior distributions $\mathrm{q}(\mathbf{u})$ and $\mathrm{q}(\mathbf{m})$.

notice that the coupling terms, which are included in $\boldsymbol{\Sigma}_b$, affect the update equations in more places across eqs. (16)–(18).

Specifically, the expression for each $\mathbb{E}_{\mathrm{q}(\mathbf{m}_b)}(\mathbf{m}_b)$ is as follows. For bags with $y_b = 0$, we have that each $\mathrm{q}(m_i)$, $i \in b$, is a univariate normal distribution $\mathcal{N}((\boldsymbol{\mu}^{\mathbf{m}_b})_i, (\boldsymbol{\Sigma}_b)_{ii})$ truncated to $(-\infty, 0)$. The expectation of such a distribution is well-known and can be obtained in closed-form [30]:

$$E_i = \mu_i - \frac{\phi(\mu_i/\sigma_i)}{1 - \Phi(\mu_i/\sigma_i)}\sigma_i, \tag{19}$$

where ϕ and Φ are, respectively, the density and cumulative distribution functions of a standard Gaussian $\mathcal{N}(0, 1)$ (recall that both are efficiently implemented in standard software packages such as Python's Scipy). We have also abbreviated $\mu_i = (\boldsymbol{\mu}^{\mathbf{m}_b})_i$ and $\sigma_i = \sqrt{(\boldsymbol{\Sigma}_b)_{ii}}$. For bags with $y_b = 1$, we proceed analogously to [19] to obtain the normalization constant Z of the distribution of interest (that is, the factorized Gaussian $\mathcal{N}(\mathbf{m}_b|\boldsymbol{\mu}^{\mathbf{m}_b}, \mathrm{diag}(\boldsymbol{\Sigma}_b))$ truncated to $\mathbb{R}^{|b|} - (-\infty, 0)^{|b|}$). Then, the expectation of each $\mathrm{q}(m_i)$, $i \in b$, is given by:

$$\mathbb{E}_{\mathrm{q}(m_i)}(m_i) = \frac{\mu_i - (1 - Z)E_i}{Z}, \tag{20}$$

where $Z = 1 - \prod_{i\in b}(1 - \Phi(\mu_i/\sigma_i))$, E_i is given by eq. (19), and we are again abbreviating $\mu_i = (\boldsymbol{\mu}^{\mathbf{m}_b})_i$ and $\sigma_i = \sqrt{(\boldsymbol{\Sigma}_b)_{ii}}$.

The full training algorithm is summarized in Algorithm 1. It is an iterative process that alternates the updates between $\mathrm{q}(\mathbf{u})$ and $\mathrm{q}(\mathbf{m})$. Details on the GP kernel and initializations used in this work are provided in Section 3.1. The code for the proposed method will be publicly available upon acceptance of the paper.

2.5 Making predictions

Suppose we are given a new bag $\mathbf{X}^\star = \{\mathbf{x}_i^\star\}_{i\in b_\star}$. As explained at the end of section 2.1, we are interested in both instance-level and bag-level predictions. For this, we first need to compute the predictive distributions over $\mathbf{m}^\star$.

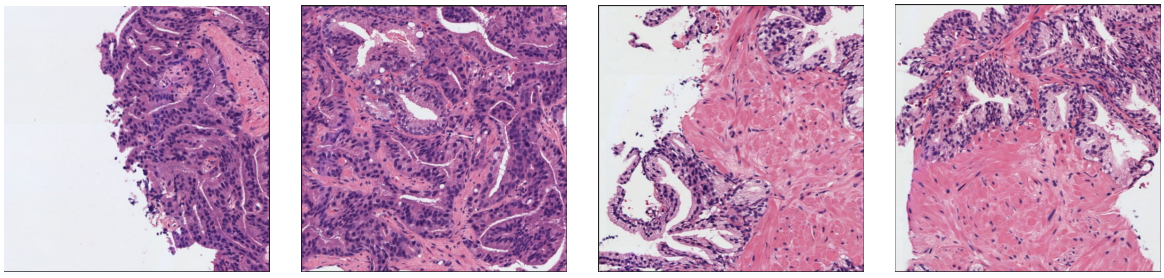
By using the learned posterior $\mathrm{q}(\mathbf{u})$ along with $\mathrm{p}(\mathbf{f}|\mathbf{u})$, we can obtain the joint distribution over $\mathbf{f}^\star$:

$$\mathrm{p}(\mathbf{f}^\star) = \int \mathrm{p}(\mathbf{f}^\star|\mathbf{u})\mathrm{p}(\mathbf{u})\mathrm{d}\mathbf{u} = \mathcal{N}(\mathbf{f}^\star|\boldsymbol{\mu}^\star, \mathbf{S}^\star), \tag{21}$$

with $\boldsymbol{\mu}^\star$ and $\mathbf{S}^\star$ given by the standard sparse GP predictions:

$$\boldsymbol{\mu}^\star = \mathbf{K}_{\star\mathbf{Z}}\mathbf{K}_{\mathbf{ZZ}}^{-1}\boldsymbol{\mu}^u, \quad \mathbf{S}^\star = \mathbf{K}_{\star\star} - \mathbf{K}_{\star\mathbf{Z}}\mathbf{K}_{\mathbf{ZZ}}^{-1}(\mathbf{K}_{\mathbf{ZZ}} - \boldsymbol{\Sigma}^u)\mathbf{K}_{\mathbf{ZZ}}^{-1}\mathbf{K}_{\mathbf{Z}\star}. \tag{22}$$

Here, $\boldsymbol{\mu}^u$ and $\boldsymbol{\Sigma}^u$ are the parameters learned during training, recall eq. (16) and (17). Naturally, the subscript $\star$ in the kernel matrices $\mathbf{K}$ indicates that we are using the new bag $\mathbf{X}^\star$. Then, since the



Cancerous Non-cancerous

Figure 3: Two examples of cancerous (left) and non-cancerous (right) patches in the SICAPv2 test set.

distribution $\mathrm{p}(\mathbf{m}|\mathbf{f})$ is also Gaussian, recall eq. (12), we can compute the joint distribution over $\mathbf{m}^\star$ in closed-form:

$$\mathrm{p}(\mathbf{m}^\star) = \int \mathrm{p}(\mathbf{m}^\star|\mathbf{f}^\star)\mathrm{p}(\mathbf{f}^\star)\mathrm{d}\mathbf{f}^\star = \mathcal{N}(\mathbf{m}^\star|\boldsymbol{\mu}_m^\star, \mathbf{S}_m^\star), \tag{23}$$

with $\boldsymbol{\mu}_m^\star$ and $\mathbf{S}_m^\star$ given by

$$\boldsymbol{\mu}_m^\star = \boldsymbol{\Sigma}_\star \cdot \boldsymbol{\mu}^\star, \quad \mathbf{S}_m^\star = \boldsymbol{\Sigma}_\star + \boldsymbol{\Sigma}_\star \cdot \mathbf{S}^\star \cdot \boldsymbol{\Sigma}_\star^\intercal. \tag{24}$$

Here, $\boldsymbol{\mu}^\star$ and $\mathbf{S}^\star$ are given by eq. (22), and $\boldsymbol{\Sigma}_\star$ is the matrix that accounts for correlation among instances in the test bag $\mathbf{X}^\star$, which is defined analogously to the training case, recall matrix $\boldsymbol{\Sigma}_b$ in eq. (12).

Once we have the joint distribution over $\mathbf{m}^\star$, the instance-level and bag-level predictions are given as:

$$\mathrm{p}(h_i^\star = 1) = \mathrm{p}(m_i^\star > 0) = \Phi\left((\boldsymbol{\mu}_m^\star)_i / \sqrt{(\mathbf{S}_m^\star)_{ii}}\right), \tag{25}$$

$$\mathrm{p}(y^\star = 1) = 1 - \int_{\mathbf{m}^\star \in (-\infty, 0)^{|b_\star|}} \mathrm{p}(\mathbf{m}^\star)\mathrm{d}\mathbf{m}^\star. \tag{26}$$

Notice that the integral in eq. (26) can be computed efficiently with the cumulative distribution function of a multivariate Gaussian, which is also available in most standard statistical packages, such as Python's Scipy.

Interestingly, these predictions generalize those obtained in VGPMIL-PR [18]. Indeed, if we do not consider correlation among instances in the test bag, i.e. $\boldsymbol{\Sigma}_\star = \mathbf{I}$, then eqs. (25) and (26) match those in [18] (last two equations before Section 3.5). Finally, although here we have detailed how to make predictions for a complete previously unseen bag $\mathbf{X}^\star$, the same process can be applied to make predictions on previously unseen individual instances $\mathbf{x}^\star$ (patches).

3 Experiments

In this section we thoroughly evaluate VGPMIL-PR-I in a real-world problem of prostate cancer detection. The experimental framework, including data, metrics and baselines, is explained in Section 3.1. The results are discussed in Section 3.2. Finally, in Section 3.3 we evaluate our method in a much larger prostate cancer detection dataset: the well-known PANDA challenge.

3.1 Experimental framework

Data description. In this paper we focus on the problem of prostate cancer detection. However, notice that the algorithm can be applied for any other type of cancer (and more generally, for any other type of image). Prostate cancer is the most commonly occurring cancer in men, and the second most commonly occurring cancer overall, according to the latest 2020 statistics on age-standardized

	λ	Accuracy	Precision	Recall	F1-score
VGPMIL	-	92.22±0.00	96.40±0.00	92.29±0.00	94.30±0.00
VGPMIL-PR	-	92.38±0.03	96.44±0.06	92.48±0.06	94.42±0.02
VGPMIL-PR-I	0.1	92.94±0.05	96.24±0.15	93.52±0.12	94.86±0.04
	0.5	**93.85±0.04**	**97.17±0.19**	**93.90±0.14**	**95.51±0.02**
	1.0	**94.58±0.03**	**97.32±0.06**	**94.83±0.04**	**96.06±0.02**
	5.0	**95.11±0.06**	**97.74±0.14**	**95.18±0.09**	**96.44±0.04**
	10.0	**95.03±0.14**	**97.72±0.16**	**95.09±0.06**	**96.39±0.10**

Table 1: Predictive performance at the level of patches (instances). In bold, we highlight the values of λ for which VGPMIL-PR-I gets better (or equal) performance than both baselines in all the metrics. The results are the mean and standard deviation over five independent runs.

incidence rate from the World Health Organisation (WHO) Global Cancer Observatory [31]. We will use the prostate cancer database presented in [32], which is called SICAPv2 and is publicly available. Although this database includes information on the Gleason score, which is used to evaluate the severity of the disease, in this work we will focus on the binary task of presence/absence of cancer.

We use the original partition of the dataset, which contains 95 training and 31 test WSIs, respectively. These very large images are split in 512x512 patches, resulting in a total amount of 15132 patches for training and 5246 for testing. Following the MIL paradigm, for the training set we only use binary labels benign/malign at the level of images (bags), but we do not have information at the level of patches (instances). In order to evaluate the predictive performance at instance-level, we do have labels for the patches in the test set. The amount of cancerous (resp. non-cancerous) images for the train set is 70 (resp. 25). For the test set, it is 25 (resp. 6). For illustration purposes, a couple of cancerous and non-cancerous patches are shown in Figure 3. In order to train our model, each patch is represented through a 128-dimensional feature vector extracted in previous work [7].

Baselines and metrics. Since our model is framed in the field of probabilistic GP-based MIL methods, we compare with the two most popular approaches VGPMIL [17] and VGPMIL-PR [18], which were reviewed in Section 2.2. For a fair comparison, the parameters used for the baselines are the same as those used for our method (described in next paragraph). For those parameters that do not have an analogous in our method (e.g. the initialization of $\mathrm{q}(\mathbf{y})$ in VGPMIL), we use the default values proposed in the original papers. To evaluate the performance of the compared methods we use four metrics: accuracy, precision, recall and F1-score (which provides a trade-off between precision and recall). For all the metrics, we use the standard implementations in the popular Python scikit-learn library [33].

Experimental details. For the underlying GPs, in this work we use the well-known squared exponential kernel [34], i.e. $\kappa(\mathbf{x}, \mathbf{y}) = \gamma \cdot \exp\left(-||\mathbf{x} - \mathbf{y}||^2/(2\ell^2)\right)$. Following [17] and [18], we use standard values for the kernel hyperparameters, i.e. $\gamma = 1$ and ℓ equals the square root of the number of features of $\mathbf{x}, \mathbf{y}$ (in this work we set $\ell = 11 \approx \sqrt{128}$). The number of inducing points is set to $M = 200$, and their locations are initialized through K-means clustering as in previous work [17, 18] (namely, 100 of them are obtained by doing clustering on the patches that belong to the positive images, and the other 100 on the patches that belong to the negative ones). The number of iterations is set to $T = 200$, which was enough to achieve convergence in practice. The expectation of the posterior distribution $\mathbb{E}_{\mathrm{q}(\mathbf{m})}(\mathbf{m})$ is initialized with a standard Gaussian for each instance independently. Notice that the initialization of $\mathrm{q}(\mathbf{u})$ is irrelevant since it gets updated first in Algorithm 1. As for the value of λ, which regulates the strength of the Ising correlation (recall eq. (8)), we will analyze five different values in the experiments, $\lambda \in \{0.1, 0.5, 1.0, 5.0, 10.0\}$. This will allow us to empirically illustrate the effect of λ.

	λ	Accuracy	Precision	Recall	F1-score
VGPMIL	-	83.87±0.00	83.33±0.00	100.00±0.00	90.91±0.00
VGPMIL-PR	-	90.32±0.00	89.29±0.00	100.00±0.00	94.34±0.00
VGPMIL-PR-I	**0.1**	**93.55±0.00**	**92.59±0.00**	**100.00±0.00**	**96.15±0.00**
	0.5	**93.55±0.00**	**92.59±0.00**	**100.00±0.00**	**96.15±0.00**
	1.0	90.32±0.00	92.31±0.00	96.00±0.00	94.12±0.00
	5.0	87.10±0.00	95.65±0.00	88.00±0.00	91.67±0.00
	10.0	83.87±0.00	95.45±0.00	84.00±0.00	89.36±0.00

Table 2: Predictive performance at the level of images (bags). In bold, we highlight the values of λ for which VGPMIL-PR-I gets better (or equal) performance than both baselines in all the metrics. The results are the mean and standard deviation over five independent runs.

		VGPMIL		VGPMIL-PR	
		Neg.	Pos.	Neg.	Pos.
Actual	Neg.	1	5	3	3
	Pos.	0	25	0	25

		VGPMIL-PR-I									
		$\lambda = 0.1$		$\lambda = 0.5$		$\lambda = 1$		$\lambda = 5$		$\lambda = 10$	
		Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.
Actual	Neg.	4	2	4	2	4	2	5	1	5	1
	Pos.	0	25	0	25	1	24	3	22	4	21

Table 3: Confusion matrices obtained at the level of images for the compared methods.

3.2 Experimental results

In this section we evaluate the performance of the compared methods on the aforementioned prostate cancer problem. We analyze eight different research questions, which are discussed in the following paragraphs.

Predictions at the level of instances (patches). Although they only use bag labels for training, the compared methods can make predictions at the level of instances, recall Section 2.5. This is important to determine more precisely in which region (patch) the cancer is present. Table 1 shows the results when making predictions at patch level. We observe that VGPMIL-PR-I outperforms both baselines in all the metrics for four out of five values of λ (and for $\lambda = 0.1$, the baselines are only better in terms of precision). We also appreciate that VGPMIL-PR-I results are robust across different runs, obtaining low values of standard deviation. This stability is important for real-world applications, where one wants to avoid high sensitivity to random initializations. Finally, notice that, as argued in [19], we also observe that VGPMIL-PR (slightly) outperforms VGPMIL.

Predictions at the level of bags (images). Table 2 shows the results when making predictions at the level of images. We observe that VGPMIL-PR-I outperforms both baselines in all the metrics when $\lambda \in \{0.1, 0.5\}$. However, when λ becomes larger, the results of VGPMIL-PR-I get worse. This fact can be explained theoretically because, whereas having low-to-moderate correlation among patches can be helpful, having strong ones tends to make the predictions too homogeneous, damaging the bag-level prediction (which takes into account the correlation among patches). Indeed, in the next research question we analyze with greater detail how λ is affecting the predictions on cancerous and non-cancerous images separately, which will provide additional insights. Finally, similar to the patch-level results, we observe that VGPMIL-PR obtains better results than VGPMIL, as expected.

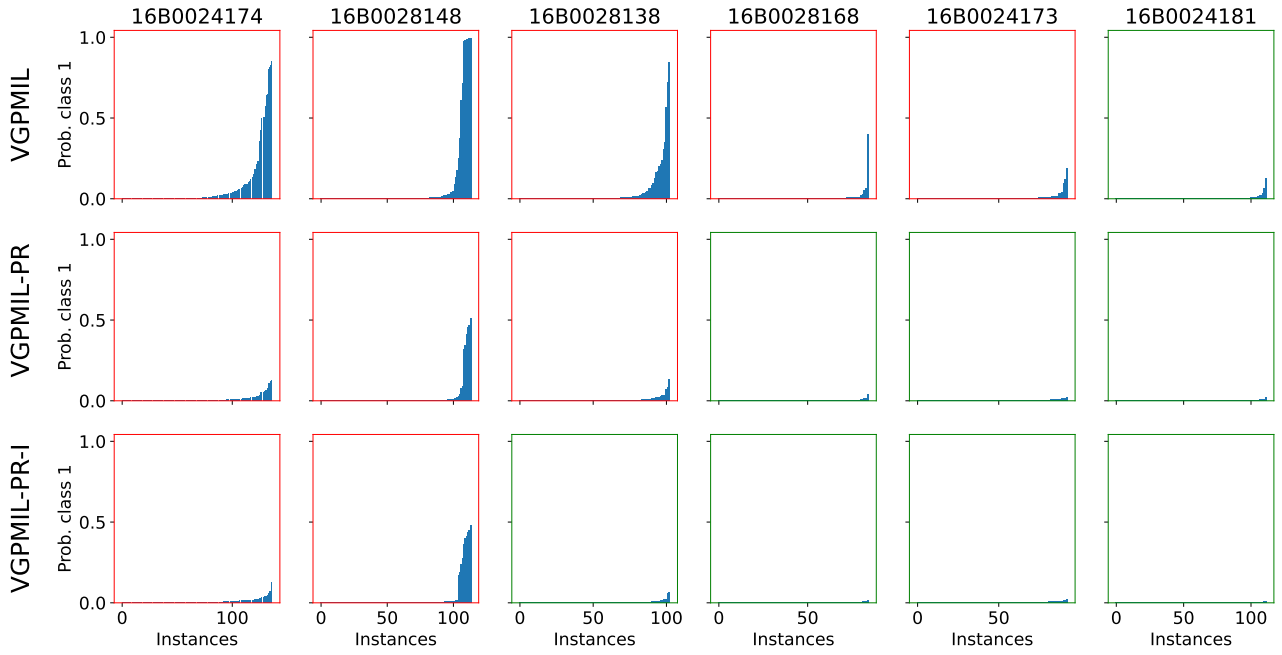


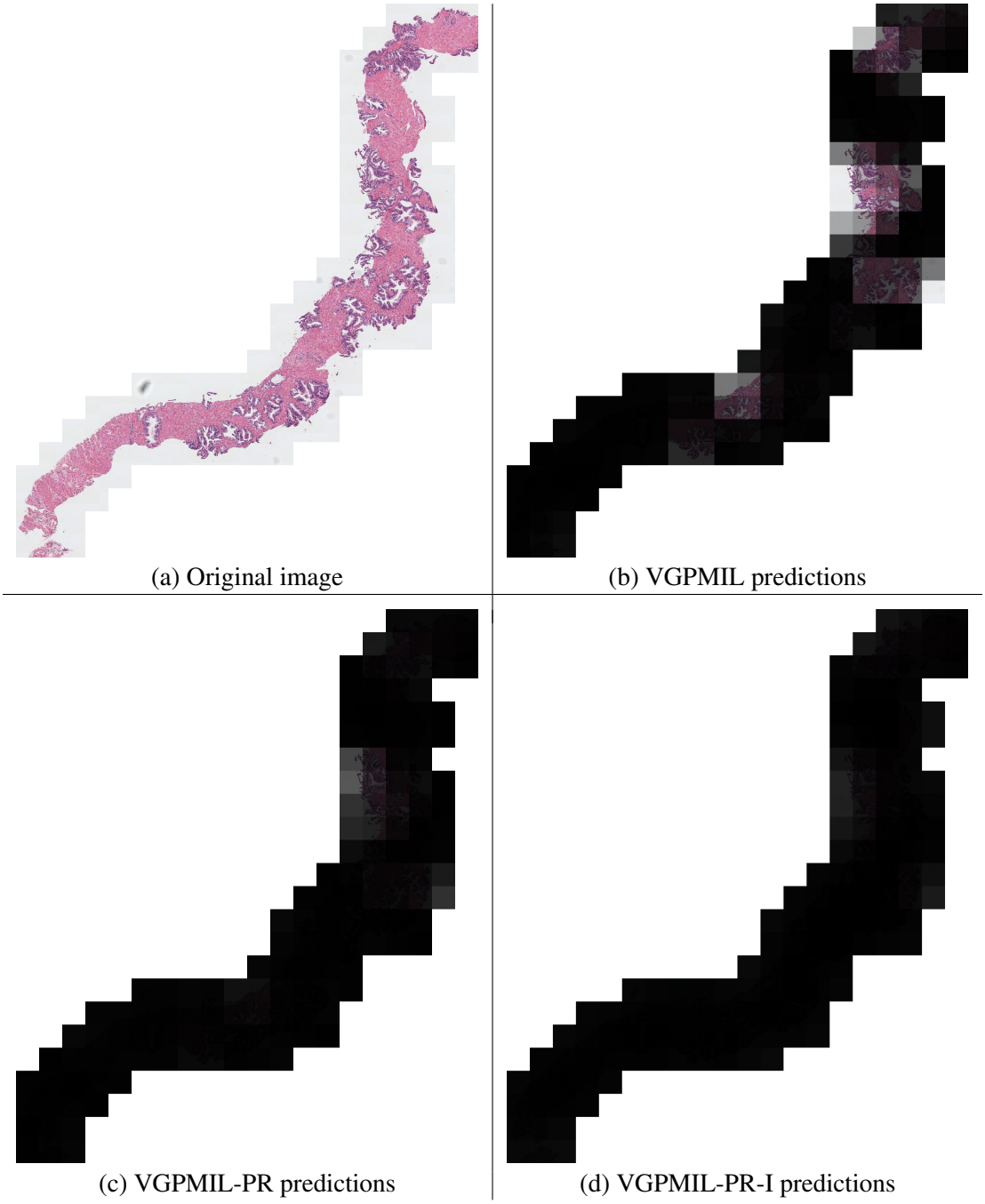
Figure 4: Patch-level predictions inside each one of the six negative (non-cancerous) WSIs in the test set. Each column is an image (the header is the image identifier in the SICAPv2 dataset). The rows refers to the three compared methods. Each subplot has red/green axis depending on whether the image is correctly classified or not by that method. The blue bars inside the subplots represent the probability of cancer for the different patches inside the image (for ease of visualization, they are sorted in increasing order).

	λ	Training time	Testing time
VGPMIL	-	15.45±0.53	3.14±0.16
VGPMIL-PR	-	11.78±0.75	2.63±0.12
VGPMIL-PR-I	0.1	12.21±0.26	2.23±0.09
	0.5	12.18±0.57	2.21±0.07
	1.0	11.79±0.44	2.24±0.15
	5.0	11.80±0.74	2.22±0.08
	10.0	11.64±0.91	2.26±0.09

Table 4: Computational cost for training and testing the compared methods (in seconds). We are using 200 iterations in all cases, recall the experimental details in Section 3.1. The results are the mean and standard deviation over five independent runs.

Also, the predictions at the level of images are very stable across runs (notice the zero standard deviation).

Analyzing the confusion matrices at bag level. Here we analyze more in detail the results presented in the previous paragraph (i.e. at the level of images). Table 3 shows the confusion matrices for the compared methods. Notice that both baselines classify all the 25 positive (cancerous) images correctly. However, the difficulties arise at the non-cancerous images. This happens because of the nature of the MIL problem: as soon as a few patches obtain a non-negligible probability of cancer, the image will be likely predicted as cancerous (recall that the MIL formulation establishes that a bag has positive class as soon as one instance inside the bag has positive class).



(a) Original image

(b) VGPMIL predictions

(c) VGPMIL-PR predictions

(d) VGPMIL-PR-I predictions

Figure 5: Patch level predictions obtained by the compared methods for image 16B0028138, which is non-cancerous. The original image is shown in (a). For predictions (b)–(d), the brightness of the patch is proportional to the probability of cancer (the brighter, the more probability).

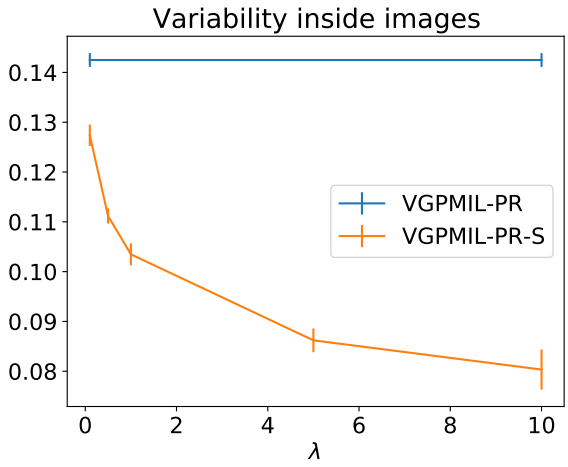


Figure 6: Variability in the patch-level predictions inside images. In VGPMIL-PR-I, the variability decreases as the strength of the Ising terms, given by λ, increases. VGPMIL-PR, which does not include Ising terms (i.e. $\lambda = 0$), gets larger variability. As explained in the text, the variability inside an image is measured as the standard deviation of the probability of cancer for all the patches inside that image.

Interestingly, the Ising model can help to avoid isolated false positive predictions on the patch-level (which lead to false positive bag predictions) by using the patch-level correlation. This is reflected in Table 3, where we observe that increasingly more negative images are predicted correctly as λ gets higher. In contrast, notice that strong correlation damage the performance in the positive class, as they penalize the appearance of positive patches (which would break the homogeneity of the bag, where most patches do not contain cancer). Therefore, we conclude that instance correlation is beneficial when used with a low-to-medium intensity. For instance, in this application $\lambda = 0.5$ is the best performing value, and it will be the one used by default in the sequel.

Analyzing the instance-level predictions for negative bags. In the previous paragraph, we have explained that negative images are incorrectly classified because a few patches inside them get classified as positive. Here we provide a visualization to support this. Figure 4 shows how the patch-level predictions are distributed inside the six negative images available in the test set. We observe that the amount of patches with a non-negligible probability of cancer gets reduced as we move from VGPMIL to VGPMIL-PR, and then to VGPMIL-PR-I. This translates into better performance at bag-level (observe that the amount of green-axis subplots increases in the same sequence VGPMIL $\rightarrow$ VGPMIL-PR $\rightarrow$ VGPMIL-PR-I). Notice that the improvement from VGPMIL to VGPMIL-PR is larger than from VGPMIL-PR to VGPMIL-PR-I. This may be due to the simplification that was introduced when computing the expectation of the truncated multivariate Gaussian in VGPMIL-PR-I, recall the third-to-last paragraph in Section 2.4.

Visualizing the predictions. In the last paragraph, we have analyzed how the patch-level predictions are distributed inside non-cancerous images *quantitatively*. Indeed, Figure 4 represents each patch through a bar. However, this hampers the *qualitative* visualization from a medical viewpoint. Here we focus on such qualitative assessment by visualizing the predictions obtained for image 16B0028138, see Figure 5. We have chosen this image because it illustrates best the effect of the coupling term. Notice that, thanks to these terms, the proposed VGPMIL-PR-I manages to keep all patches with a probability closer to zero than VGPMIL and VGPMIL-PR. As a consequence, VGPMIL-PR-I is the only method that correctly classifies this image as non-cancerous, recall Figure 4. For the other two methods, there are some patches that trigger the image prediction to be cancerous.

An explicit analysis on the role of λ. The hyperparameter λ is at the core of the novel VGPMIL-PR-I. It was introduced in the probabilistic model to regulate the strength of the coupling term,

recall eq. (8). This role has been confirmed *indirectly* in Table 3: when λ gets higher, the predictive performance improves for negative bags and degrades for positive ones. This can be explained because a higher λ homogenizes the patch-level predictions and difficulties the appearance of positive patches. Here we perform a more *direct* measure to gain insights into the role of λ. Specifically, we define the "variability inside a bag" as the standard deviation of the probability of cancer for all the patches inside that bag. Therefore, this metric measures the dispersion in the patch-level predictions obtained inside a bag. Figure 6 shows the evolution of this metric for VPGMIL-PR-I as λ increases. As theoretically expected, the metric decreases as λ gets higher. Also, notice that the metric value for VGPMIL-PR is higher. This is explained because VGPMIL-PR does not incorporate Ising correlation, i.e. $\lambda = 0$. The value for VPGMIL is even higher, 0.30, and it is not included in Figure 6 for ease of visualization. This greater value is probably due to the additional approximations that VGPMIL involves, which deepens the independence among patches.

Computational cost. Finally, we report the computational training and testing time for the compared methods, see Table 4. The results are in the same order of magnitude in all cases, which justifies the practical utility of the novel VGPMIL-PR-I, which obtained better predictive performance, recall Tables 1 and 2. In fact, VGPMIL-PR-I is slightly faster than VGPMIL, since the Jaakola bound approximation leveraged in the latter introduces additional parameters ξ to be estimated. As theoretically expected, the computational cost of VGPMIL-PR-I and VGPMIL-PR is analogous, since the update equations for the former are just a generalization of those for the latter, recall Sections 2.4 and 2.5. Finally, notice that the value of λ does not affect the computational cost of VGPMIL-PR-I, as λ only regulates the intensity of the Ising terms (but it does not introduce any additional computation).

Comparison to other related MIL approaches. So far we have focused on the comparison of VGPMIL-PR-I with VGPMIL and VGPMIL-PR. Since VGPMIL-PR-I builds on the same type of GP-based modeling, this is the most meaningful comparison in order to evaluate our main contribution (the Ising term to account for correlations among patches). However, to provide a wider perspective, it is interesting to compare the novel VGPMIL-PR-I to other state-of-the-art and popular families of MIL methods. We consider three families: attention-based methods, where the two algorithms proposed in [35] are the most popular approaches; MIL methods based on pseudo-labels such as the recent [7]; and classical pooling/aggregation methods such us the mean aggregation [36]. These will be referred to as Att-MIL, Gated-Att-MIL, PS-MIL and Mean-Agg, respectively.

Let us discuss the results both at instance (patch) and bag (image) levels. For the former, notice that the formulation of attention-based methods (Att-MIL and Gated-Att-MIL) and classical aggregation methods (Mean-Agg) do not allow for making predictions at instance level in a natural way. Namely, the instance-level labels are not modelled explicitly in this type of methods, and this is precisely one of their main limitations [37]. Compared to PS-MIL, which does model instance labels explicitly, the novel VGPMIL-PR-I achieves higher predictive performance (95.11 vs 85.01 in accuracy and 96.44 vs 88.05 in F1-Score). Regarding bag-level performance, the results are shown in Table 5. We observe that the best results are obtained by attention-based methods Att-MIL and Gated-Att-MIL, followed by the novel VGPMIL-PR-I.

In conclusion, we observe that the results are quite different depending on the nature of the model and the user requirements. If one is interested in predictions at patch level, then the novel VGPMIL-PR-I is the best choice. However, if one is only interested in image level performance, then attention-based approaches are the best option for this data. Indeed, we hypothesise that the performance of attention-based approaches could be even enhanced by leveraging correlation among patches in a similar way to VGPMIL-PR-I. This is a interesting line of future research, see Section 4.

3.3 Evaluation on a larger dataset: PANDA

The SICAPv2 dataset used so far is of medium-size (total amount of 126 WSIs, leading to 20378 patches; recall Section 3.1). This has allowed us to carry out a very detailed analysis of the results. In this section we show that the novel VGPMIL-PR-I also performs well on larger datasets, such

	Accuracy	Precision	Recall	F1-score
VGPMIL-PR-I	93.55±0.00	92.59±0.00	**100.00**±0.00	96.15±0.00
Att-MIL	**96.80**±0.00	**96.20**±0.00	**100.00**±0.00	**98.00**±0.00
Gated-Att-MIL	**96.80**± 0.00	**96.20**±0.00	**100.00**±0.00	**98.00**±0.00
Mean-Agg	87.10±0.00	95.70±0.00	88.00±0.00	91.70±0.00
PS-MIL	90.32±NA	89.28±NA	**100.00**±NA	94.33±NA

Table 5: Comparison with other related MIL methods which are not based on the GP modeling. The predictive performance at the level of images (bags) is shown. The results are the mean and standard deviation over five independent runs. The algorithm PS-MIL was run only once because of its high computational training cost.

	Accuracy	Precision	Recall	F1-score
VGPMIL	74.87±0.08	74.25±0.06	**99.77±0.00**	85.13±0.04
VGPMIL-PR	90.64±0.06	90.97±0.05	96.60±0.04	93.70±0.04
VGPMIL-PR-I	**92.57±0.14**	**95.42±0.24**	94.22±0.11	**94.82±0.09**
Att-MIL	90.92±0.01	94.17±0.01	93.43±0.01	93.79±0.01
Gated-Att-MIL	91.70±0.01	94.31±0.01	94.40±0.01	94.34±0.01
Mean-Agg	88.09±0.00	91.57±0.01	92.27±0.01	91.91±0.00
PS-MIL	88.36±NA	87.99±NA	97.11±NA	92.33±NA

Table 6: Predictive performance at the level of images (bags) in the PANDA dataset. The results are the mean and standard deviation over five independent runs. The algorithm PS-MIL was run only once because of its high computational training cost.

us the well-known PANDA set. Although scalability is not an issue from a theoretical perspective, since the model is based on sparse GPs, it is important to verify it in practice.

PANDA also tackles the problem of prostate cancer detection, and was presented at the MICCAI 2020 conference as a challenge[1]. Since the test set of PANDA is not publicly available, we use the train/test split proposed in [38], where each split follows the overall class proportions. Namely, the dataset used here features a total amount of 10503 WSIs, which leads to 1107931 patches. Notice that this is much larger than SICAPv2 (83 times larger in terms of WSIs).

Table 6 shows the predictive performance at image level, an aspect where attention-based methods stood out in the previous dataset. In this case, we observe that VGPMIL-PR-I obtains consistently better results. Additionally, as outlined in the last research question in Section 3.2, VGPMIL-PR-I is able to provide instance-level predictions, which is not the case for attention-based models. We conclude that, for the PANDA dataset, the proposed method is the best choice in comparison to the other tested approaches.

4 Conclusions, limitations and future work

In this work we have introduced VGPMIL-PR-I, a novel MIL methodology that incorporates instance label correlation through a coupling term inspired by the Ising model. VGPMIL-PR-I is a generalization of another probabilistic MIL method, whose formulation is theoretically recovered when the influence of the Ising term converges to zero. In the experimental section, we have shown that VGPMIL-PR-I outperforms other related state-of-the-art probabilistic MIL approaches in two real-world problems of prostate cancer detection, effectively reducing false positive bag predictions and providing instance-level predictions. We have also provided different visualizations to better understand the behavior of the proposed model, specially the influence of the new coupling term.

[1] `https://panda.grand-challenge.org/`

As discussed along the paper, our model presents several limitations which we summarize next. Firstly, we needed to introduce a diagonal approximation to compute the expectation of the truncated mulivariate Gaussian in VGPMIL-PR-I, recall Section 2.4. This is probably reflected in the empirical performance, as the improvement when moving from VGPMIL to VGPMIL-PR is generally larger than when moving from VGPMIL-PR to VGPMIL-PR-I. Secondly, we have observed that the behaviour of VGPMIL-PR-I depends on the value of λ, which regulates the strength of the coupling term. Although we have discussed the role of λ and tested different values, it remains a hyperparameter that has to be found empirically using the validation set. We believe that its value (or even distribution over it) could be estimated from the data by introducing λ in the probabilistic modeling. Even more, λ could be estimated per image, since the level of correlation could be image-dependent. Thirdly, we have observed that the image-level performance of VGPMIL-PR-I is not generally better than that of attention-based methods. This is probably due to the different nature of the models. Indeed, the explicit modeling of instance label in GP-based models, which allows them to provide instance-level predictions, may come at the cost of less accurate bag-level predictions.

In addition to the aforementioned ideas, this work opens other future research lines. First, seeing the performance boost obtained in GP-based methods through the novel coupling term, and taking into account the good results of attention-based methods in bag-level prediction, it is very interesting to explore the modeling of instance label correlations in the context of attention-based methods. Second, notice that we are using mean-field variational inference to estimate the model parameters in VGPMIL-PR-I. A promising alternative is to estimate them by directly optimizing the evidence lower bound (ELBO). Finally, although we have focused on modeling correlation between neighboring patches in histopathogical images, we expect that the ideas behind our proposal can boost further research in MIL, by exploiting the particular structure of the data used in different applications.

References

[1] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329–353, 2018.

[2] M. Cui and D. Y. Zhang, "Artificial intelligence and computational pathology," *Laboratory Investigation*, vol. 101, no. 4, pp. 412–422, 2021.

[3] S. Huang, Z. Liu, W. Jin, and Y. Mu, "Bag dissimilarity regularized multi-instance learning," *Pattern Recognition*, vol. 126, p. 108583, 2022.

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

[5] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[6] M. López-Pérez, M. Amgad, P. Morales-Álvarez, P. Ruiz, L. A. Cooper, R. Molina, and A. K. Katsaggelos, "Learning from crowds in digital pathology using scalable variational gaussian processes," *Scientific reports*, vol. 11, no. 1, pp. 1–9, 2021.

[7] A. Schmidt, J. Silva-Rodríguez, R. Molina, and V. Naranjo, "Efficient cancer classification by coupling semi supervised and multiple instance learning," *IEEE Access*, vol. 10, pp. 9763–9773, 2022.

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

[9] A. Schmidt, J. Silva-Rodríguez, R. Molina, and V. Naranjo, "Coupling semi-supervised and multiple instance learning for histopathological image classification," *IEEE Access*, 2022.

[10] Y. Wu, A. Schmidt, E. Hernández-Sánchez, R. Molina, and A. K. Katsaggelos, "Combining attention-based multiple instance learning and gaussian processes for ct hemorrhage detection," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 582–591.

[11] M. Kandemir, F. A. Hamprecht *et al.*, "Instance label prediction by dirichlet process multiple instance learning." in *UAI*, 2014, pp. 380–389.

[12] J. Read, L. Martino, and J. Hollmén, "Multi-label methods for prediction with sequential data," *Pattern Recognition*, vol. 63, pp. 45–55, 2017.

[13] J. Read, L. Martino, and D. Luengo, "Efficient monte carlo methods for multi-dimensional learning with classifier chains," *Pattern Recognition*, vol. 47, no. 3, pp. 1535–1546, 2014.

[14] J. Read and L. Martino, "Probabilistic regressor chains with monte carlo methods," *Neurocomputing*, vol. 413, pp. 471–486, 2019.

[15] M. Kim and F. De la Torre, "Gaussian processes multiple instance learning," in *ICML*, 2010.

[16] M. Kandemir, M. Haußmann, F. Diego, K. T. Rajamani, J. Van Der Laak, and F. A. Hamprecht, "Variational weakly supervised gaussian processes," in *Proceedings of the British Machine Vision Conference (BMVC)*, E. R. H. Richard C. Wilson and W. A. P. Smith, Eds. BMVA Press, September 2016, pp. 71.1–71.12.

[17] M. Haußmann, F. A. Hamprecht, and M. Kandemir, "Variational bayesian multiple instance learning with gaussian processes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6570–6579.

[18] F. Wang and A. Pinar, "The multiple instance learning gaussian process probit model," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 3034–3042.

[19] Z. Wang, A. Lamb, E. Saveliev, P. Cameron, J. Zaykov, J. M. Hernandez-Lobato, R. E. Turner, R. G. Baraniuk, C. Barton, S. P. Jones *et al.*, "Results and insights from diagnostic questions: The neurips 2020 education challenge," in *NeurIPS 2020 Competition and Demonstration Track*. PMLR, 2021, pp. 191–205.

[20] N. Ding, J. Deng, K. P. Murphy, and H. Neven, "Probabilistic label relation graphs with ising models," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1161–1169.

[21] P. Qin and J. Zhao, "A polynomial-time algorithm for image segmentation using ising models," in *2011 Seventh International Conference on Natural Computation*, vol. 2, 2011, pp. 932–935.

[22] F. Wang and A. Pinar, "The multiple instance learning gaussian process probit model," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Banerjee and K. Fukumizu, Eds., vol. 130. PMLR, 13–15 Apr 2021, pp. 3034–3042.

[23] E. Snelson and Z. Ghahramani, "Sparse gaussian processes using pseudo-inputs," in *Advances in Neural Information Processing Systems*, vol. 18, 2006.

[24] M. Girolami and S. Rogers, "Variational bayesian multinomial probit regression with gaussian process priors," *Neural Computation*, vol. 18, no. 8, pp. 1790–1817, 2006.

[25] B. D. Ripley, *Spatial statistics*. John Wiley & Sons, 2005.

[26] J. Hensman, A. De G. Matthews, and Z. Ghahramani, "Scalable variational Gaussian process classification," in *International conference on artificial intelligence and statistics*, 2015, pp. 351–360.

[27] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[28] S. Wilhelm and B. Manjunath, "tmvtnorm: A package for the truncated multivariate normal distribution," *R Journal*, vol. 2, no. 2, pp. 1–25, 2010.

[29] Y. Li and S. K. Ghosh, "Efficient sampling methods for truncated multivariate normal and student-t distributions subject to linear inequality constraints," *Journal of Statistical Theory and Practice*, vol. 9, no. 4, pp. 712–732, 2015.

[30] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous univariate distributions, volume 2*. John wiley & sons, 1995, vol. 289.

[31] "Global cancer observatory, world health organisation," https://gco.iarc.fr/, accessed: 2010-03-15.

[32] J. Silva-Rodríguez, A. Colomer, M. A. Sales, R. Molina, and V. Naranjo, "Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection," *Computer Methods and Programs in Biomedicine*, vol. 195, p. 105637, 2020.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[34] C. Williams and C. Rasmussen, *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006, vol. 2.

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

[36] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2017, pp. 603–611.

[37] A. Schmidt, P. Morales-Álvarez, and R. Molina, "Probabilistic attention based on gaussian processes for deep multiple instance learning," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2023.

[38] J. Silva-Rodriguez, A. Colomer, J. Dolz, and V. Naranjo, "Self-learning for weakly supervised gleason grading of local patterns," *IEEE journal of biomedical and health informatics*, vol. 25, no. 8, pp. 3094–3104, 2021.

Chapter 4

Combining Semi Supervised and Multiple Instance Learning

4.1 Publication details

Authors: Arne Schmidt, Julio Silva-Rodríguez, Rafael Molina, Valery Naranjo
Title: Efficient Cancer Classification by Coupling Semi Supervised and Multiple Instance Learning
Reference: IEEE Access, vol. 10, pp. 9763-9773, 2022, doi: 10.1109/ACCESS.2022.3143345.
Status: Published
Quality indices:

- Impact Factor (JCR 2022): 3.9
  - Rank 72/158 (Q2) in Computer Science, Engineering Systems
  - Rank 100/275 (Q2) in Engineering, Electrical and Electronic

4.2 Main contributions

- We propose a novel method that allows to train with bag labels, an arbitrary amount of instance labels and unlabeled data. The model uses pseudo labels derived by the bag labels and predicted probabilities to leverage the information of the unlabeled instances.
- A new labeling paradigm is introduced. We show that the model is more efficient if a pathologist provides some patch labels for each available WSI instead of labeling complete WSIs extensively with the same labeling resources.
- For three public datasets of breast and prostate cancer, the model shows highly accurate local predictions and only 5 patch labels per WSI were enough to obtain a performance similar to a fully supervised model with all patch labels.
- The feature extractor with the proposed training procedure was used to extract features for several probabilistic models.

EFFICIENT CANCER CLASSIFICATION BY COUPLING SEMI SUPERVISED AND MULTIPLE INSTANCE LEARNING

Arne Schmidt
Dep. of Computer Science and AI
University of Granada
Granada, Spain

Julio Silva-Rodríguez
Inst. of Transport and Territory
Universitat Politècnica de València
Valencia, Spain

Rafael Molina
Dep. of Computer Science and AI
University of Granada
Granada, Spain

Valery Naranjo
Inst. of Research and Innovation in Bioengineering
Universitat Politècnica de València
Valencia, Spain

ABSTRACT

The annotation of large datasets is often the bottleneck in the successful application of artificial intelligence in computational pathology. For this reason recently Multiple Instance Learning (MIL) and Semi Supervised Learning (SSL) approaches are gaining popularity because they require fewer annotations. In this work we couple SSL and MIL to train a deep learning classifier that combines the advantages of both methods and overcomes their limitations. Our method is able to learn from the global WSI diagnosis and a combination of labeled and unlabeled patches. Furthermore, we propose and evaluate an efficient labeling paradigm that guarantees a strong classification performance when combined with our learning framework. We compare our method to SSL and MIL baselines, the state-of-the-art and completely supervised training. With only a small percentage of patch labels our proposed model achieves a competitive performance on SICAPv2 (Cohen's kappa of 0.801 with 450 patch labels), PANDA (Cohen's kappa of 0.794 with 22,023 patch labels) and Camelyon16 (ROC AUC of 0.913 with 433 patch labels). Our code is publicly available at `https://github.com/arneschmidt/ssl_and_mil_cancer_classification`.

Keywords Cancer Classification · Histopathology · Multiple Instance Learning · Semi-Supervised Learning · Whole Slide Images

1 Introduction

The analysis of histopathological biopsies is the gold standard for the diagnosis of many different cancer types. In the last years, Computer-Aided Diagnosis (CAD) systems based on artificial intelligence have gained attention as a promising tool to reduce pathologists' workload, improve the repeatability and to avoid the variability of diagnostic processes. For the training of deep learning algorithms, initially many approaches relied on detailed local-level annotations of the digitized biopsies by pathologists [1]. Unfortunately, due to the large size of the WSIs, this process is a time-consuming task which makes it difficult to obtain large and heterogeneous annotated datasets. This recently led to the rise of approaches that do not need detailed local-level annotations. Instead,

they utilize the MIL assumption where the image patches form the instances and the complete WSI forms the bag [2]. In this setting, no patch-level annotations are needed and only the diagnosis of the biopsies are used for training. Another strategy to learn with fewer patch-level annotations is SSL where only a subset of the image patches must be labeled. Still, existing methods have some common limitations: while SSL techniques do not incorporate the WSI diagnosis (global label) and therefore show a limited performance, MIL methods often can not make accurate patch-level predictions or have to be trained on very large datasets. For example, in [2] the authors conclude that at least 10,000 slides are necessary for a good performance. These limitations encourage the development of novel data-efficient methodologies which balance the amount of patch-level annotations and size of the required datasets and can flexibly adapt to different scenarios.

1.1 Contributions

We propose a new machine learning method based on MIL and SSL and an efficient labeling strategy to perform cancer classification with fewer annotations and reduced human workload. The contributions of this work are:

- A novel cancer classification method utilizing the global WSI diagnosis, unlabeled image patches and a limited number of labeled image patches for training. The proposed method exploits pseudolabeling techniques to combine both global labels in the MIL perspective and scarce patch-level annotations under the SSL setting. This combined approach overcomes current limitations of existing MIL and SSL methods and shows a significant improvement in comparison to the SSL and MIL baselines.
- An Efficient Labeling (EL) technique to achieve the best possible performance with a limited amount of annotations. Instead of annotating complete WSIs we propose to annotate only some cancerous patches per WSI for each cancer class.

We make an extensive quantitative validation of the performance on three different datasets and show that our deep learning framework achieves very competitive results without the need for detailed patch labels or an excessive amount of WSIs. With just a few patch labels per WSI we get a similar performance as in a supervised setting, even on relatively small datasets. The success of our algorithm supports the following labeling paradigm: A good performance of deep learning algorithms is already possible if pathologists only point out a few cancerous image patches per WSI instead of spending a lot of time with the detailed annotation.

1.2 Related Work

To structure the related work into Multiple Instance Learning (MIL) and Semi Supervised Learning (SSL) approaches, we first clarify the definition of both, following the terminology of Cheplygina et al. [3]. Under the MIL assumption, instances (patches) are grouped into bags (WSIs), where only the label of the entire bag is known and the instance labels remain unobserved. In this paradigm, learning is driven by known global information (WSI diagnosis). SSL describes a learning scenario with two sets of samples: a labeled set and an unlabeled set. SSL methods use the unlabeled set (additionally to the labeled set) to find a better decision boundary and improve the classifier. In the given use-case, this means SSL methods use labeled and unlabeled patches for training, but not the WSI labels.

MIL approaches for histopathological images are becoming more and more popular because they do not require detailed local annotations, but only bag labels for training. [2] Usually, a bag-level representation is obtained by the aggregation of either the instance-level features (embedding-based) or their predictions (instance-based). Recently, the classical aggregation functions based on max or average pooling have been replaced by more advanced mechanisms, such as learnable attention methods [4]. Campanella et al. [2] showed promising results processing the top-ranked positive instance features with an RNN. In other works, the use of instance-based aggregations based on top and bottom ranked instances [5] or min-max aggregation [6] have been proposed.

Further approaches use embedding-based MIL via multi-head attention mechanisms [7] or combine instance-level predictions with embeddings [8]. Hashimoto et al. [9] use multiple scales with attention mechanisms and domain adversarial training for malignant lymphoma subtype classification. Common limitations of existing approaches are the requirement of very large datasets [2] and the incapability to make class predictions at instance level [4] [2] [9]. Further, recent approaches often include complex multi-stage training procedures with multiple models [2] [9]. This motivates the development of well performing, but simpler approaches for an easy application in clinical practice.

SSL approaches use labeled and unlabeled patches for training. For histopathological images, most existing SSL approaches rely on pseudo-labeling techniques such as Pulido et al. [10] who apply MixMatch [11] and FixMatch [12] under a highly noisy and imbalanced data setting. Jaiswal et al. [13] combine pseudo-labeling techniques with a novel learning rate schedule (one cycle policy). The approaches of Shaw et al. [14] and Marini et al. [15] are based on teacher-student models, where the teacher model trains with the labeled set of images. The SSL component of our work is related to FixMatch and Unsupervised Data augmentation (UDA) [16]: UDA proposes to use unlabeled images for so-called consistency regularization. Fixmatch extends the idea of consistency regularization with pseudolabels: Based on weak image augmentations, pseudo labels are assigned to confident predictions while the network is trained with strong image augmentations. The common drawback of all the mentioned SSL methods is that they do not make use of global information (bag labels) and always require a certain amount of labeled instances.

SSL+MIL approaches were proposed very recently for histopathological images, but existing methods show some major differences to our work. Otalora et al. [17] propose an SSL+MIL method based on teacher-student networks, but it is specialized for prostate cancer and uses micro tissue arrays for pre-training. Although this approach is theoretically interesting, the performance gap to the supervised state-of-the-art models is quite large in practice (listed in Table 2). Li et al. [18] and Lu et al. [19] also propose hybrid models of SSL+MIL, but the applications are not comparable to our work: while the first approach is applied to binary semantic segmentation of WSIs, the latter is used for binary classification of histopathological images of 2048×1536 pixels that are much smaller than WSIs.

Our method takes advantage of both SSL and MIL learning strategies and is able to perform multi-class classification on WSIs for different cancer types. It incorporates the augmentation strategy of FixMatch [12] and the consistency regularization of Unsupervised Data augmentation (UDA) [16] while the pseudo label assignment is driven by the MIL perspective. As a result, the proposed method inherits the advantages of SSL and MIL while overcoming their existing limitations: our method achieves competitive results on small datasets, provides multi-class instance-level predictions, only needs one training procedure, one stage and one model that performs the common mini-batch training but still has the capability to include the bag label information.

1.3 Paper Structure

The rest of the paper is organized as follows: In section 2 we describe the problem in theoretical terms, the proposed efficient labeling strategy (2.1), the image augmentation strategy (2.2), the training framework (2.3) and the theoretical background of the proposed method (2.4). In the experiment section 3 we first outline the description of dataset (3.1) and implementation (3.3). In the ablation studies (3.4) we show experimentally the effect of the different loss components. Finally, we highlight the effect of the proposed efficient labeling strategy (3.5) and compare with state-of-the-art methods (3.7) before concluding our article (4).

2 Model description

Let us consider a WSI classification problem where images are assigned a single class Y or a primary and secondary class Y^1 and Y^2. We refer to these WSI labels as 'bag labels' in the context of MIL and to the image patches as 'instances'. Each patch can be either non-cancerous (NC) or contain one of the cancer classes. There are many problems that can be formulated this way. For

the example of prostate cancer, the tissue is classified as non-cancerous (NC), Gleason grade 3 (GG3), Gleason grade 4 (GG4) or Gleason grade 5 (GG5). The primary Gleason grade Y^1 and the secondary Gleason grade Y^2 of a WSI are assigned based on the two most prominent grades. In other cancer classification tasks like the lymph node detection of the Camelyon16 challenge, just one global label is assigned. Our approach works in both cases.

To translate the problem into a mathematical notation, we denote the bag indices as $B = \{1, 2, .., N\}$ where N is the number of WSIs in the training set. Let further $I_b = \{1, 2, .., M_b\}$ be the index set for the image patches (instances) in bag b. The complete set of image patches and their true cancer class can now be defined as

$$\{x_{bi}, y_{bi}\} \;\; b \in B, i \in I_b \tag{1}$$

To describe the labels let us first define the subset of non-cancerous WSIs $B^- \subset B$ and cancerous WSIs B^+.
Following the MIL assumption we know that for each negative bag $b \in B^-$:

$$y_{bi} = NC \; \forall i \in I_b \tag{2}$$

For all positive bags we know that some patches must contain the pattern of the present cancer class Y_b. For each bag $b \in B^+$:

$$\exists i \in I^b : y_{bi} = Y_b \tag{3}$$

which in the case of a primary and secondary label applies to both Y_b^1 and Y_b^2.

Note that the targets y are represented as a C-dimensional probability vector with each dimension representing one class probability and the class labels are described as one-hot vectors.

2.1 Efficient Labeling

We propose a data setting that we name Efficient Labeling (EL): For each cancerous WSIs the pathologist only points out a few cancerous patches instead of annotating the whole WSI. For each global label Y_b some corresponding patch labels $y_{bi} = Y_b$ are assigned. We consider the annotation of a few cancerous patches per WSI a realistic and time-efficient strategy for the annotation of a new dataset from scratch or the data collection in already deployed CAD systems. In the latter case, the pathologist provides labels during the diagnostic process (human-in-the-loop, see f.e. [20]).

In our experiments, this data setting is simulated by picking randomly a certain amount of patch labels and hide the others during model training. This allows us to systematically study the effect of a varying amount of patch labels.
We divide the indices of each positive bag into the set of labeled ($L \subset I_b$) and the set of unlabeled ($U \subset I_b$) instances such that all labels $\{y_{bi} | i \in L_b\}$ are available due to pathologists annotation, while the labels $\{y_{bi}, | i \in U_b\}$ remain unknown.

2.2 Image Augmentation

Our image augmentation strategy is related to FixMatch [12] and Unsupervised Data augmentation (UDA) [16]: UDA proposes to use unlabeled images for so-called consistency regularization: for two versions of a randomly augmented image the network is trained to predict the same class probabilities. The FixMatch algorithm combines consistency regularization with pseudolabeling. Here, a weak image augmentation is applied to the unlabeled images, the class is estimated by a CNN and pseudo labels are assigned to the images with confident class predictions. Then the network is trained to predict these pseudo labels given a strongly augmented version of the unlabeled images. Both approaches have in common that random image augmentation is a key component.

Similar to [12] the weak and strong image augmentation for the image patches play an important role in our approach. The strong image augmentation in our implementation uses a very strong

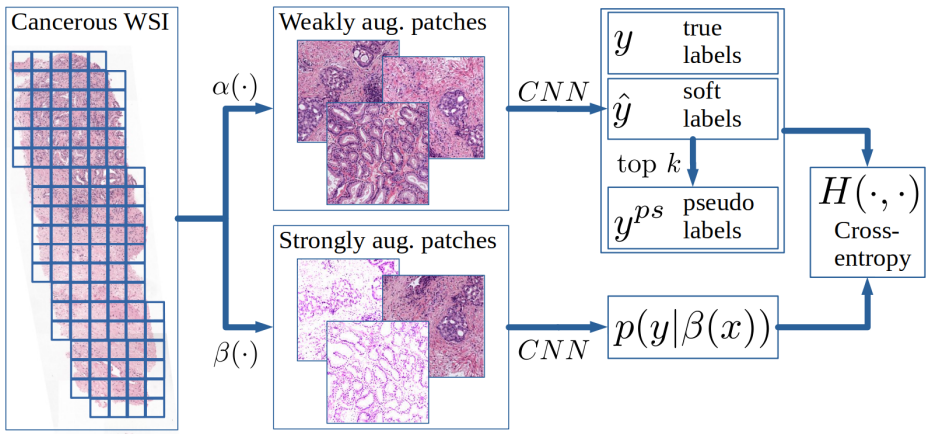


Figure 1: Proposed training framework for cancerous WSI, combining MIL and SSL. We take all patches of the WSI and apply a weak augmentation to obtain soft labels and pseudo labels by the CNN predictions. Based on these labels, we train the same CNN with the strongly augmented patches.

random brightness shift that leads to substantially darker and brighter versions of the original image. The weak augmentation only applies a mild version of the brightness shift, leading to images similar to the original. Applying only a weak augmentation makes it easier for the network to obtain a correct prediction and is therefore used to estimate pseudo labels and soft labels. The strongly augmented images are more challenging to predict and are therefore used to train the network. We denote $\alpha(\cdot)$ as the operator of weak random image augmentation and $\beta(\cdot)$ as a strong random image augmentation. For more details, we refer to the theoretical background (2.4) and the implementation details (3.3).

2.3 Proposed Training Framework

The goal is to train a patch classifier $p_\theta(y|x)$ which predicts class probabilities y for a given patch x and is parametrized by the model weights θ (following the notation of [12]). The training procedure (Figure 1) can be applied to any classification model and is divided into three steps that are repeated for each training epoch:

Step 1 Obtain the CNN predictions of the weakly augmented image patches in the positive bags. For a given image patch x_{bi} of the positive bag $b \in B^+$, we apply the weak image augmentation α. The weakly augmented image patch $\alpha(x_{bi})$ is used to predict the CNN output probability vector $p_\theta(y|\alpha(x_{bi}))$ which we define as $\hat{y}_{bi}$:

$$\hat{y}_{bi} := p_\theta(y|\alpha(x_{bi})) \; \forall b \in B^+ \tag{4}$$

As some of these vectors of probabilities will serve later as training targets, we will call $\hat{y}$ soft labels.

Step 2 Calculate pseudo labels for each positive bag $b \in B^+$. We know from equation (3) that some patches have the same class as the WSI. Given the global label Y_b of the bag, we assign this pseudo label to the k patches whose class probabilities of class Y_b are the largest of all instances in the bag. Concretely, this is done by the following steps:

(i) Create a list of probability vectors $\hat{y}_{bi}$ ordered with respect to class Y_b.

(ii) Select the k first items of this list to define the index set $P_b \subset I_b$.

(iii) Assign the one-hot class label Y_b to the patches indexed by P_b as a pseudo label y^{ps}:

$$y^{ps}_{bi} = Y_b \; \forall i \in P_b, b \in B^+ \tag{5}$$

In the case of two or more global labels, this pseudo label assignment is performed for each of them.

Step 3 Use the strongly augmented image patches $\beta(x)$ and a combination of groundtruth labels, pseudo labels and soft labels to train the CNN. Mathematically, the loss function is described as:

$$\begin{aligned}\mathcal{L}(\theta) = &\sum_{b\in B^-} \underbrace{\sum_{i\in I_b} H(y_{bi}, p_\theta(y|\beta(x_{bi}))}_{A} \\ &+ \sum_{b\in B^+} \Bigg(\underbrace{\sum_{i\in L_b} \lambda H(y_{bi}, p_\theta(y|\beta(x_{bi}))}_{B} \\ &\quad + \underbrace{\sum_{i\in P_b} H(y^{ps}_{bi}, p_\theta(y|\beta(x_{bi}))}_{C} \\ &\quad + \underbrace{\sum_{i\in U_b\setminus P_b} H(\hat{y}_{bi}, p_\theta(y|\beta(x_{bi}))}_{D} \Bigg)\end{aligned} \tag{6}$$

Here, $H(\cdot,\cdot)$ denotes the cross-entropy loss for classification and λ is a hyperparameter to assign a higher weight to the groundtruth labels of the cancer classes. Note that all terms of the loss function (A, B, C, D) split into sums over the instances. Training can therefore be performed in minibatches via stochastic gradient descent. In comparison to semi-supervised methods, our algorithm is still able to train without any patch labels (MIL setting): In this case, the loss term of positive instance labels (B) can be simply omitted, and the training can be performed based only on negative, pseudo and soft labels (terms A, C and D).

The proposed training framework is summarized in Algorithm 1. For notational coherence, we describe the algorithm for an instance-wise optimization. In practice, the prediction of step 1 and the gradient update of step 3 can be performed in common mini-batches for efficient computational parallelization.

2.4 Background

In the following subsection, we want to explain the derivation and theoretical background of the different loss components and the image augmentation.

The MIL component of our method enables the model to incorporate information from the global WSI labels during training and constitutes the loss terms A and C of equation 6. The loss term A uses the MIL property of equation 2: all instances in a negative bag must be negative. With these negative instances, the model can perform supervised training. Further, the pseudo labels of term C are derived by the MIL perspective: From equation (3) we know that some instance labels are equal to the bag label Y_k. A natural assumption is that instances with the highest class probabilities (of class Y_k) are the best candidates for the assignment of label Y_k. When no positive instance labels are available, this label assignment enables the model to learn the positive classes at instance level through the bag labels. The proposed MIL component can be seen as an extension of the max-pooling which is used for example in [2] in the first training phase. Instead of assigning the global label just to one instance with the highest probability, we assign it to multiple instances (k in total) with the highest probabilities. Further, we extend the binary case [2] to multiple classes using the class probabilities of the given global label, as described in step 2. The empirical improvement of our algorithm over max-pooling is discussed in section 3.4.

Algorithm 1 Proposed Training Procedure

Input: For each bag $b = 1, .., N$: Image patches $\{x_{bi}\}_{i=1,..,M_b}$, a reduced number of patch labels $\{y_{bi}\}_{i \in L_b}$, WSI labels $\{Y_b\}$, number of epochs E, learning rate η
Output: Optimal model parameters θ

```
for e = 1 to E do
  # Step 1
  for b = 1 to N do
    for i = 1 to M_b do
      estimate ŷ_bi ← p_θ(y|x_bi) (eq. 4)
    end for
    # Step 2
    Order {ŷ_bi} regarding class Y_b (Step 2 (i))
    Define P_b as the k max. probabilities (Step 2 (ii))
    Assign y_bi^ps ← Y_b for i ∈ P_b (Step 2 (iii))
  end for
  # Step 3
  for b = 1 to N do
    for i = 1 to M_b do
      θ ← θ − η dL(θ)/dθ (eq. 6, using {y_bi}, {y_bi^ps}, {ŷ_bi})
    end for
  end for
end for
return θ
```

The SSL component of our method ensures that the labeled (term A and B of equation 6) and unlabeled (term C and D of equation 6) patches are used to improve the classifier. From a theoretical point of view, it has been shown that pseudo labels (loss term C) can be interpreted as a form of entropy minimization [21].

As the conditional entropy of class probabilities is a measure of class overlap, the optimization will favor putting the class decision boundary in a low density area and leads to a better separation of classes [22]. The loss term D with soft labels serves as an additional consistency regularization: for two randomly augmented versions of the image, the classifier is trained to predict the same output. This technique has been proven to lead to better generalization and stability of the classifier [23]. Further, we want to discuss the role of weak and strong image augmentations for label propagation. The basic assumption of semi-supervised learning algorithms is that the data distribution of unlabeled datapoints can help a model to find a better decision boundary between the classes. One strategy is to propagate label information from one datapoint to nearby unlabeled datapoints during the model training (so called 'label propagation', see f.e. [24] or [16]). The final goal is to assign a consistent label in high density areas provided by some labeled datapoints. Label propagation with loss terms C and D in combination with weak and strong image augmentation (α and β) can be explained in the following way: Let $V_\alpha(x)$ and $V_\beta(x)$ be the space of all possible image augmentations with α and β, respectively, for a given image patch x. As the strong image augmentation β leads to a higher distortion of the image, the image space $V_\beta(x)$ is larger than $V_\alpha(x)$ and we assume $V_\alpha(x) \subset V_\beta(x)$ when the same random augmentations are applied for α and β. As the pseudo and soft labels are predicted on $\alpha(x)$ and the network is trained on $\beta(x)$, label information is propagated from $V_\alpha(x)$ to $V_\beta(x)$ during training, as shown in Figure 2. Other unlabeled datapoints that are in or close to $V_\beta(x)$ are more likely to be assigned the same class as x in the next iteration. Therefore, the available patch labels are propagated to unlabeled patches in areas of high data density. As a result, the model is encouraged to assign a similar label to all instances in a data cluster and to define the decision boundaries between those data clusters.
The SSL component of our work is inspired by Fixmatch [12] and UDA [16] and in the following, we briefly discuss similarities and differences. Fixmatch has a similar augmentation strategy as

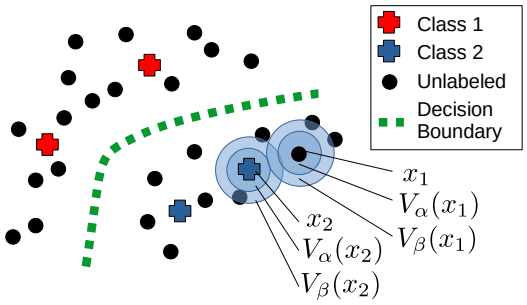


Figure 2: Simplified illustration of label propagation with weak and strong image augmentation. The datapoints correspond to image patches in our case. Shown are two example points x_1 and x_2 with the corresponding regions of weak and strong image augmentation ($V_\alpha(x)$ and $V_\beta(x)$ respectively).

the proposed method, but we extend it with soft labels and a MIL-driven pseudo label assignment instead of using a probability threshold as in the original work. This enables the model to incorporate bag labels during training while maintaining the benefits of SSL. The soft label assignment is inspired by UDA, such that loss terms B and D are similar to the UDA training. The idea of label propagation by consistency regularization was presented in the context of UDA, together with a theoretical proof based on graph theory. Apart from the soft-label assignment, UDA is lacking the other components of our proposed method (weak/strong augmentation, pseudo label assignment, bag label incorporation).

3 Experiments and Discussion

We performed extensive experiments to evaluate our proposed training framework as well as the proposed efficient labeling (EL) strategy.

3.1 Datasets

The experiments were conducted using three different public WSIs datasets of prostate and breast cancer. The gigapixel WSIs are sliced into smaller patches that form the instances of the MIL problem, while the WSI diagnosis is the bag label. The used datasets have both biopsy-level and patch-level annotations available, which made them particularly suitable for the validation of the proposed method. The SICAPv2[1] [25] dataset was used to validate the proposed method on prostate cancer for the multiclass Gleason grading scenario. This dataset contains 155 prostate WSIs which are sliced into 512x512 overlapping patches. The primary and secondary Gleason grade for all WSIs as well as patch-level labels are included for a large number of instances in the dataset. In our work, we maintained the proposed partitions of the original dataset for training, validation and testing.
Additionally, we use the PANDA dataset [2] for prostate cancer classification, which is substantially larger than SICAPv2, to test our method on a dataset with a different size. It consists of 10,415 WSIs and was presented at the MICCAI 2020 conference as a challenge. As the test set of the PANDA dataset is not public, we use the available WSIs to generate a train/validation/test split of 8469, 353 and 1794 WSIs, respectively. We extract 512x512 patches with a 50% overlap from the WSIs. The data was collected from two datacenters ('Radboud' and 'Karolinska') but only the WSIs from Radboud have local annotations of the Gleason grade while the annotations from

[1] Available at: https://data.mendeley.com/datasets/9xxm58dvs3/1
[2] Available at: https://www.kaggle.com/c/prostate-cancer-grade-assessment

Karolinska distinguish non-cancerous and cancerous. The exact classes of these cancerous patches remains unknown, so they can only be used as unlabeled data for training. We disregard all patches with less than 50% tissue and assign the label 'non-cancerous' to all patches that have at least 95% pixels annotated as non-cancerous or background. To the cancerous patches of Radboud we assign the Gleason grade which has the highest amount of pixels in comparison to the other cancer classes. The breast cancer experiments were conducted on Camelyon16[3] which contains 130 WSIs for testing and 270 WSIs for training/validation. We split them into 80% for training and 20% for validation. For the training/validation set, detailed annotations are available while the testing is done only at the WSI level in a binary manner (cancer vs no-cancer). For our experiments, we sliced the WSIs into non-overlapping 512x512 patches at 20x magnification and filtered out the patches that contain less than 5% tissue.

3.2 Metrics

To compare our method, we use the metrics that are reported for other state-of-the-art methods for the different datasets. For prostate cancer (SICAPv2 and PANDA), the common metric for comparison is Cohen's quadratic kappa, which measures the inter-rater reliability between the pathologist's annotations and the model's predictions. It is calculated based on the confusion matrix, and a kappa value of 0.0 indicates agreement by chance while 1.0 means complete agreement. This metrics takes into account that, in a set of ordered classes, error between consecutive classes should be less penalized and therefore it is especially suitable for Gleason grading. Further, we report the average F1 score, as in [25, 26]. The F1 score is based on the recall and precision per class and then averaged over the classes. For the breast cancer dataset Camelyon16, the commonly reported metric is the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC). The ROC curve is created by plotting the true positive rate against the false positive rate at various threshold settings and measures the diagnostic ability of a binary classifier. The AUC of the ROC is 0.5 for a random classifier and 1.0 for a perfect classifier. As our model uses pseudo labels, it is especially important to prove the reliability and robustness of our model. We perform multiple independent runs on the independent test sets to assure a reliable high performance of our method. The results are therefore reported as mean and standard deviation of the above described metrics.

3.3 Implementation Details

We implemented our model in tensorflow 2.3 and used one TITAN X (Pascal) GPU with 12 Gb for training with minibatches of 16 patches. The training until convergence with the EfficientNet-B5 backbone took approximately 7 hours (30 epochs) for SICAPv2, 7 days (10 epochs) for PANDA and 6 days (15 epochs) for Camelyon16. The time to perform predictions for inference is negligible for applications in clinical practice and lies below 2 seconds for a complete WSI on average for all used datasets. The model selection and hyperparameter tuning was performed on the four-fold cross validation set of SICAPv2. For the classification backbone, this work utilized the state-of-the-art image classification model EfficientNet [27] which was pre-trained on ImageNet and can scale with 8 different levels of complexity (B0-B7). We used the four-fold cross validation of SICAPv2 for the model selection and observed that an increasing complexity of the model led indeed to a better performance until EfficientNet-B5. Models B6 and B7 did not show any further improvements, so we chose EfficientNet-B5 as our backbone. The hyperparameters of the model were set to $k = 5$ (tested: $k = 1, 3, 5, 10$, used in Step 2 in 2.3) and $\lambda = 3$ (tested $\lambda = 1, 2, 3, 5$, used in equation 6 D) which showed the best results. The network was fine-tuned with stochastic gradient descent and the learning rate 0.01. For the relatively small dataset SICAPv2, class balanced loss was used (based on true y and estimated labels $\hat{y}$) to stabilize the training as we sometimes observed the convergence to 'bad' local minima (for the experiments $P = 0$ and $P = 1$ in Figure 3). We resized the image patches to 250x250 which is the input resolution for our model.

[3] Available at: https://camelyon16.grand-challenge.org/Data/

Table 1: Ablation studies on SICAPv2 with 5 patch labels per cancerous WSI and global label. The results are reported as the mean and standard deviation of 5 independent runs.

Model	Cohen's quadr. kappa	avg. F1 Score
GT	0.768 ± 0.009	0.688 ± 0.012
GT + PL	0.774 ± 0.012	0.697 ± 0.007
GT + SL	0.780 ± 0.012	0.698 ± 0.012
GT + SL + PL	$\mathbf{0.801 \pm 0.013}$	$\mathbf{0.700 \pm 0.011}$
MIL (Max-pooling)	0.545 ± 0.038	0.492 ± 0.026
SSL (Fixmatch)	0.774 ± 0.031	0.676 ± 0.009

For image augmentation brightness shift, random flip and rotation were used. The difference between weak and strong augmentation in our experiment was the intensity of the brightness shift (multiplication of the alpha channel with a factor) which led to a darker or brighter version of the original image. While the weak augmentation α uses random brightness shift factors between 0.9 and 1.1, the strong augmentation β uses a range from 0.5 to 1.5. The stronger the brightness shift, the harder it gets to visually recognize the pattern in the images.

3.4 Ablation Studies

To study the effect of the different loss components and the improvement over the SSL and MIL baselines, we performed an ablation study for the SICAPv2 dataset with efficient labeling (see section 2.1) and 5 patch labels per WSI and global label (equal to the experiment P=5 of section 3.5). In Table 1 we first compare 4 different label settings: only the available ground truth labels (GT), ground truth and soft labels (GT + SL), ground truth and pseudo labels (GT + PL) and ground truth, soft and pseudo labels (GT + SL + PL) which is our proposed setting. In the loss equation, terms A and B represent the ground truth, term C the pseudo labels and term D the soft labels. The model trained only with ground truth labels can be seen as a baseline because it simply uses all available labels in a supervised fashion. We observe that pseudo label as well as soft labels improved this baseline in both metrics. The best result was obtained using ground truth, pseudo and soft labels, and we therefore proved that all loss terms are relevant in practice.

We also compared our method to the SSL and MIL baselines to highlight the improvement of our combined solution. The chosen baseline implementations Max-pooling and Fixmatch are the algorithms that are the most related approaches in the fields of SSL and MIL (for details, see section 2.4). For the MIL baseline, we disregarded the available patch labels for training. Max-pooling inspired our pseudo-label assignment and is commonly used, f.e. in [2]: the global label was assigned to one patch with the highest class probability. The poor results of only 0.545 (Cohen's kappa) and 0.492 (F1 Score) highlight that the dataset is too small for this MIL-baseline method. Including some patch labels with our proposed method performs much better. To compare with the SSL baseline, we implemented the Fixmatch algorithm [12], which uses the available patch-labels but can not integrate the global WSI labels for training. For a fair comparison, we assigned negative patch labels to all patches of a negative WSI, although this is already beyond SSL in a strict sense. As proposed in the original paper, pseudo labels were assigned for cancer class predictions higher than 0.95. In this setup, the Fixmatch baseline showed a comparable performance to our proposed pseudo-label assignment (GT+PL) in terms of Cohen's kappa, but the F1 score was significantly lower. In comparison to our proposed final model (GT+SL+PL), the SSL baseline performed approximately 2.5 percentage points worse in both Cohen's kappa and F1 score. Overall, we see that utilizing a reduced number of patch labels and WSI labels with our approach achieved a substantial improvement over the SSL and MIL baselines.

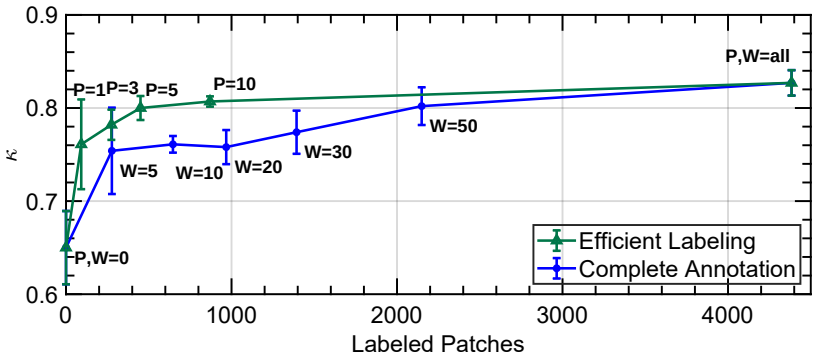


Figure 3: Comparison of data label settings. Efficient Labeling (EL) with P annotated patches per primary and secondary Gleason grade of all WSIs and Complete Annotation (CA) of W WSIs with all the patch labels. We plot the mean and standard deviation of Cohen's quadratic kappa (patch level) of five runs against the total amount of annotated training patches of SICAPv2.

3.5 Efficient Labeling vs. Complete Annotation

In the next experiment, we compared two different data settings for the prostate cancer dataset SICAPv2: We wanted to study whether with limited resources it is better to use Efficient Labeling (EL, see section 2) with a few patch labels from all available WSI or a few WSI with the Complete Annotation (CA). In the first case (EL) we randomly sampled a certain amount P of patch labels for the primary and secondary Gleason grade of each WSI. For the second approach (CA) we randomly selected W WSIs and used all patch labels of this selection for training. In Figure 3 we observe that the EL setting required substantially fewer labels than CA to obtain good results. We explain this by the higher variability of the annotated patches with EL that allows the network to learn from more diverse examples. The annotated patches of CA have a higher co-similarity and therefore contribute less information to the model training. The steep ascent of the performance from $P = 0$ to $P = 5$ proofs the efficiency of our learning approach and EL. To estimate the saved time and resources (to annotate a dataset) for the model training with our approach, we use the total amount of local annotations. Concretely, we count the total number of labeled patches used for training. We compare settings with a reduced number of patches with supervised training using all available patch labels. In the case of SICAPv2, our model with $P = 5$ and EL showed a performance close to the supervised one with only using 450 of the 4384 available patch labels. This means that approximately 10 times less labeled patches were needed for training.
For PANDA, the ratio of saved labeling effort is comparable: the model trained with $P = 5$ uses approximately 10 times less patch annotations for training than the supervised setup, but with much higher absolute numbers: while the supervised model is trained with 205,111 labeled patches, the model with $P = 5$ obtained 22,023 patch labels. For Camelyon16 (Table 3), the advantage is even bigger: the model with P=5 and EL used 433 patch labels while the supervised model trained with 21437 patch labels. This means, that only 2% of the complete training data was needed for the proposed approach, while the result remains close to the supervised performance, as reported in the next section.

3.6 Qualitative Evaluation

We qualitatively assessed the WSI predictions for SICAPv2 and show visual examples in Figure 4. For comparison, the predictions of our proposed model trained without any patch labels ($P = 0$), with some patch labels ($P = 5$) and all patch labels ($P = all$) are depicted as well as the ground truth annotations. We observe that the model trained without any patch labels in a MIL setting correctly marks the cancerous areas but has problems to assign the right classes to the tissue. This highlights the limitations of MIL models trained on relatively small datasets for complex multi-class scenarios. The model with some patch annotations ($P = 5$) shows a robust performance which is

Table 2: Comparison with previous works of prostate cancer patch-level Gleason grading. We report the average result of 5 independent runs.

Method	Learning	Dataset	Cohen's quadr. kappa	avg. F1 Score
Arvaniti et al. [26] (2018)	Supervised	(other)*	$0.55/0.49$	–
Nir et al. [28] (2018)	Supervised	(other)*	0.60	–
Otálora et al. [17] (2020)	MIL + SSL	(other)*	$0.59/0.55$	–
Silva-Rodríguez et al. [25] (2020)	Supervised	SICAPv2	0.77	0.66
Ours (P=5; 450 positive patch labels)	MIL + SSL	SICAPv2	0.801	0.700
Ours (P=10; 870 positive patch labels)	MIL + SSL	SICAPv2	0.807	0.710
Ours (P=all; 4,384 pos. patch labels)	Supervised	SICAPv2	0.827	0.718
Ours (P=5; 22,023 positive patch labels)	MIL + SSL	PANDA	0.794	0.739
Ours (P=10; 41,910 positive patch labels)	MIL + SSL	PANDA	0.830	0.735
Ours (P=all; 205,111 pos. patch labels)	Supervised	PANDA	**0.891**	**0.812**
Inter-Pathologists [26]*			0.65	–

* Results reported on different datasets, patch size and resolutions, see [26], [28] and [17] for details.

Table 3: Comparison with previous works of metastasis detection in sentinel lymph nodes of breast cancer patients (Camelyon16). Our reported results are the average of 3 independent train and test runs.

Method	Learning	ROC AUC
Camelyon16 Winner [29]	Supervised	0.923
Camelyon16 Best on Leaderboard [29, 30]	Supervised	**0.994**
Campanella et al. [2] **	MIL	0.899
Campanella et al. [2] ***	MIL	0.965
Ours (P=5; 433 positive patch labels)	MIL + SSL	0.913
Ours (P=all; 21,437 pos. patch labels)	Supervised	0.933
Pathologists with time constraints [30]		0.810
Pathologists without time constr. [30]		0.966

** Tested on Camelyon16, trained on MSK breast dataset (total 9894 WSIs, see [2] for details)

*** Trained and tested on MSK breast dataset (total 9894 WSIs, see [2] for details)

close to the prediction of the supervised model ($P = all$). This confirms the reliability of the proposed method, which uses pseudo labels to complement a small amount of patch labels. Note that both models, $P = 5$ and $P = all$, highlight some areas as Gleason Grade 3 that are annotated as non-cancerous. This can be explained by the interpolation in between patches to produce the graphic and the ambiguity in the Gleason grading task: even between pathologists, a complete agreement on the exact cancerous regions is rare, as reported in Table 2.

3.7 Comparison with State of the Art

In this section we report the results for the three datasets: SICAPv2, PANDA and Camelyon16, and compare our proposed method with efficient labeling (EL, see section 2.1) to other state-of-the-art approaches. In Table 2 we show the performance of patch level classifiers of Gleason grades. We observe that our model is able to achieve competitive results with only 5 patch labels per WSI and global label. For the relatively small dataset SICAPv2, our model with $P = 5$ achieves a remarkable result of 0.807 Cohen's kappa, outperforming the existing supervised state-of-the-art [25] for this dataset. In this setting, the model only required a total of 433 labeled patches. Our model in the completely supervised setting reached a slightly better result, but using approximately 10x more patch labels. For a larger prostate cancer dataset, PANDA, we observe similar results. The model with $P = 10$ achieved a remarkable Cohen's kappa value of 0.830 and an average F1 score of 0.735. Note that the gap in comparison to the supervised model (with a Cohen's kappa of 0.891 and an average F1 Score of 0.812) is slightly larger than for the SICAPv2 experiment. This can

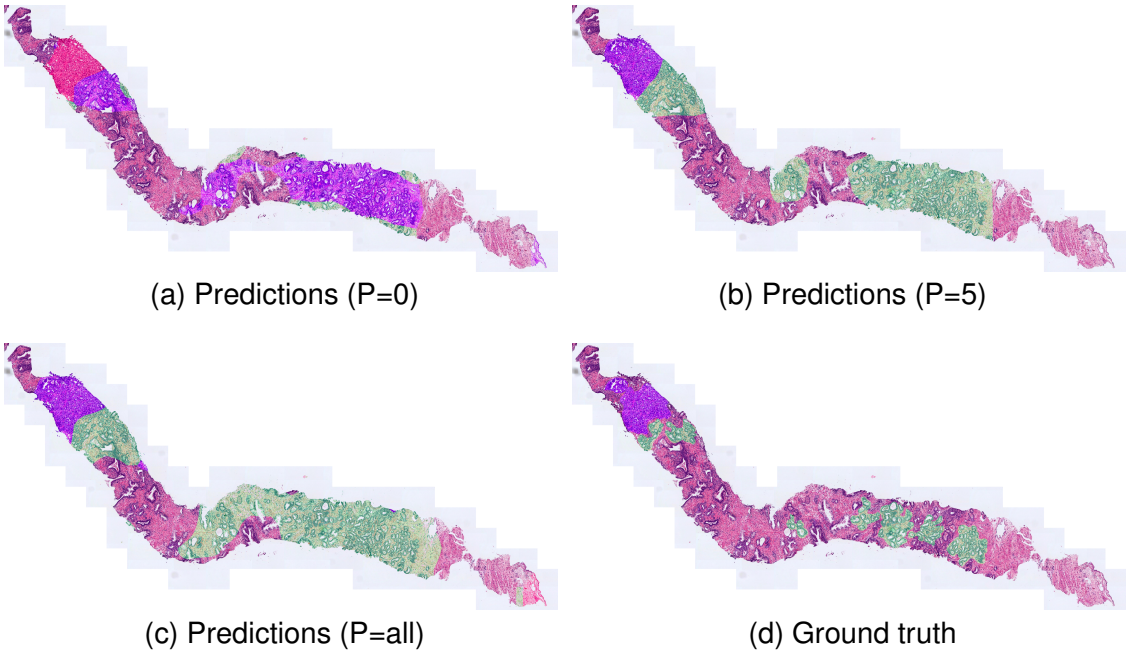


Figure 4: Visual example of model predictions for a test WSI of SICAPv2. The cancerous areas are marked in green (Gleason Grade 3), blue (Gleason Grade 4) and red (Gleason Grade 5). We compare the model predictions trained with $P = 0$ (MIL), $P = 5$ (some patch labels), $P = all$ (supervised), corresponding to the patch labels per class and WSI available during training. The marked areas of predictions are interpolated from patch-level predictions and therefore not as fine-grained as the ground-truth annotation. While the model trained in a MIL setting (a) correctly identifies the cancerous areas, the predicted classes are incorrect. The model predictions with the setting $P = 5$ depicted in (b) are very similar to those of the supervised model (c) and the ground truth (d).

be explained by the much higher absolute number of labeled patches for the supervised setting (205,111 patch labels). In this case, the model learns to mimic the pathologist's annotation very accurately. It is noteworthy to mention that, as the inter-pathologist agreement for this task lies round 0.65 [26], all Cohen's kappa values above 0.8 indicate a very high agreement with the given annotation. The proposed SSL+MIL approach with $P = 10$ shows a very good performance, while 163,201 less patch labels were used than in the supervised approach (P=10: 41,910 patch labels; supervised: 205,111 patch labels).

Table 3 shows the results for the detection of lymph node metastasis of breast cancer with the dataset Camelyon16. As Camelyon16 allows only the evaluation of the global WSI labels, we derived the cancer probability simply from the highest patch probability per WSI. Although our model's primary strength is the instance (patch-level) classification, we obtained a competitive Camelyon16 result with $P = 5$ (ROC AUC $= 0.913$) close to the supervised performance $P = all$ (ROC AUC $= 0.933$) while using approximately 50 times less patch labels during training. Further, the results with $P = 5$ are still more than 10 percentage points above pathologists with realistic time constraints (ROC AUC $= 0.810$). The strong performance proves the model's good generalization to different cancer types and the high accuracy of the instance predictions: bag labels can reliably be derived from them by a simple heuristic. Note that the MIL approach of Campanella et al. [2] had strong results but has some limitations: the method trained on a 20 times larger dataset and only predicted binary labels on the bag level. Our model, trained only on the Camelyon16 training set, is able to provide patch-level predictions and extendable to multiclass-settings.

3.8 Advantages and Limitations of the Proposed Method

The proposed method has several advantages in comparison to other existing approaches. First of all, it showed a high performance while training with limited resources: A total of 450 labeled patches of 155 WSIs for prostate cancer and 433 labeled patches of 400 WSIs for breast cancer were sufficient to obtain competitive results. This confirms the effectiveness of the proposed combination of MIL and SSL techniques. Furthermore, it can adapt flexibly to any amount of available patch labels, as shown in the experiments of Figure 3. Depending on the available annotations, even unlabeled WSIs or completely annotated WSIs can be easily integrated in the training procedure. Regarding the best labeling strategy, the proposed efficient labeling strategy showed very good results with limited annotations, as highlighted in subsection 3.5. It can be recommended for the future annotation of datasets. Still, there are some limitations of our method. When no patch labels are available, the proposed method can still be used for training, but the performance was not comparable to the supervised training result, as shown in Figure 3. In the default MIL setting, other specialized MIL methods might provide a better performance [2,9]. Furthermore, our method assumes that the label classes on instance and bag level are the same. For problems where the local cancer class differs from the overall WSI labels, the proposed algorithm needs to be adjusted.

4 Conclusions

We have presented a flexible deep learning framework for cancer classification which is able to make very accurate local as well as global predictions while requiring significantly fewer annotations than supervised approaches. The success of this approach can be attributed to the combination of semi-supervised and multiple instance learning as well as the proposed efficient labeling strategy, which was experimentally quantified. The work of the pathologist in our setting reduces to the annotation of some cancerous patches in each WSI and the final diagnosis. With this work, we hope to significantly contribute to the efforts of improving cancer diagnosis with the help of deep learning. By reducing the dependency on large, completely annotated datasets, we lower the threshold for new applications of artificial intelligence. With our approach, researchers and engineers can train deep learning models for cancer classification problems for which deep learning was not yet applied because of data limitations. To further improve our approach, we propose two future research directions: (i) active learning algorithms to choose the most discriminative patches for labeling and (ii) the use of an additional bag-level classifier based on the models feature maps to obtain even better WSI-level results.

References

[1] N. Dimitriou, O. Arandjelović, and P. D. Caie, "Deep learning for whole slide image analysis: An overview," *Frontiers in Medicine*, vol. 6, p. 264, 2019.

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

[3] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Medical Image Analysis*, vol. 54, pp. 280–296, 2019.

[4] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," pp. 3376–3391, 2018.

[5] M. Lerousseau, M. Vakalopoulou, M. Classe, J. Adam, E. Battistella, A. Carré, T. Estienne, T. Henry, E. Deutsch, and N. Paragios, "Weakly Supervised Multiple Instance Learning Histopathological Tumor Segmentation," 2020, pp. 470–479.

[6] G. Xu, Z. Song, Z. Sun, C. Ku, Z. Yang, C. Liu, S. Wang, J. Ma, and W. Xu, "CAMEL: A Weakly Supervised Learning Framework for Histopathology Image Segmentation," *International Conference for Commputer Vision (ICCV)*, no. cMIL, 2019.

[7] Y. Huang and A. C. Chung, "Evidence localization for pathology images using weakly supervised learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 613–621.

[8] P. Chikontwe, M. Kim, S. J. Nam, H. Go, and S. H. Park, "Multiple instance learning with center embeddings for histopathology classification," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2020, pp. 519–528.

[9] N. Hashimoto, D. Fukushima, R. Koga, Y. Takagi, K. Ko, K. Kohno, M. Nakaguro, S. Nakamura, H. Hontani, and I. Takeuchi, "Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3851–3860.

[10] J. V. Pulido, S. Guleria, L. Ehsan, M. Fasullo, R. Lippman, P. Mutha, T. Shah, S. Syed, and D. E. Brown, "Semi-Supervised Classification of Noisy, Gigapixel Histology Images," in *International Conference on Bioinformatics and Bioengineering (BIBE)*, 2020, pp. 563–568.

[11] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *International Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[12] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 596–608.

[13] A. K. Jaiswal, I. Panshin, D. Shulkin, N. Aneja, and S. Abramov, "Semi-supervised learning for cancer detection of lymph node metastases," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[14] S. Shaw, M. Pajak, A. Lisowska, S. Tsaftaris, and A. O'Neil, "Teacher-student chain for efficient semi-supervised histology image classification," in *International Conference on Learning Representations (ICLR)*, 2020.

[15] N. Marini, S. Otálora, H. Müller, and M. Atzori, "Semi-supervised training of deep convolutional neural networks with heterogeneous data and few local annotations: An experiment on prostate histopathology image classification," vol. 73, p. 102165, 2021.

[16] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," in *International Conference on Neural Information Processing Systems (NeurIPS)*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33, 2020, pp. 6256–6268.

[17] S. Otálora, N. Marini, H. Müller, and M. Atzori, "Semi-weakly supervised learning for prostate cancer image classification with teacher-student deep convolutional networks," *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, vol. 12446 LNCS, pp. 193–203, 2020.

[18] J. Li, W. Chen, X. Huang, S. Yang, Z. Hu, Q. Duan, D. Metaxas, H. Li, and S. Zhang, "Hybrid supervision learning for pathology whole slide image classification," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2021, pp. 309–318.

[19] M. Y. Lu, R. J. Chen, and F. Mahmood, "Semi-supervised breast cancer histology classification using deep multiple instance learning and contrast predictive coding," in *Medical Imaging 2020: Digital Pathology*, vol. 11320, 2020.

[20] S. Budd, E. C. Robinson, and B. Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis," *Medical Image Analysis*, vol. 71, p. 102062, 2021.

[21] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *International Conference on Neural Information Processing Systems (NeurIPS)*, 2005, p. 17.

[22] D. Lee, "Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks," in *International Conference on Machine Learning (ICML)*, 2013.

[23] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," in *International Conference on Neural Information Processing Systems (NeurIPS)*, 2016, pp. 1171–1179.

[24] A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Label propagation for deep semi-supervised learning," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5065–5074.

[25] J. Silva-rodríguez, A. Colomer, M. A. Sales, R. Molina, and V. Naranjo, "Going deeper through the Gleason scoring scale : An automatic end-to-end system for histology prostate grading and cribriform pattern detection," *Computer Methods and Programs in Biomedicine*, vol. 195, 2020.

[26] E. Arvaniti, K. S. Fricker, M. Moret, N. Rupp, T. Hermanns, C. Fankhauser, N. Wey, P. J. Wild, J. H. Rüschoff, and M. Claassen, "Automated Gleason grading of prostate cancer tissue microarrays via deep learning," *Scientific Reports*, vol. 8, no. 1, pp. 1–11, 2018.

[27] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning (ICML)*, vol. 97, 2019, pp. 6105–6114.

[28] G. Nir, S. Hor, D. Karimi, L. Fazli, B. F. Skinnider, P. Tavassoli, D. Turbin, C. F. Villamil, G. Wang, R. S. Wilson, K. A. Iczkowski, M. S. Lucia, P. C. Black, P. Abolmaesumi, S. L. Goldenberg, and S. E. Salcudean, "Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts," pp. 167–180, 2018.

[29] "Camelyon16 challenge results," https://camelyon16.grand-challenge.org/Results/, accessed: 2021-12-21.

[30] B. E. Bejnordi, M. Veta, P. J. van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J. A. W. M. van der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol, O. Geessink, N. Stathonikos, M. V. van Dijk, P. Bult, F. Beca, A. Beck, D. yong Wang, A. Khosla, R. Gargeya, H. Irshad, A. Zhong, Q. Dou, Q. Li, H. Chen, H. Lin, P. Heng, C. Hass, E. Bruni, Q. J. J. Wong, U. Halici, M. Ü. Öner, R. Cetin-Atalay, M. Berseth, V. Khvatkov, A. Vylegzhanin, O. Z. Kraus, M. Shaban, N. Rajpoot, R. Awan, K. Sirinukunwattana, T. Qaiser, Y. Tsang, D. Tellez, J. Annuscheit, P. Hufnagl, M. Valkonen, K. Kartasalo, L. Latonen, P. Ruusuvuori, K. Liimatainen, S. Albarqouni, B. Mungal, A. George, S. Demirci, N. Navab, S. Watanabe, S. Seno, Y. Takenaka, H. Matsuda, H. A. Phoulady, V. Kovalev, A. Kalinovsky, V. Liauchuk, G. Bueno, M. M. Fernández-Carrobles, I. Serrano, Ó. Déniz, D. Racoceanu, and R. Venâncio, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Journal of the American Medical Association*, vol. 318, p. 2199–2210, 2017.

Chapter 5

Probabilistic Active Learning for the Uncertainty-based Acquisition of Image Patches

5.1 Publication details

Authors: Arne Schmidt, Pablo Morales-Álvarez, Lee A. D. Cooper, Lee A. Newberg, Andinet Enquobahrie, Rafael Molina, and Aggelos K. Katsaggelos
Title: Focused Active Learning for Histopathological Image Classification **Status:** Under review **Reference:** Medical Image Analysis
Quality indices:

- Impact Factor (JCR 2022):
  -

5.2 Main contributions

- We propose the novel FocAL (Focused Active Learning) algorithm, a combination of a Bayesian neural network and Out of Distribution (OoD) detection. It is able to acquire new unlabeled images by estimating the informativeness of each image by the epistemic uncertainty. Furthermore, images with a high aleatoric uncertainty and a high OoD score are actively avoided because they correspond to images with artifacts or ambiguities.
- In experiments with MNIST and histopathological images we show that existing state-of-the-art active learning algorithms are prone to acquire uninformative images which harms the training and acquisition procedure. FocAL avoids these images and shows the overall best performance.

FOCUSED ACTIVE LEARNING FOR HISTOPATHOLOGICAL IMAGE CLASSIFICATION

Arne Schmidt
Dep. of Computer Science and AI
University of Granada
Granada, Spain

Pablo Morales-Álvarez
Dep. of Statistics and Operations Research
University of Granada
Granada, Spain

Lee A.D. Cooper
Dep. of Pathology
Northwestern University
Chicago, USA

Lee A. Newberg
Kitware Inc.
Carrboro, USA

Andinet Enquobahrie
Kitware Inc.
Carrboro, USA

Aggelos K Katsaggelos
Dep. of Electrical Computer Engineering
Northwestern University
Evanston, USA

Rafael Molina
Dep. of Computer Science and AI
University of Granada
Granada, Spain

ABSTRACT

Active Learning (AL) has the potential to solve a major problem of digital pathology: the efficient acquisition of labeled data for machine learning algorithms. However, existing AL methods often struggle in realistic settings with artifacts, ambiguities, and class imbalances, as commonly seen in the medical field. The lack of precise uncertainty estimations leads to the acquisition of images with a low informative value. To address these challenges, we propose Focused Active Learning (FocAL), which combines a Bayesian Neural Network with Out-of-Distribution detection to estimate different uncertainties for the acquisition function. Specifically, the weighted epistemic uncertainty accounts for the class imbalance, aleatoric uncertainty for ambiguous images, and an OoD score for artifacts. We perform extensive experiments to validate our method on MNIST and the real-world Panda dataset for the classification of prostate cancer. The results confirm that other AL methods are 'distracted' by ambiguities and artifacts which harm the performance. FocAL effectively focuses on the most informative images, avoiding ambiguities and artifacts during acquisition. For both experiments, FocAL outperforms existing AL approaches, reaching a Cohen's kappa of 0.764 with only 0.69% of the labeled Panda data.

Keywords Active Learning · Cancer Classification · Histopathological Images · Bayesian Deep Learning

1 Introduction

Artificial Intelligence (AI) methods have obtained impressive results in digital pathology and in some cases, AI models even outperformed expert pathologists in cancer classification [1, 2, 3]. The

hope is that AI can make the diagnosis more accurate, objective, reproducible, and faster in the future [4].

To achieve this goal, trained, specialized AI models for each subtask are required, for example for the quantification of tumor-infiltrating lymphocytes in lung cancer [5], metastasis detection of breast cancer in lymph nodes [3, 6] or Gleason grading of prostate cancer [7, 8]. Openly available, labeled datasets are limited to certain subtasks and for many future applications, the aggregation of large amounts of labeled data remains challenging because the annotation requires medical experts. This makes the labeling process time-consuming and expensive. A common approach to labeling is to divide a region or Whole Slide Image (WSI) into small patches that are individually labeled [4]. The model is then trained to make local patch-level predictions that can be aggregated for the final diagnosis. The problem with supervised deep learning methods is the need for large amounts of detailed (patch-level) annotations for training to obtain a satisfying predictive performance. To alleviate this burden, *semi-supervised learning* [9, 10, 11, 6, 8] and *multiple instance learning* [12, 13, 14] have become major fields of interest in the recent years. Despite the importance of these fields, there is another approach to efficiently handle labeling resources with several advantages over semi-supervised and multiple instance learning.

Active Learning (AL) describes machine learning methods that actively query the most informative labels. In the AL setting, the AI model starts training with a small set of labeled images and iteratively selects images from a large pool of unlabeled data. These selected images are labeled in each iteration by an 'oracle,' in our application a medical expert. AL has several advantages over semi-supervised and multiple instance learning: (i) The model training and dataset creation go hand-in-hand. The performance of the model is constantly monitored to assess if the collected labeled data is enough - or if more labeled data is needed to reach the desired performance. (ii) The model looks for the most informative images automatically in the acquisition step. In semi-supervised learning in comparison (and other methods that require labeling), finding these informative, salient images requires a lot of manual searching. (iii) AL is very data-efficient while multiple instance learning often requires large datasets to compensate for missing instance labels [12]. Furthermore, the AL model can be trained to make accurate local (patch-level) predictions while multiple instance learning models often do not provide those and therefore lack explainability. In other cases, the multiple instance learning setting is not applicable at all because the global classes differ from the local classes. As an example consider images with global labels 'cat' and 'dog'. It would not be possible to train a multiple instance learning model for classifying 'paws' and 'ears' at the image-patch level because this information simply can not be deducted from the global labels. AL models can be trained to make any kind of local predictions with reduced labeling effort.

Related work In AI research, different AL strategies have been proposed to determine the most informative images. Early approaches used the uncertainty estimation of support vector machines [15], Gaussian processes [16] or Gaussian random fields [17] to rate the image informativeness. With the rise of deep learning, the focus shifted to Bayesian Neural Networks (BNNs) for AL [18], which was adapted several times for histopathological images [19, 20, 21]. This approach has the advantage of a probabilistic uncertainty estimation which is not only used for acquisition, but it is also crucial for diagnostic predictions in medical applications. BNNs allow the application of several different uncertainty-based acquisition functions, such as *BALD* [22], *Max Entropy* [19, 23], and *Mean Std* [24]. Other publications focus on the user interface and server application of AL [25, 26] rather than the AL model itself. In the existing literature, the uncertainty estimation is often only used to determine the amount of new information in each image. We extend this idea by using complementary uncertainty measures to *avoid* labeling uninformative, ambiguous, or artifactual images. In digital pathology, several data-related challenges like artifacts, ambiguities, and the typical huge class imbalance hinder the application of AL (see "Problem analysis" paragraph below). Our proposed method tackles these problems successfully by precise uncertainty estimations which leads to improved performance.

BNNs are not only of interest for the AL acquisition, their capacity to estimate the predictive uncertainty is highly important in safety-critical areas like medicine [27] or autonomous driving [24]. The

uncertainty estimation helps to distinguish confident predictions from risky ones. In our case, we aim to decompose uncertainty into *epistemic uncertainty* and *aleatoric uncertainty* describing the model and data uncertainty, respectively [28]. Epistemic uncertainty describes uncertainty in model parameters that can be reduced by training with additional labeled data. Therefore it can serve as a measure of informativeness in the active learning process. Unfortunately, the epistemic uncertainty is not only high for informative, in-distribution images, but also for OoD images. In fact, epistemic uncertainty has recently been used explicitly for OoD detection [29, 30, 31]. Aleatoric uncertainty describes irreducible uncertainty in the data due to ambiguities that cannot be improved with additional labeling. Studies have shown that training with ambiguous data can harm the performance of the algorithm considerably if not taken into account [32, 33]. In the Panda challenge, label noise associated with the subjective grading assigned by pathologists was considered to be a major problem [7].

To estimate these uncertainties with BNNs, Kendall *et al.*[34] proposed a network with two final probabilistic layers, corresponding to the two uncertainty measures. A theoretically sound, more stable, and efficient approach (relying on a single probabilistic layer) was proposed by Kwon *et al.*[27]. We base our BNN for uncertainty estimations on the latter method due to the mentioned advantages. In Section 3.2 we outline how the uncertainty estimations can be interpreted in the context of clinical applications like pathology.

To avoid acquiring image patches with artifacts, we apply OoD detection. Commonly, OoD data refers to data that originates from a different distribution than the training data (in-distribution) [35]. In the context of AL and pathology, we define the in-distribution as the distribution of patches containing (cancerous or non-cancerous) tissue. All the images with artifacts (such as pen markings, tissue folds, blood, or ink) [36] will be considered OoD. These artifacts are inevitable in real-world data and there are several reasons to exclude them from the distribution of interest for acquisition: (i) It is impossible to learn all possible artifacts explicitly due to their wide variability. We argue that a model should reliably classify tissue and predict a high uncertainty for everything it does not know. (ii) It harms the performance of AL algorithms to acquire images with artifacts, as we show empirically in Section 3. (iii) The model should focus on learning what *is* cancerous instead of everything that *is not* cancerous. By learning cancerous patterns it automatically learns what is not cancerous (everything else).

In OoD detection, early methods used the depth [37, 38] or distance [39, 40] of datapoints, represented by low-dimensional feature vectors. With the rise of deep learning, OoD metrics were often applied to the features extracted by a deep neural network [41, 35, 42]. In this line with previous research, we utilize extracted feature vectors and implement a density-based OoD scoring method [43] to detect artifacts in the data.

Problem Analysis Although AL has a huge potential for digital pathology, we analyze several challenges that hinder its application in practice:

- Medical imaging problems like pathology often have a high class imbalance. For example, in prostate cancer grading, the highest Gleason patterns may be underrepresented which needs to be taken into account during acquisition. Other AL algorithms treat each class equally and are not able to acquire a sufficient number of images of this underrepresented class in our experiments (Section 3).
- Many patches are ambiguous. There may be patches for which even subspecialists disagree on their label, or patches containing multiple classes. Assigning labels to these patches is difficult and may be detrimental to the quality of the dataset and the algorithm's performance. This not only slows the labeling process down, but it can also add noise to the training data as only one label per patch is assigned. In fact, label noise associated with the subjective grading assigned by pathologists was considered one key problem in the Panda challenge [7].
- WSIs can contain many different artifacts, such as pen markings, tissue folds, ink, or cauterized tissue. Existing AL algorithms often assign a high informativeness to these patches

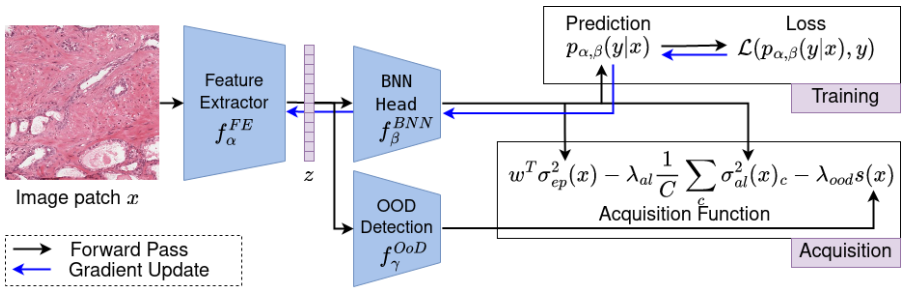


Figure 1: Model overview of the proposed FocAL method. It consists of three main components: feature extractor, BNN, and OoD detection (blue boxes). The figure shows how the different components are combined for training and acquisition.

although they do not contain important information for model training, as we show empirically in the experimental section 3.

We want to stress that similar problems of class imbalance, ambiguities, and artifacts are present in many other medical imaging applications, such as CT scans for hemorrhage detection [44], dermatology images for skin cancer classification [45] or retinal images for the detection of retinopathy [46].

Contribution To address these challenges we propose Focused Active Learning (FocAL), a probabilistic deep learning approach that focuses on the underrepresented malignant classes while ignoring artifacts and ambiguous images. More specifically, we combine a Bayesian Neural Network (BNN) with Out of Distribution (OoD) Detection to estimate the three major elements of the proposed acquisition function. The *weighted epistemic uncertainty* rates the image informativeness, taking the class imbalance into account. The *aleatoric uncertainty* is used to avoid ambiguous images for acquisition. The *OoD score* helps to ignore outliers (like artifacts) that do not contribute information for the classification of tissue. We show empirically that these precise uncertainty estimations help to focus on labeling salient, informative images while other methods often fail to address this realistic data setting.

The article is structured as follows. We outline the theory of the proposed model, including the BNN and OoD components of the acquisition function in Section 2. In Section 3, we perform an illustrative MNIST experiment to analyze the behavior of existing AL approaches when artifacts and ambiguities are present. Furthermore, we demonstrate that each of our model components works as expected to avoid acquiring images with ambiguities and artifacts, overcoming the problems of the existing approaches. For the Panda prostate cancer datasets we perform an ablation study about the introduced hyperparameters, analyze the uncertainty estimations, and report in the final experiments that our method can reach a Cohen's kappa of 0.763 with less than 1% of the labeled data (4400 labeled image patches). Finally, in Section 4 we conclude our article and give an outlook of future research.

2 Methods

Here we describe the three elements of FocAL: the feature extractor, the Bayesian Neural Network, and the Out-of-distribution score. The final paragraph of this section outlines the acquisition function and algorithm of the novel FocAL method. An overview of the model components is depicted in Figure 1.

Active Learning (AL) In AL we assume that at the beginning a small set of labeled data $\mathcal{D}_{train} = \{x_i, y_i\}_{i=1,..,N}$ of images x_i and labels y_i and a pool of unlabeled data $\mathcal{D}_{pool}$ is available. We assume each y to be a C-dimensional, one-hot encoded vector, where C stands for the number of classes. A machine learning model $\mathcal{M}$ trains with a labeled set $\mathcal{D}_{train}$ and then chooses

a subset $\mathcal{A} \subset \mathcal{D}_{pool}$ of unlabeled images to be labeled (by a specialist, such as a pathologist in the given application). The choice is made with the help of an *acquisition function* $a(x, \mathcal{M})$ which estimates the informativeness of each image x. Probabilistic Active Learning The images with the highest acquisition scores are labeled and added to $\mathcal{D}_{train}$. Then, the model is retrained with the updated training set. This acquisition step is repeated iteratively such that the model performance increases while more and more labeled data is aggregated.

Feature Extraction The feature extractor f_α^{FE}, with model parameters α, is the first component of the proposed model. We use a Convolutional Neural Network (CNN) to extract high-level 128-dimensional features $z = f_\alpha^{\mathrm{FE}}(x)$ from each image patch. The exact architecture of the feature extractor depends on the image data and task, see the implementation details for each experiment in Section 3. The feature extractor is trained during the AL process end-to-end with the BNN by gradient descent, see Figure 1. Although training the feature extractor is important for obtaining good final results, the emphasis of this article lies on the development of the BNN and OoD detection which perform the high-level reasoning, as described in the next paragraphs.

Bayesian Neural Network (BNN) The BNN model f_β^{BNN}, with model parameters β, allows probabilistic reasoning based on the extracted feature vectors $Z = \{z_i\}$. Note that the feature extractor and BNN head together could also be interpreted as a large, convolutional BNN with Bayesian layers near the output. Previous studies have shown that this combination of deterministic convolutions and Bayesian fully connected layers is the most effective way to introduce Bayesian uncertainty in the AL context [47]. Here, we treat the feature extractor and BNN separately, because the features are also used later for OoD detection. The BNN is not only able to make accurate classification predictions, but can also estimate epistemic and aleatoric uncertainty. These estimated uncertainties will be further described below. They play a crucial role in the proposed acquisition function (see the last paragraph of this section).

The BNN in our model consists of two fully connected layers with 128 units and a final softmax output layer. In comparison to deterministic networks with weight parameters ω, BNNs treat the model weights as random variables with a probability distribution $p(\omega)$. As the true posterior distribution $p(\omega|\mathcal{D}_{train})$ is intractable, it has to be approximated. Following the success of similar approaches in recent studies [18, 24, 27], we use variational inference to approximate the posterior distribution by a tractable variational distribution $q_\beta(\omega)$, where β describe the variational parameters of the distribution. Specifically, we define q as a product of independent Gaussian distributions over each model weight, parametrized by mean and variance. To approximate the real posterior, the minimization of the KL divergence $KL(q_\beta(\omega)|p(\omega|\mathcal{D}_{train}))$ is achieved by maximizing the evidence lower bound (ELBO) utilizing the reparametrization trick [48]. Gradient descent allows the optimization of the BNN and the feature extractor end-to-end. We denote the ELBO loss function as $\mathcal{L}(p_{\alpha,\beta}(y|x), y)$. The predictive distribution $p_{\alpha,\beta}(y|x)$ is obtained by applying the feature extractor $z = f_\alpha^{\mathrm{FE}}(x)$ and integrating the BNN through Monte Carlo sampling as:

$$
\begin{aligned}
p_\beta(y|z) &= \int p(y|\omega, z) q_\beta(\omega)\, d\omega \\
&\approx \frac{1}{T} \sum_{t=1}^{T} p(y|z, \omega_t),
\end{aligned}
\tag{1}
$$

where we use Monte Carlo sampling by drawing T realizations $\{\omega_t\}_{t=1,..,T}$ of the variational weight distribution. The *argmax* over classes of the vector $p_\beta(y|z)$ defines the predicted class. For notational convenience, we will drop the parameters α and β when not needed. We show in Figure 1 an overview of the forward and backward pass with gradient descent.

Bayesian Uncertainty Estimations In Addition to the class prediction (eq. 1), the BNN is able to estimate the uncertainty, measured by the predictive covariance matrix $\mathrm{Cov}_{p(y^*|z^*,Z,Y)}(y^*)$. This variance can be further decomposed into epistemic and aleatoric uncertainty.

Epistemic uncertainty (model uncertainty) measures the uncertainty introduced by the model parameters ω and can be reduced with more labeled training data. A high epistemic uncertainty

indicates a high informativeness of a given image. Unfortunately, using only the epistemic uncertainty for acquisition can lead to an unwanted outcome. If the data is contaminated with outliers such as artifacts, this can result in acquiring only outliers that do not contribute any value towards learning the classes of interest, as shown empirically in the experimental section 3.

Aleatoric uncertainty (data uncertainty) captures the uncertainty inherent in the data due to ambiguities. It can not be reduced with more labeled data. Images with a high aleatoric uncertainty should be avoided during acquisition as the chance of mislabeling (due to ambiguous image content) or inherent data noise is higher for these images.

There are different possibilities for estimating epistemic and aleatoric uncertainty. Here, we follow the approach of Kwon *et al.*[27], which does not require additional parameters, is numerically stable, and has a strong theoretical background. The covariance matrix is decomposed into

$$\begin{aligned}\mathrm{Cov}_{p(y^*|z^*,Z,X)}(y^*) = &\underbrace{\frac{1}{T}\sum_{t=1}^{T}\{p(y^*|z^*,\omega_t) - \hat{p}(y^*|z^*)\}^{\otimes 2}}_{:=\text{epistemic uncertainty}} \\ &+ \underbrace{\frac{1}{T}\sum_{t=1}^{T}\mathrm{diag}\{p(y^*|z^*,\omega_t)\} - p(y^*|z^*,\omega_t)^{\otimes 2}}_{:=\text{aleatoric uncertainty}}\end{aligned} \tag{2}$$

where $\hat{p}(y^*|z^*) = \frac{1}{T}\sum_{t=1}^{T} p(y^*|z^*,\omega_t)$, $diag\{p(y^*|z^*,\omega_t)\}$ is the diagonal matrix formed by the vector entries of $p(y^*|z^*,\omega_t)$ in the diagonal, and the outer product $v^{\otimes 2} = vv^T$.
Note that both epistemic and aleatoric uncertainties are given as $C \times C$ covariance matrices with values of the *uncertainty per class* on the diagonal. We define the C- dimensional vectors of class-wise uncertainties as σ_{ep}^2 for the epistemic and σ_{al}^2 for the aleatoric uncertainty.

Out-of-Distribution (OoD) Detection For the OoD detection, we use an unsupervised, density-based model f_γ^{OoD} with parameters γ [1], based on the extracted features z. Instead of having a binary decision (in/out of distribution), we want to score each feature vector of the unlabeled images with a Local Outlier Factor (LOF) [43]. The LOF is based on the k-nearest neighbors $N_k(z)$ of a vector z and the *local reachability density* $lrd_k(z)$, a density measure based on the distance to the k-nearest-neighbors.
The LOF of a vector z is defined as

$$\mathrm{LOF}_k(z) = \frac{1}{|N_k(z)|}\sum_{z' \in N_k(z)} \frac{lrd_k(z')}{lrd_k(z)} \tag{3}$$

with hyperparameter k which should be set to the minimum amount of expected datapoints in a cluster [43]. Intuitively, the LOF is high if a feature vector lies in a region with a lower density than its neighbors (indicating an outlier). If the region of a feature vector has the same density as its neighbors, its LOF is close to 1. The upper bound depends on the characteristics of the data, i.e. the distances between feature vectors. Empirically we observed that scaling the LOF by 0.1 leads to an OoD score that is in same the range as the other uncertainty measures (epistemic and aleatoric uncertainty). Therefore, we define the outlier scoring function as

$$s(x) = 0.1\ LOF_k(f^{\mathrm{FE}}(x)). \tag{4}$$

Note that the scaling factor does not introduce an additional hyperparameter. It is inherently tuned by manipulating the weighting factor λ_{ood} in eq. 5 for which we perform experiments in section 3.2. The hyperparameter k can be set to a rough estimate of the minimum number of initial images

[1] Note that the parameters γ consist of the locations and densities of the currently labeled feature vectors. These are not model parameters in the strict sense but we follow this notation for coherence.

$x \in \mathcal{D}_{train}$ that are not affected by ambiguities or artifacts. We set it to $k = 10$ for MNIST and $k = 50$ for the Panda dataset.

Focused Active Learning (FocAL) We propose an acquisition function that combines the uncertainty-based measures and the OoD scoring discussed above:

$$a(x, \mathcal{M}) = \underbrace{w^T \sigma^2_{ep}(x)}_{\text{weighted ep. unc.}} - \lambda_{al} \underbrace{\frac{1}{C} \sum_c \sigma^2_{al}(x)_c}_{\text{aleatoric unc.}} - \lambda_{ood} \underbrace{s(x)}_{\text{OoD score}} \tag{5}$$

with the (calculated) class weight vector $w = [w_1, w_2, .., w_C]^T$ (see eq. 6) and hyperparameters $\lambda_{al} \in \mathbb{R}^+$, $\lambda_{ood} \in \mathbb{R}^+$. The images with the highest scores of the acquisition function are selected for labeling in each step. Each component fulfills a specific task in the acquisition process:
1) Weighted Epistemic Uncertainty With the BNN we can calculate the informativeness of each unlabeled image measured by the epistemic uncertainty. This measure has a sound theoretical background and a proven track record in practice. The advantage of BNNs is that this approach estimates the epistemic uncertainty for each class independently. Existing approaches based on epistemic uncertainty often just take the sum over all classes. For our proposed model, we want to emphasize the informativeness of underrepresented classes. Therefore, we multiply the epistemic uncertainty of each class with a class-weight w_c which is calculated by

$$w_c = \frac{N_{train}}{N_c * C} \tag{6}$$

with N_{train} being the number of labeled images, N_c being the number of labeled images of class c and C the number of classes. The class weights are recalculated at each acquisition step, depending on the given label distribution of the current set D_{train}. This allows the algorithm to automatically adjust to class imbalances in $\mathcal{D}_{train}$.
2) Aleatoric Uncertainty The BNN measures the aleatoric uncertainty of each unlabeled image. We down-weight the informativeness of images based on their aleatoric uncertainty estimate to avoid labeling ambiguous patches. Although in the existing literature the aleatoric uncertainty is described as a measure of data uncertainty, we found that it does not capture data uncertainty for images with a different appearance (OoD). Therefore, an additional measure for the OoD images is necessary.
3) OoD Score To avoid the acquisition of outliers we apply an OoD algorithm on extracted image features. We down-weight the informativeness of images with a high OoD score. This allows the network to focus on the in-distribution data and acquire informative image patches.
The active learning procedure is summarized in Algorithm 1.
After the acquisition steps are completed, the trained models are not only able to give accurate classification predictions for each new test image, but also the epistemic and aleatoric uncertainty and OoD score which is very useful for the pathologist in the diagnostic process. In the regions where all three uncertainty measures are low, the prediction is reliable and the pathologist can trust the classification result.

3 Experiments

For the empirical validation, we use two publicly available datasets, Panda and MNIST. In the MNIST dataset, we artificially introduce ambiguities and artifacts to demonstrate the functionality of the different model components. The proposed FocAL strategy avoids ambiguities, and artifacts and outperforms other approaches. In the second experiment with the Panda dataset, we apply the model on real-world data. We perform a study about the introduced hyperparameters, analyze the uncertainty estimations of the model and compare different AL methods. compare to the following other AL strategies that were used in the recent literature:
RA [18, 20, 19]: Random Acquisition (RA) is a simple baseline method that uses a uniform distribution over the images instead of an informativeness measure.

Algorithm 1 FocAL algorithm

Input: Start training set $\mathcal{D}^0_{train}$, pool of unlabeled data $\mathcal{D}^0_{pool}$, models f^{FE}_α, f^{BNN}_β, f^{OoD}_γ, number of acquisition steps S.
Output: Optimal model parameters α, β, γ; training dataset $\mathcal{D}^S_{train}$
for $s = 0$ to S **do**
Train f^{FE}_α, f^{BNN}_β with $\mathcal{D}^s_{train}$.
Predict features $Z_{train} \leftarrow f^{\mathrm{FE}}_\alpha(X_{train})$
Update f^{OoD}_γ with Z_{train}
Predict features $Z_{pool} \leftarrow f^{\mathrm{FE}}_\alpha(X_{pool})$
Estimate unc. $\sigma^2_{ep}(X_{pool}), \sigma^2_{al}(X_{pool}) \leftarrow f^{\mathrm{BNN}}_\beta(Z_{pool})$
Estimate OoD scores $s(X_{pool}) \leftarrow f^{\mathrm{OoD}}_\gamma(Z_{pool})$ (eq. 4)
Select acq. set A^s with $a(X_{pool}, \{f^{\mathrm{FE}}_\alpha, f^{\mathrm{BNN}}_\beta, f^{\mathrm{OoD}}_\gamma\})$ (eq. 5)
Label A^s
Add A^s to labeled data $\mathcal{D}^{s+1}_{train} \leftarrow \mathcal{D}^s_{train} \cup A^s$
Remove A^s from pool $\mathcal{D}^{s+1}_{pool} \leftarrow \mathcal{D}^s_{pool} \setminus A^s$
end for
return Optimal model parameters α, β, γ; training dataset $\mathcal{D}^S_{train}$.

EN [18, 47, 19]: The maximum entropy (EN) is used for acquisition. As entropy is a measure of new information, the most informative images should be obtained.
BALD [22, 18, 20, 19]: Acquisition with Bayesian Active Learning by Disagreement (BALD). The idea of BALD is to select the images which maximize the mutual information between predictions and model posterior. This is one of the most popular methods adapted in recent literature.
MS [18]: The Mean Std (MS) measures the uncertainty by the average standard deviation of the predictive distribution. The idea is to acquire images with the least confident predictions.
EP [49]: BNN using only the epistemic uncertainty as calculated in equation 2. This method is similar to FocAL, but without weighting the epistemic uncertainty and without the aleatoric uncertainty and OoD scoring.
FocAL: The proposed FocAL method as described in Section 2.

3.1 MNIST

The goal of this experiment is to illustrate the functionality of FocAL in a controlled environment with an intuitive dataset with artificial artifacts and ambiguities.
Dataset The well-known MNIST dataset [50] contains 60,000 training and 10,000 test images of handwritten digits with 28x28 greyscale pixels. Of the original training split, we randomly sample 2000 images of which 20 images are initially labeled (D_{train}) while 1980 images remain initially unlabeled (D_{pool}). This relatively low number of images is chosen for better visualization of the data distribution (Fig. 3 and 4). In each acquisition step, 10 images are acquired (labeled) until $\mathcal{D}_{train}$ contains 200 labeled images. We also sample 200 validation images (from the training split) to use a reasonably small validation set in the context of limited labeled data [51]. For testing, we use the original test split of 10000 images. Furthermore, we adjust this dataset to mimic the problems in digital pathology that we want to tackle. The class imbalance is obtained by reducing the original 10 classes to only 3 classes: Digit '0', digit '1', and 'all other digits'. The classes '0' and '1' represent the malignant classes (10% portion of the whole dataset each) while 'all other digits' represent the healthy tissue (80% portion of the whole dataset).

Artifacts and Ambuiguities The artificial artifacts and ambiguities are obtained by adding perturbations to the input images, as depicted in Figure 2. We use three different perturbations that mimic artifacts and ambiguities in histopathological images: *Black dots* are randomly added to 75% of the total image pixels by setting the greyscale value to 0. This simulates pen marker or ink

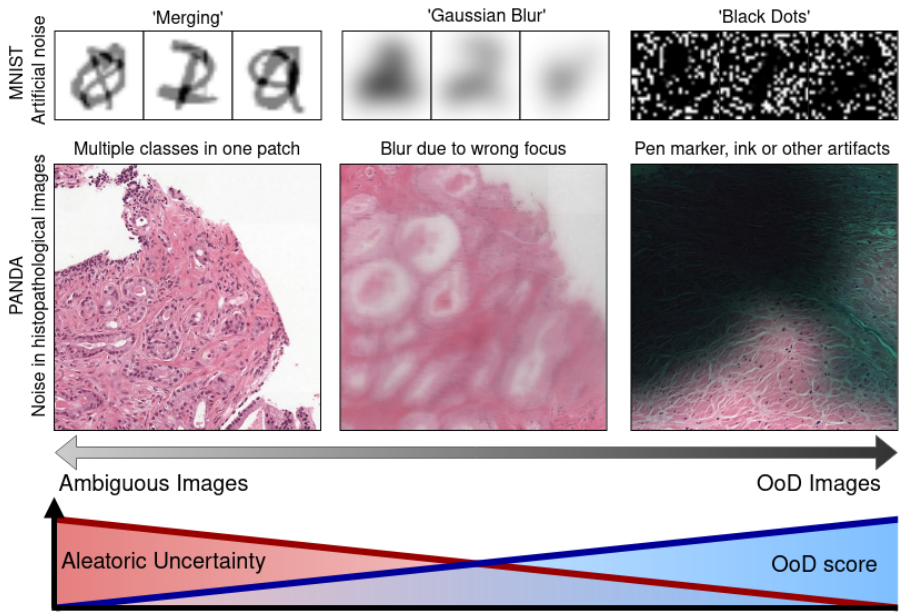


Figure 2: Images with ambiguities and artifacts that should be avoided during AL acquisition. The top row shows MNIST images with the artificial noise types 'Merging', 'Gaussian Blur', and 'Black Dots'. They simulate the artifacts and ambiguities encountered in histopathological images (bottom row) in the Panda dataset. The left Panda patch contains two different classes (Gleason Grade 3 and 4), the middle patch is blurry due to wrong microscope focus and the right patch is covered by pen marker, obscuring most tissue parts. Although a clear categorization is difficult, we propose the following scale: The images on the left side show ambiguities but the images are in-distribution because their appearance (color distribution and shapes) is normal. The images on the right side can be considered OoD because the color distribution and shapes substantially differ from the 'normal' images of interest. The blurry images are in between these two extremes as the color distribution and appearance is slightly OoD and they contain ambiguities due to blurry edges and patterns. We will see that with the proposed FocAL method, the shown images are avoided thanks to the aleatoric uncertainty and OoD score.

in histopathological images that can cover large parts of image patches. *Gaussian blur* filter with a standard deviation of $\sigma = 4$ is used to simulate the blur caused by wrong focus. *Merging* by randomly blending one image with another image of a different class together leads to ambiguous images with two different plausible labels (while maintaining the original one-hot encoded label). This simulates ambiguities by the presence of two cancerous classes in one image patch or edge cases with unclear ground truth label. Each of the three perturbations is applied to a total of 200 unlabeled images.

Implementation Details As the images of MNIST are very small, we use a simple feature extractor consisting of one convolutional layer with 4 filters (stride 3x3), max pooling (stride 2x2), and one fully connected layer with 128 units. The BNN consists of two fully connected layers with 128 units each and a final softmax layer with three output units, corresponding to the three classes. We use the cross-entropy loss and the Adam optimizer [52] for 1000 epochs before each acquisition step. The learning rate is set to 0.0001 and multiplied by 0.5 if the validation accuracy does not increase for 50 epochs. The combination of a high number of epochs and learning rate reduction assures complete convergence at each acquisition step. We experimentally set the weight factors to $\lambda_{al} = 0.5$ and $\lambda_{ood} = 2.0$ since it showed the best results (tested: 0.5, 1.0, and 2.0 for each hyperparameter). Note, that an extensive ablation study of these hyperparameters is included in Section 3.2 for the Panda dataset.

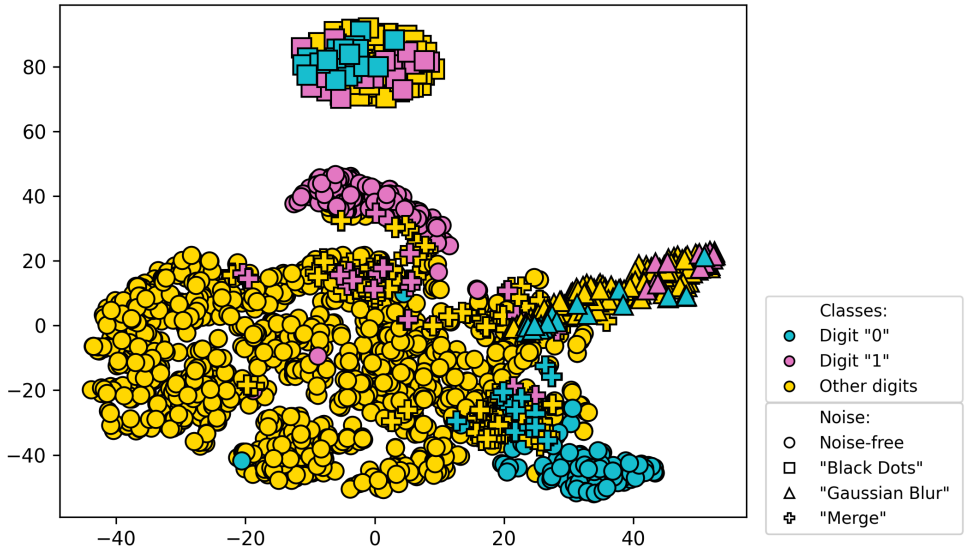
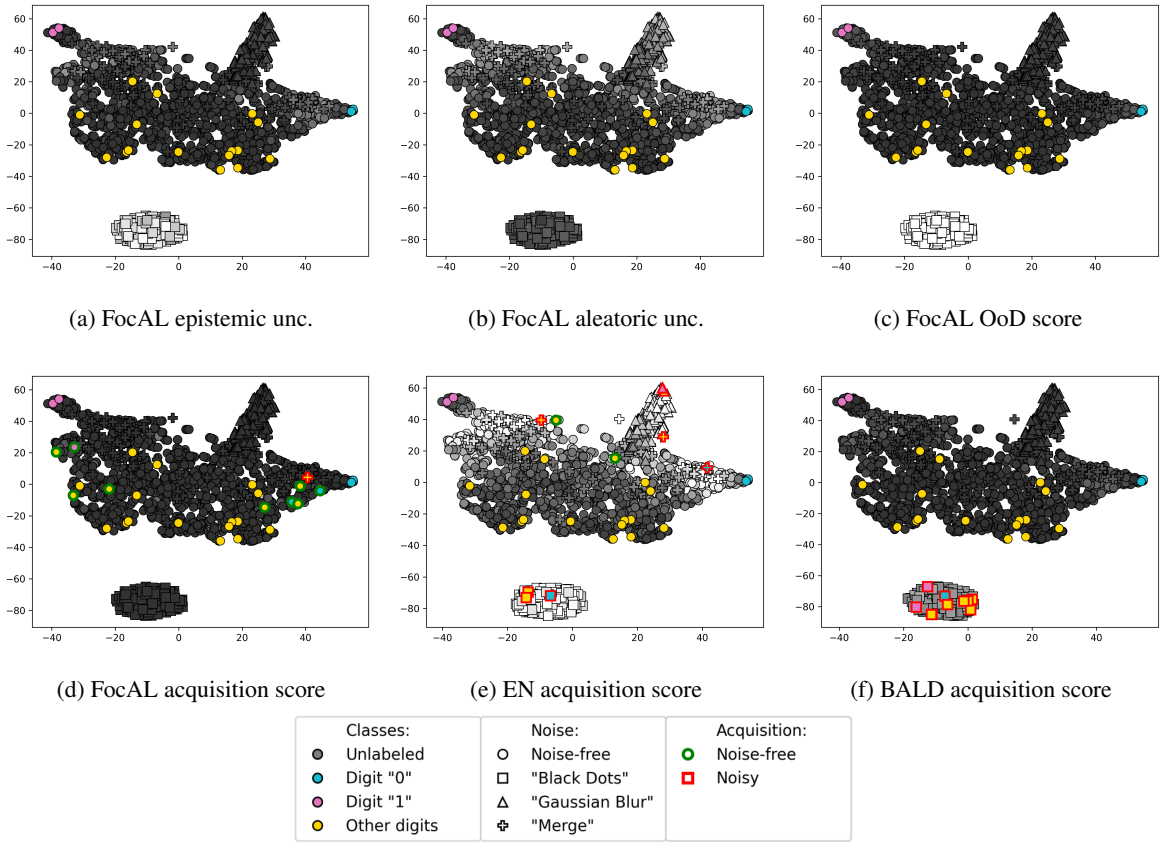


Figure 3: Feature distribution of the 2000 images (X_{train} and X_{pool}) after the last acquisition step with 200 labeled images. Each point represents the feature vector z of an MNIST image, reduced to two dimensions by t-SNE. The distribution supports our categorization of artifacts and ambiguities (Fig. 2). The images with 'black dots' (depicted as squares) are OoD while the 'merged' images are ambiguous and therefore close to the class boundaries. Blurred images show both characteristics (OoD and ambiguities) as some images are far away from the distribution of interest, while others lie close to the class boundaries.

Data Distribution First, we empirically analyze the data distribution with respect to the artificial artifacts and ambiguities. For this purpose, we plot the feature vectors z after the complete AL process using the FocAL method. We reduce the features to a two-dimensional distribution with t-SNE and depict the data in Figure 3. The distribution empirically reflects the categorization shown in Fig. 2. The images with 'black dots' are OoD because they are far away from the data distribution of interest. The images with 'Gaussian blur' are partially OoD and the 'merged' images are completely in-distribution. Similarly, ambiguities can be identified. The 'merged' images and a part of the images with 'Gaussian blur' are ambiguous and therefore close to the class boundaries. The images with 'Black dots' are not ambiguous. Apart from this data-related observation, the figure shows that the model learns to separate the classes during the active learning procedure. This class-separation is a necessary step for a good final classification.

Acquisition Figure 4 shows the acquisition behavior of FocAL and competing methods. It confirms that the FocAL model components work as expected in practice. The weighted epistemic uncertainty of the FocAL method (4a) is high (bright greyscale color) for images at the class boundaries, but also for noisy images (especially of the noise type 'Black dots' and 'Merging'). This means that it is a good measure of informativeness, but 'distracted' easily by artifacts and ambiguities. The aleatoric uncertainty (4b) captures images with 'Merging' and 'Gaussian blur' while the OoD score (4c) highlights the images with 'Black dots'. These images with ambiguities and artifacts are avoided during acquisition. As a result, the FocAL method (4d) acquires only 1 ambiguous image while 9 acquired images are informative and contain several images of the minority classes '0' and '1'. Furthermore, the acquisition analysis highlights the problems of existing AL methods. EN in Figure 4e and BALD in Figure 4f are highly 'distracted' by images with ambiguities and artifacts and do not acquire any informative data in this step. Ambiguous images are close to class bound-



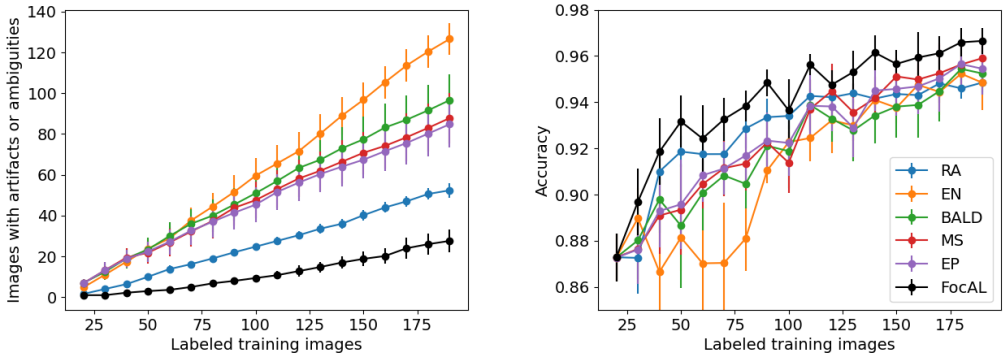
(a) FocAL epistemic unc. (b) FocAL aleatoric unc. (c) FocAL OoD score

(d) FocAL acquisition score (e) EN acquisition score (f) BALD acquisition score

Figure 4: Feature distribution, uncertainty, and acquisition scores for the first acquisition step (best viewed with zoom), similar to Figure 3. Labeled images are dots filled with turquoise (Digit ”0”), pink (Digit ”1”), or yellow (other digits). Unlabeled images are dots with greyscale color, representing the uncertainty or acquisition score (the higher the brighter). The datapoints with a green edge represent noise-free images (good for training) and datapoints with a red edge represent images with artifacts, blur or ambiguities. For the proposed FocAL method, the epistemic uncertainty (a) measures the image informativeness, but it is easily distracted by artifacts and ambiguities. These uninformative images can be captured by a high aleatoric uncertainty (b) or a high OoD score (c). Therefore, in the final FocAL acquisition (d), 9 noise-free images are acquired (and only 1 ambiguous image). The competing methods EN (e) and BALD (f) in comparison acquire almost only images with artifacts or ambiguities in this step which add less information to the training.

aries while OoD images have a ’novel’ image content due to their different appearance. Therefore, the acquisition scores of other AL methods for these images are usually high. Although here we depict only the data distribution of features from the first acquisition step, these observations are representative for all further acquisition steps as well.

Model Comparison Figure 5a confirms the previous observation (Fig. 4e and 4f) that other methods acquire many images with artifacts and ambiguities, even more than the model with random acquisition (RA). Figure 5b shows that the acquisition of many images with ambiguities and artifacts harms the test performance, as the model EN acquires the most images with ambiguities and artifacts and shows the worst performance. The other AL methods also acquire a substantial amount of images with ambiguities and artifacts and their performance is on par or even below the baseline model with random acquisition RA. Note that in other studies AL methods outperform random acquisition but many are conducted on clean, highly curated datasets. This often does not



(a) Acquired images with artifacts or ambiguities

(b) Test accuracy

Figure 5: Results of the MNIST Experiments with mean and standard error of five independent runs. In 5a the total acquired noisy images are plotted and in 5b and the accuracy. The proposed FocAL algorithm effectively avoids acquiring images with ambiguities and artifacts and shows the strongest performance.

apply to real-world data (like histopathological images). The proposed FocAL method effectively avoids acquiring uninformative images which leads to the overall best performance.

To compare to the supervised baseline, we try two different settings: one trained with all 2000 images (including the 600 images with artifacts and ambiguities) and another one with only the 1400 images without perturbations. Again, we report the average results over 5 independent runs. The supervised model trained without artifacts and ambiguities performed substantially better (accuracy 0.965; mean f1 score: 0.940) than the model trained with all 2000 images (accuracy 0.939; mean f1 score: 0.901). This confirms our hypothesis that avoiding artifacts and ambiguities is essential for a good performance. We observe that the FocAL method reaches the supervised performance (with all 2000 images) with only 90 labeled training images (4.5% of all images) and even outperforms this supervised performance due to the successful avoidance of perturbed images. Therefore, FocAL not only alleviates the labeling process - it can further save resources usually necessary for data curation. The performance of the supervised model trained on the highly curated dataset with 1400 images is reached by FocAL with 190 labeled images (9.5% of all training data).

3.2 Panda

The Panda dataset is a large open dataset for the classification of prostate cancer. We use this dataset for hyperparameter tuning, analysis of the different uncertainty measures, and finally for a comparison of the different AL methods.

Dataset The Panda dataset [7] consists of 10,616 WSIs and was presented at the MICCAI 2020 conference as a Kaggle challenge. Two institutes participated in labeling the WSIs. The images from Radboud University Medical Center come with detailed label masks of all tissue parts in the Gleason Grading (GG) scheme. The classes are 'Non-cancerous' (NC), 'Gleason 3', 'Gleason 4', and 'Gleason 5' depending on the architectural growth patterns of the tumor. The second institute, the Karolinska Institute, only assigned binary (cancer vs. healthy) labels and we therefore disregard their images for our experiments. After sorting out corrupted images, a total of 5058 WSIs are left of which we used 1000 WSIs for testing and 30 WSIs for validation that were randomly chosen. The WSIs were divided into 50% overlapping 512x512 patches. To use a multi-scale feature extractor (see Implementation paragraph below), we determine the class of each patch by its center segment (256x256 square), depending on the majority class of the pixels if at least 5% of the pixels are annotated as cancerous. If less than 5% of the pixels are annotated as cancerous, the patch is assigned the non-cancerous label. If a patch contains more than 95% of background (according to

the annotation mask), it is disregarded. Therefore, the dataset is already curated because several artifacts are already excluded in this step. Note that in other experiments, where the dataset curation is more difficult, the advantage of FocAL might be even bigger.

For the empirical validation, we design two different experiments, *small-Panda* for studying the hyperparameter setting due to computational constraints and *big-Panda* with all available images for the final experiments. In the small-Panda experiment, we take a subset of 200 WSIs for training. For the AL start, we randomly choose 20 WSIs with 5 labeled patches each, resulting in 100 labeled patches in total (equally distributed over the classes with 25 patches per class). We perform 16 acquisition steps of 50 patches each. The big-Panda setup includes all available 4028 WSIs from the Radboud center for training. To reach a competitive performance with limited computational resources, we start with 400 randomly extracted patches of 100 WSIs (100 from each class). In each acquisition step, we acquire 400 more patches. Validation and test sets are equal for both experiments, see above. The common metric to measure the classification of prostate cancer is Cohen's quadratic kappa which measures the similarity of the ground truth and the predictions, taking the class order into account (misclassifying Gleason 5 as Gleason 4 has less impact than misclassifying Gleason 5 as NC).

Implementation For the feature extraction, we use the EfficientNetB3 model [53] as a CNN backbone in a multi-scale architecture. Remember that the patch classes were assigned based on the center 256x256 square of each patch with resolution 512x512 (see dataset paragraph above). The center square (256x256) is cropped and fed into the CNN. At the same time, the complete 512x512 patch is resized to 256x256 and fed into a parallel CNN. The two feature vectors (of 1536 dimensions each) are concatenated, followed by a dropout layer and a fully connected layer with 128 units. This approach has the following two advantages. First, the WSI can be segmented with a high level of detail because the classification is performed for relatively small patch centers of 256x256. Second, the context (surrounding tissue) can still be taken into account by the model.

The complete feature extractor has fewer than 22 million trainable parameters (for comparison, a ResNet50 [54] has over 23 million parameters). The BNN consisted, as in the MNIST experiment, of two fully connected layers with 128 units and a final softmax layer with one output unit per class. We train the model with the Adam optimizer [52] for 200 epochs for each acquisition step. The learning rate is set to $1\mathrm{e}{-4}$ for the first 100 epochs, then reduced to $1\mathrm{e}{-5}$ for the other 100 epochs.

Hyperparameter Tuning First, we perform experiments regarding the newly introduced hyperparameters. We analyze λ_{al} and λ_{ood} of the FocAL acquisition function (eq. 5) that weight the importance of avoiding ambiguities and OoD images, respectively. For this purpose, we use the small-Panda setup (as described in the dataset paragraph above). Figure 6 shows the results for different hyperparameter settings. If the factors are too high, the model performance decreases, as the worst performances are given for the models with $(\lambda_{al} = 2.0,\ \lambda_{ood} = 1.0)$ and $(\lambda_{al} = 2.0,\ \lambda_{ood} = 2.0)$. We assume that in this case, the model focuses too much on the avoidance of ambiguities such that the novelty (measured by the epistemic uncertainty) is not given enough importance. Especially a high λ_{al} can harm the performance, as this measure sometimes mistakenly shows high values for informative images at the class boundary. We choose the model with $\lambda_{al} = 0.5,\ \lambda_{ood} = 1.0$ for the final experiments because it is the overall best-performing model. Note that a comparable performance was obtained for the models with $(\lambda_{al} = 0.5,\ \lambda_{ood} = 2.0)$ and $(\lambda_{al} = 1.0,\ \lambda_{ood} = 2.0)$. Overall the performance is robust for all models with $\lambda_{al} < 2$.

Uncertainty Estimations Figure 7 illustrates the uncertainty estimates in the first acquisition step of the FocAL method. It shows that each component of the acquisition function works as expected. The area with a high acquisition score (7g) is based on a high epistemic uncertainty in the circle A in Figure 7d and a low aleatoric uncertainty and OoD score in Figures 7e and 7f, respectively. Indeed, the area contains cancerous tissue and the model shows some misclassifications here (Gleason 4 instead of Gleason 3), as is clear by comparing Figures 7b and 7c. Therefore, labeling these patches can improve the overall model performance. Other parts of the image show a high aleatoric uncertainty, for example, circle B in Figure 7e. This indicates ambiguous patches and therefore, these images are avoided. The acquisition score in this area is low. In circle C in Figure 7f we see

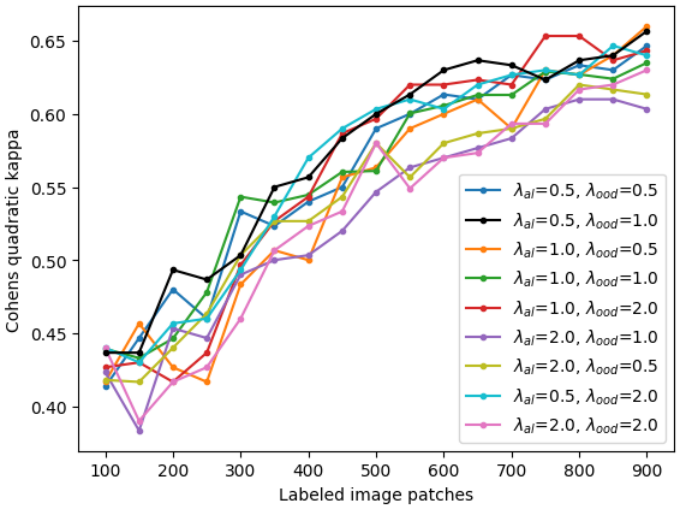
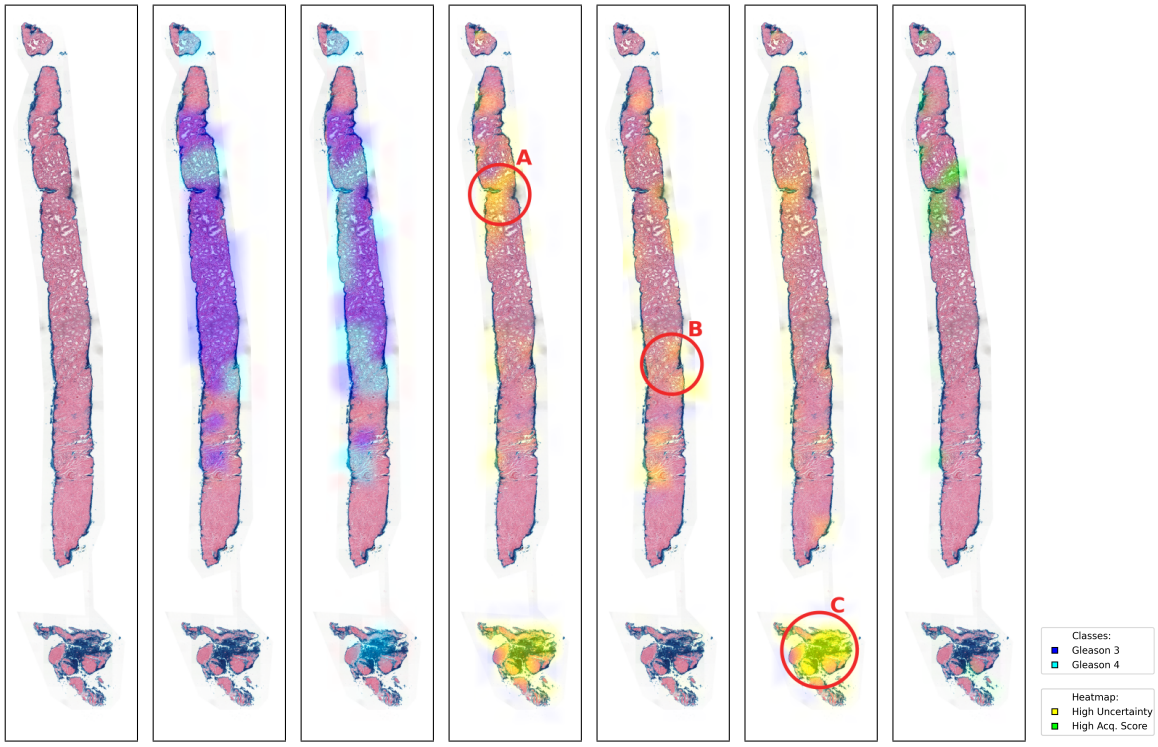


Figure 6: Hyperparameter Tuning of the FocAL method. The two hyperparameters λ_{al} and λ_{ood} are analyzed which weight the aleatoric uncertainty and OoD detection, respectively. The best performing model with $\lambda_{al} = 0.5,\ \lambda_{ood} = 1.0$ is used for the final experiments. If the weight λ_{al} is chosen too high, the performance of the model drops because the avoidance of ambiguities is given too much importance.

a region containing mainly artifacts of dark ink. These artifacts are detected by the OoD detection and therefore also avoided in the final acquisition, although the epistemic uncertainty is high in this region. As the images are produced at an early stage of the active learning process, the model's predictions (7c) are not yet accurate which is also reflected by high uncertainty values (7d - 7f).

In the diagnostic process, these uncertainties are important to identify unreliable predictions. A high epistemic uncertainty means that the model has to be further trained on this specific tissue type. A high aleatoric uncertainty indicates data ambiguities that might lead to a wrong class prediction. The OoD score shows that the indicated region contains artifacts or content that is substantially different from the learned data distribution. Low uncertainties - of all measures - indicate that the model probably made a correct classification in those areas.

Acquisition In Figure 8 we depict five patches with the highest acquisition score of the methods EN, BALD, and FocAL. They are taken from the third acquisition step but represent the general tendency that can be observed throughout the active learning process. The proposed FocAL method avoids artifacts and ambiguities and acquires representative patches containing cancerous tissue. All five patches with the highest acquisition score contain either Gleason 3 or Gleason 4 in this acquisition step. Overall, in the complete AL process (big-Panda setup), the FocAL method acquired the highest number of Gleason 5 patches. It is the most severe grade and at the same time the most underrepresented one. In total 525 patches of Gleason 5 are acquired on average in the three runs (while EN acquired 403 and all other methods below 400 each). This shows, that the class imbalance was successfully addressed. The EN acquisition assigns a high score for ambiguous patches that contain multiple different classes, like patches 8b and 8e. Both patches include glands of both Gleason 3 and Gleason 4, but for the classification task, only one label per patch is assigned, which is Gleason 4 for patch 8b and Gleason 3 for patch 8e. We argue that these ambiguities can slow down the labeling process because the pathologists take longer for the decision in comparison to the annotation of representative, non-ambiguous patches. The BALD method, which is commonly used and has shown impressive results on clean datasets [18], fails to find informative patches. All five patches with the highest acquisition score contain artifacts and none of them contains cancerous tissue. Similar observations can be made for the MS and EP methods. All three methods



(a) WSI (b) Ground tr. (c) Prediction (d) Epist. unc. (e) Aleat. unc. (f) OoD Score (g) Acquis. Sc.

Figure 7: Visualization of predictions and uncertainties for a test slide at the first acquisition step of the Panda dataset (after training on the initial 400 patches). It was split into 120 overlapping patches and all heatmaps were produced by cubic interpolation of the patch center predictions. Uncolored parts correspond to non-cancerous tissue (in 7b, 7c) and areas without (or with low) uncertainties (in 7d - 7g). The red circle A in Fig. 7d marks an area with high epistemic uncertainty. As the aleatoric uncertainty and OoD score are low in this area, this results in a high acquisition score (green in Fig. 7g). The area of circle A is informative. The red circle B in Fig. 7e shows an area with high aleatoric uncertainty due to slight blur and ambiguities. Therefore, the final acquisition score is low in this area. The red circle C in Fig. 7f shows an area with artifacts (blue ink) that results in a high OoD score. Although the epistemic uncertainty is high in this area, the acquisition of these non-cancerous and uninformative patches is avoided.

(BALD, MS and EP) are highly 'distracted' by artifacts resulting in an acquired dataset in which many patches contain little or no tissue at all.

Model Comparison The proposed FocAL algorithm shows an overall strong performance as reported in Figure 9 with a final Cohen's quadratic kappa of 0.764 with 4400 image patches corresponding to only 0.69% of all patches of the dataset. After the second acquisition step (with 1200 labeled patches and more), the result is constantly better than RA, MS, EP, and BALD, because the acquisition of artifacts and ambiguities is actively avoided. As the FocAL acquisition selects representative patches of the classes, including many images of the most severe grade (Gleason 5) which is highly underrepresented, the created dataset is of high quality. Therefore, our algorithm has successfully addressed the challenges of histopathological labeling.

The methods BALD, MS, and EP, which are acquiring the highest number of artifacts, show a weak performance in the first acquisition steps. Their performance remains significantly below the baseline of random acquisition (RA) until 2400 patches are acquired. Afterwards, their performance is comparable to RA, but not significantly better. This confirms the previous findings of the

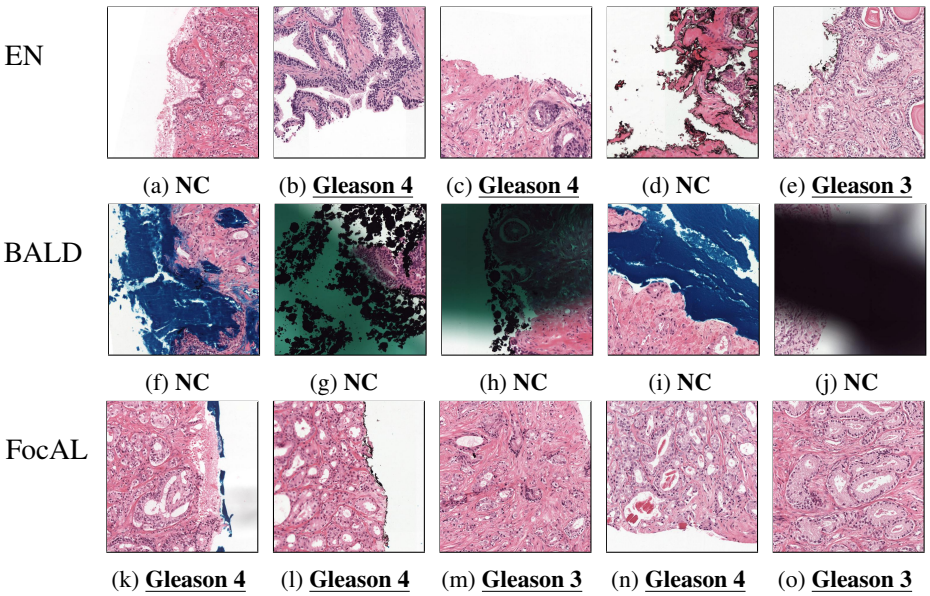


Figure 8: Acquired image patches. Each row shows the five patches with the highest acquisition score of the methods EN, BALD, and FocAL at the third acquisition step (after training with 800 labeled patches). Below each patch, we report the class and underline cancerous classes to highlight them. While the EN strategy favors ambiguous patches (like patches 8b and 8e) and BALD gets distracted by artifacts that do not contain cancerous tissue, the proposed FocAL method acquires informative patches that represent the cancerous classes well.

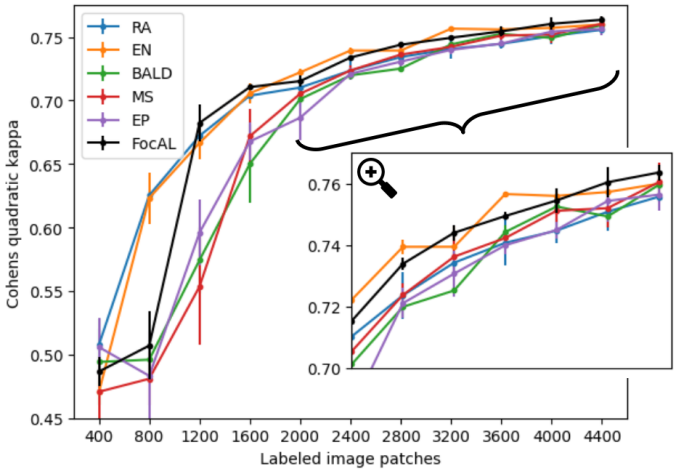


Figure 9: Model Comparison on the Panda dataset (setting big-Panda). Several existing methods (BALD, MS, and EP) are below or on par with random acquisition (RA). This is the consequence of the acquisition of artifacts in the active learning process. The EN method acquires fewer artifacts, but more ambiguous images. This seems to be less harmful to the final result but leads to a more difficult labeling process and a less representative dataset. The proposed FocAL method actively avoids artifacts and ambiguities and reaches a satisfying performance with a final Cohen's quadratic kappa of 0.764.

MNIST experiments: these commonly applied acquisition methods perform well on clean datasets but artifacts have a major impact on their performance.

The EN method is less affected by the acquisition of artifacts but acquires more ambiguous images as already seen in Figure 8 and the MNIST experiment 4. Interestingly, this seems to not have a major impact on the performance for the Panda dataset and the final Cohen's quadratic kappas values are comparable to FocAL over several acquisitions. However, we want to stress some disadvantages of the EN method. For the MNIST experiment 5, this method showed the worst performance and acquired the highest number of patches with ambiguities and artifacts. Therefore, it might not be suitable for other applications. Furthermore, the extensive acquisition of ambiguous images in the Panda dataset can lead to problems in the labeling process: the annotation by pathologists might take longer and the chances of wrong annotations are higher. Also, the final dataset might be less valuable because it mainly consists of edge cases of this specific model instead of representative patches of each class. The final Cohen's quadratic kappa value of EN is 0.759 (FocAL: 0.764). The supervised model with access to all 633.235 patch labels reaches a Cohen's quadratic kappa of 0.841 for this task.

Model Limitations Although FocAL successfully addresses the analyzed problems, there are some limitations worth discussing. We mimic the AL process with an already labeled dataset but it needs to be validated in a human study in the future. In practice, new problems but also advantages of FocAL might appear that are not notable in the experiments with an already labeled dataset. Additionally, we see some possibilities to further improve the model. The aleatoric uncertainty is a widely adopted uncertainty measure for data uncertainty and overall captures ambiguities in the data well, but it sometimes shows false-positive values assigned to images close to the class boundary, as seen in the MNIST experiment (Figure 4). Indeed, we showed in the hyperparameter tuning for Panda (Figure 6) that the weight of the aleatoric uncertainty λ_{al} must be carefully chosen. To further improve the model, a more precise uncertainty estimation for ambiguities could be a promising direction. Another limitation is that the model is currently trained from scratch for each acquisition step, as adapted from Gal *et al.*[18]. Incrementally updating the model parameters at each step can reduce the overall training time, especially in the later acquisition steps.

4 Conclusions

Our analysis of existing AL approaches for datasets with ambiguities and artifacts shows that these methods do not perform as expected. The widely used BALD algorithm, for example, acquires large amounts of images with artifacts, leading to performance that is often on par with or even below that of a random acquisition strategy. Furthermore, the resulting dataset is not a good representative of the classes of interest. Our proposed FocAL method addresses this issue by using precise uncertainty measures combined with OoD detection to avoid these ambiguities and artifacts while accounting for the class imbalance. In our experiments, we showed that each model component works as expected and that the overall results improve considerably. The acquired images are representative of the classes of interest and form a high-quality dataset. In the future, it would be interesting to analyze AL methods for other types of medical images, such as CT scans, dermatology images, or retinal images, with regard to artifacts, ambiguities, and class imbalance. It is likely that state-of-the-art methods such as BALD encounter similar problems, and the proposed FocAL method could provide a possible solution. In addition to these future applications, human studies are needed to validate the method in a real labeling process.

References

[1] Z. Zhang, P. Chen, M. McGough, F. Xing, C. Wang, M. Bui, Y. Xie, M. Sapkota, L. Cui, J. Dhillon, N. Ahmad, F. K. Khalil, S. I. Dickinson, X. Shi, F. Liu, H. Su, J. Cai, and L. Yang, "Pathologist-level interpretable whole-slide cancer diagnosis with deep learning," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 236–245, 2019.

[2] A. Hekler, J. S. Utikal, A. H. Enk, W. Solass, M. Schmitt, J. Klode, D. Schadendorf, W. Sondermann, C. Franklin, F. Bestvater, M. J. Flaig, D. Krahl, C. von Kalle, S. Fröhling, and T. J.

Brinker, "Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images," *European Journal of Cancer*, vol. 118, pp. 91–96, 2019.

[3] B. Ehteshami Bejnordi, M. Veta, P. Johannes van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J. A. W. M. van der Laak, and a. t. C. Consortium, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *JAMA*, vol. 318, no. 22, pp. 2199–2210, 2017.

[4] N. Dimitriou, O. Arandjelović, and P. D. Caie, "Deep learning for whole slide image analysis: An overview," *Frontiers in Medicine*, vol. 6, p. 264, 2019.

[5] N. Shvetsov, M. Grønnesby, E. Pedersen, K. Møllersen, L.-T. R. Busund, R. Schwienbacher, L. A. Bongo, and T. K. Kilvaer, "A pragmatic machine learning approach to quantify tumor-infiltrating lymphocytes in whole slide images," *Cancers*, vol. 14, no. 12, p. 2974, 2022.

[6] A. Schmidt, J. Silva-Rodriguez, R. Molina, and V. Naranjo, "Efficient cancer classification by coupling semi supervised and multiple instance learning," *IEEE Access*, vol. 10, pp. 9763–9773, 2022.

[7] W. Bulten, K. Kartasalo, P.-H. C. Chen, P. Ström, H. Pinckaers, K. Nagpal, Y. Cai, D. F. Steiner, H. van Boven, R. Vink, C. Hulsbergen-van de Kaa, J. van der Laak, M. B. Amin, A. J. Evans, T. van der Kwast, R. Allan, P. A. Humphrey, H. Grönberg, H. Samaratunga, B. Delahunt, T. Tsuzuki, T. Häkkinen, L. Egevad, M. Demkin, S. Dane, F. Tan, M. Valkonen, G. S. Corrado, L. Peng, C. H. Mermel, P. Ruusuvuori, G. Litjens, M. Eklund, the PANDA challenge consortium, A. Brilhante, A. Çakır, X. Farré, K. Geronatsiou, V. Molinié, G. Pereira, P. Roy, G. Saile, P. G. O. Salles, E. Schaafsma, J. Tschui, J. Billoch-Lima, E. M. Pereira, M. Zhou, S. He, S. Song, Q. Sun, H. Yoshihara, T. Yamaguchi, K. Ono, T. Shen, J. Ji, A. Roussel, K. Zhou, T. Chai, N. Weng, D. Grechka, M. V. Shugaev, R. Kiminya, V. Kovalev, D. Voynov, V. Malyshev, E. Lapo, M. Campos, N. Ota, S. Yamaoka, Y. Fujimoto, K. Yoshioka, J. Juvonen, M. Tukiainen, A. Karlsson, R. Guo, C.-L. Hsieh, I. Zubarev, H. S. T. Bukhar, W. Li, J. Li, W. Speier, C. Arnold, K. Kim, B. Bae, Y. W. Kim, H.-S. Lee, and J. Park, "Artificial intelligence for diagnosis and gleason grading of prostate cancer: the PANDA challenge," *Nature Medicine*, vol. 28, no. 1, pp. 154–163, 2022.

[8] S. Otálora, N. Marini, H. Müller, and M. Atzori, "Semi-weakly supervised learning for prostate cancer image classification with teacher-student deep convolutional networks," vol. 12446 LNCS, pp. 193–203, 2020.

[9] J. Li, W. Speier, K. C. Ho, K. V. Sarma, A. Gertych, B. S. Knudsen, and C. W. Arnold, "An EM-based semi-supervised deep learning approach for semantic segmentation of histopathological images from radical prostatectomies," *Comput. Medical Imaging Graph.*, vol. 69, pp. 125–133, 2018.

[10] N. Marini, S. Otálora, H. Müller, and M. Atzori, "Semi-supervised training of deep convolutional neural networks with heterogeneous data and few local annotations: An experiment on prostate histopathology image classification," *Medical Image Analysis*, vol. 73, p. 102165, 2021.

[11] M. Y. Lu, R. J. Chen, and F. Mahmood, "Semi-supervised breast cancer histology classification using deep multiple instance learning and contrast predictive coding (conference presentation)," in *Medical Imaging 2020: Digital Pathology*, 2020, p. 18.

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

[13] P. Chikontwe, M. Kim, S. J. Nam, H. Go, and S. H. Park, "Multiple instance learning with center embeddings for histopathology classification," in *Medical Image Computing and Computer Assisted Intervention – MICCAI*, 2020, pp. 519–528.

[14] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," in *Conference on Computer Vision and Pattern Recognition - CVPR*, 2021, pp. 14 313–14 323.

[15] A. J. Joshi, F. Porikli, and N. Papanikolopoulos, "Multi-class active learning for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2372–2379.

[16] X. Li and Y. Guo, "Adaptive active learning for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 859–866.

[17] X. Zhu, J. Lafferty, and Z. Ghahramani, "Combining active learning and semi-supervised learning using gaussian fields and harmonic functions," in *International Conference on Machine Learning - ICML*, 2003, pp. 58–65.

[18] Y. Gal, R. Islam, and Z. Ghahramani, "Deep bayesian active learning with image data," *International Conference on Machine Learning - ICML*, pp. 1183–1192, 2017.

[19] L. Raczkowski, M. Możejko, J. Zambonelli, and E. Szczurek, "ARA: accurate, reliable and active histopathological image classification framework with bayesian deep learning," *Scientific Reports*, vol. 9, no. 1, p. 14347, 2019.

[20] J. Carse and S. McKenna, "Active learning for patch-based digital pathology using convolutional neural networks to reduce annotation costs," in *Digital Pathology*, 2019, vol. 11435, pp. 20–27.

[21] A. L. Meirelles, T. Kurc, J. Saltz, and G. Teodoro, "Effective active learning in digital pathology: A case study in tumor infiltrating lymphocytes," *Computer Methods and Programs in Biomedicine*, vol. 220, p. 106828, 2022.

[22] N. Houlsby, F. Huszar, Z. Ghahramani, and M. Lengyel, "Bayesian active learning for classification and preference learning," *CoRR*, vol. abs/1112.5745, 2011. [Online]. Available: http://arxiv.org/abs/1112.5745

[23] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[24] V. B. Alex Kendall and R. Cipolla, "Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," in *British Machine Vision Conference - BMVC*, 2017, pp. 57.1–57.12.

[25] S. Lee, M. Amgad, P. Mobadersany, M. McCormick, B. P. Pollack, H. Elfandy, H. Hussein, D. A. Gutman, and L. A. Cooper, "Interactive classification of whole-slide imaging data for cancer researchers," *Cancer Research*, vol. 81, no. 4, pp. 1171–1177, 2021.

[26] R. Marée, L. Rollus, B. Stévens, R. Hoyoux, G. Louppe, R. Vandaele, J.-M. Begon, P. Kainz, P. Geurts, and L. Wehenkel, "Collaborative analysis of multi-gigapixel imaging data using cytomine," *Bioinformatics*, vol. 32, no. 9, pp. 1395–1401, 2016.

[27] Y. Kwon, J.-H. Won, B. J. Kim, and M. C. Paik, "Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation," *Computational Statistics & Data Analysis*, vol. 142, p. 106816, 2020.

[28] A. Der Kiureghian and O. Ditlevsen, "Aleatory or epistemic? does it matter?" *Structural safety*, vol. 31, no. 2, pp. 105–112, 2009, publisher: Elsevier.

[29] T. Xiao, A. Gomez, and Y. Gal, "Wat heb je gezegd? detecting out-of-distribution translations with variational transformers," 2019.

[30] J. Mukhoti, A. Kirsch, J. van Amersfoort, P. H. S. Torr, and Y. Gal, "Deep deterministic uncertainty: A simple baseline," 2021. [Online]. Available: 10.48550/ARXIV.2102.11582

[31] A. T. Nguyen, F. Lu, G. L. Munoz, E. Raff, C. Nicholas, and J. Holt, "Out of distribution data detection using dropout bayesian neural networks," 2022. [Online]. Available: 10.48550/ARXIV.2202.08985

[32] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, "Deep label distribution learning with label ambiguity," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2825–2838, 2017.

[33] M. Bernhardt, D. Coelho de Castro, R. Tanno, A. Schwaighofer, K. Tezcan, M. Monteiro, S. Bannur, M. Lungren, A. Nori, B. Glocker, J. Alvarez-Valle, and O. Oktay, "Active label cleaning for improved dataset quality under resource constraints," *Nature Communications*, vol. 13, 2022.

[34] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Neural Information Processing Systems - NeurIPS*, vol. 30, 2017.

[35] Y. Sun, Y. Ming, X. Zhu, and Y. Li, "Out-of-distribution detection with deep nearest neighbors," in *International Conference on Machine Learning - ICML*, 2022.

[36] N. Kanwal, F. Pérez-Bueno, A. Schmidt, R. Molina, and K. Engan, "The devil is in the details: Whole slide image acquisition and processing for artifacts detection, color variation, and data augmentation: A review," *IEEE Access*, vol. 10, 2022.

[37] T. Johnson, I. Kwok, and R. T. Ng, "Fast computation of 2-dimensional depth contours." in *International Conference on Knowledge Discovery and Data Mining*, 1998, pp. 224–228.

[38] I. Ruts and P. J. Rousseeuw, "Computing depth contours of bivariate point clouds," *Computational Statistics & Data Analysis*, vol. 23, no. 1, pp. 153–168, 1996.

[39] E. M. Knorr and R. T. Ng, "Algorithms for mining distance-based outliers in large datasets." in *International Conference on Very Large Databases - VLDB*, A. Gupta, O. Shmueli, and J. Widom, Eds., 1998, pp. 392–403.

[40] ——, "Finding intensional knowledge of distance-based outliers." in *International Conference on Very Large Databases - VLDB*, M. P. Atkinson, M. E. Orlowska, P. Valduriez, S. B. Zdonik, and M. L. Brodie, Eds., 1999, pp. 211–222.

[41] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, "Latent space autoregression for novelty detection," in *Conference on Computer Vision and Pattern Recognition - CVPR*. IEEE Computer Society, 2019, pp. 481–490.

[42] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *International Conference on Neural Information Processing Systems - NeurIPS*, 2018, pp. 7167–7177.

[43] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *ACM SIGMOD International Conference on Management of Data*, ser. Association for Computing Machinery, 2000, pp. 93–104.

[44] Y. Wu, A. Schmidt, E. Hernández-Sánchez, R. Molina, and A. K. Katsaggelos, "Combining attention-based multiple instance learning and gaussian processes for CT hemorrhage detection," in *Medical Image Computing and Computer Assisted Intervention – MICCAI*, 2021, vol. 12902, pp. 582–591.

[45] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[46] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *Journal of the American Medical Association - AMA*, vol. 316, no. 22, pp. 2402–2410, 2016-12.

[47] J. Zeng, A. Lesnikowski, and J. M. Alvarez, "The relevance of bayesian layer positioning to model uncertainty in deep bayesian active learning," *Neural Information Processing Systems - NeurIPS*, 2018.

[48] D. P. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," in *International Conference on Neural Information Processing Systems - NIPS*, 2015, pp. 2575–2583.

[49] V.-L. Nguyen, S. Destercke, and E. Hüllermeier, "Epistemic uncertainty sampling," in *International Conference on Discovery Science*, P. Kralj Novak, T. Šmuc, and S. Džeroski, Eds., 2019, pp. 72–86.

[50] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[51] A. Oliver, A. Odena, C. Raffel, E. D. Cubuk, and I. J. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," in *International Conference on Neural Information Processing Systems - NIPS*. Curran Associates Inc., 2018, pp. 3239–3250.

[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2015.

[53] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning - ICML*, vol. 97, 2019, pp. 6105–6114.

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Conference on Computer Vision and Pattern Recognition - CVPR*, 2016, pp. 770–778.

Chapter 6

Probabilistic Generative Segmentation to Capture Crowdsourcing Labels

6.1 Publication details

Authors: Arne Schmidt, Pablo Morales-Álvarez, Rafael Molina

Title: Probabilistic Modeling of Inter- and Intra-observer Variability in Medical Image Segmentation

Reference:

Status: Published

Quality indices:

- GGS Rating (2021): A++
- GGS Class (2021): 1
- CORE: A++

6.2 Main contributions

- The different diagnostic assessment of histopathological images leads to a high inter- and intra-observer variability and imposes a huge challenge for deep learning models. We propose the Probabilistic Inter-Observer and iNtra-Observer variation NetwOrk (Pionono) to model this variability in a principled way.
- The model learns a probability distribution for each annotator which explicitly captures the labeling uncertainty. Therefore, the model can simulate each annotator as well as "gold predictions", reflecting the experts' agreement.
- In extensive experiments of two public prostate cancer datasets we demonstrate that the model provides accurate segmentations, uncertainty predictions, and annotator simulations. It outperforms baselines and other state-of-the-art approaches.

PROBABILISTIC MODELING OF INTER- AND INTRA-OBSERVER VARIABILITY IN MEDICAL IMAGE SEGMENTATION

Arne Schmidt
Department of Computer Science and AI
University of Granada
Granada, Spain

Pablo Morales-Álvarez
Department of Statistics and Operations Research
University of Granada
Granada, Spain

Rafael Molina
Department of Computer Science and AI
University of Granada
Granada, Spain
*

ABSTRACT

Medical image segmentation is a challenging task, particularly due to inter- and intra-observer variability, even between medical experts. In this paper, we propose a novel model, called Probabilistic Inter-Observer and iNtra-Observer variation NetwOrk (Pionono). It captures the labeling behavior of each rater with a multidimensional probability distribution and integrates this information with the feature maps of the image to produce probabilistic segmentation predictions. The model is optimized by variational inference and can be trained end-to-end. It outperforms state-of-the-art models such as STAPLE, Probabilistic U-Net, and models based on confusion matrices. Additionally, Pionono predicts multiple coherent segmentation maps that mimic the rater's expert opinion, which provides additional valuable information for the diagnostic process. Experiments on real-world cancer segmentation datasets demonstrate the high accuracy and efficiency of Pionono, making it a powerful tool for medical image analysis.

Keywords Probabilistic Deep Learning · Medical Images · Semantic Segmentation · Probabilistic Generative Models

1 Introduction

Artificial Intelligence (AI) algorithms have shown remarkable progress in image analysis, holding great promise for faster and more accurate diagnostic procedures [1, 2, 3, 4]. Nevertheless, in medical practice, there exists a high degree of variability among the opinions of different medical experts, even when the same expert assesses the same data at different times. This inter- and intra-observer variability has been reported across various tasks, including MRI-based segmentation of

*This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska Curie grant agreement No 860627 (CLARIFY Project), from the Spanish Ministry of Science and Innovation under project PID2019-105142RB-C22, and by FEDER/Junta de Andalucía-Consejería de Transformación Económica, Industria, Conocimiento y Universidades under the project P20_00286.

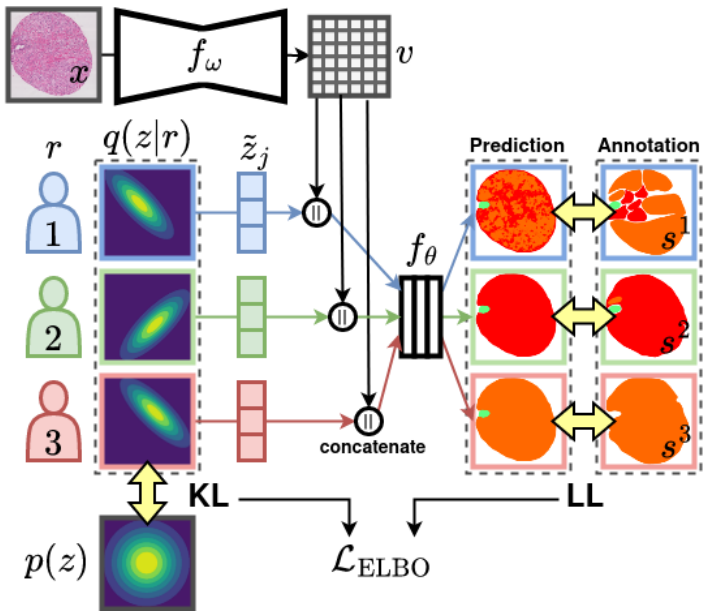


Figure 1: The proposed Pionono model. The labeling behaviour of each rater r is represented by a multivariate Gaussian distribution $q(z|r)$. The drawn samples $\tilde{z}_j$ are concatenated with the extracted features v of f_ω and then fed into the segmentation head f_θ. The output simulates the inter- and intra-observer variability of annotations and is optimized using the real annotations s^r of each rater. The model is trained end-to-end with a combination of log-likelihood loss (LL) and Kulback Leibler (KL) divergence between posterior and prior, combined in the overall loss $\mathcal{L}_{\text{ELBO}}$.

HCC lesions [5], lung cancer segmentation in CT scans [6], and multiple fields in pathology [7, 8, 9, 10]. It leads to uncertainties when applying AI models because in contrast to other classification tasks, there is not a single ground truth.

Especially in the medical domain, the careful modeling of uncertainties in its different forms has a high priority to minimize the risk of relying on incorrect predictions [11, 3, 12, 1, 13]. In recent years, probabilistic methods, such as Bayesian Neural Networks [14] and sparse Gaussian processes [13, 15] have gained more and more attention, because they are able to account for uncertainties in a sound manner. They showed promising results when modeling uncertainty in the network weights [14], data ambiguities[12] or attention weights [3]. Although inter- and intra-observer variability is often mentioned as a key challenge when applying AI to medical data [7, 16, 17, 18], to the best of our knowledge there is no method that explicitly models these two types of uncertainty for medical image segmentation.

To address this gap, we propose a novel approach called the Probabilistic Inter-Observer and iNtra-Observer variation NetwOrk (Pionono), depicted in Figure 1. This model accurately accounts for inter- and intra-observer variability using probabilistic deep learning. Specifically, each rater's labeling behavior is represented as a probability distribution in latent space, and optimized using the log Evidence Lower BOund (ELBO) in an end-to-end training process. The variance of each rater's distribution models the intra-observer variability, while the differences between the distributions models the inter-observer variability. When two raters exhibit similar labeling behavior, their probability distributions overlap substantially, while different labeling behavior results in a small overlap of distributions.

The approach is validated in extensive experiments of prostate and breast cancer segmentation, using 'gold' labels. They reflect the expert agreement to show that our probabilistic modeling improves the predictive performance and estimates the predictive uncertainty. Furthermore, we also test its capability to model each rater's labeling behavior. As shown in the experiments, it

Method	(i) Prob. Uncert.	(ii) Coh. Segm.	(iii) Exp. Opinion	(iv) Scale
STAPLE	✗	✗	✗	✓
Prob U-Net	✓	✓	✗	✓
CM global	✗	✗	✓	✓
CM pixel	✗	✗	✓	✗
Pionono	✓	✓	✓	✓

Table 1: For AI segmentation models to achieve the best possible diagnostic support, they should address four key issues: (i) provide a *probabilistic uncertainty* estimation, not only a single prediction for a test image; (ii) provide multiple *coherent segmentation* hypotheses; (iii) simulate different *expert opinions* for better explainability and decision support; (iv) *scale* to a higher amount of raters in the case that more data from different hospitals can be integrated.

can simulate expert opinions for a given test image in a consistent manner, providing a realistic estimation of "what expert X would say in this case". Our contributions can be summarized as follows:

- We propose Pionono, a probabilistic deep learning model that uses probability distributions in latent space to represent inter- and intra-observer variability. It can be trained with labels of multiple raters.
- The model is able to provide accurate segmentation predictions (compared to the expert agreement and different expert opinions), outperforming existing state-of-the-art algorithms such as STAPLE, Probabilistic U-Net and models based on global or local confusion matrices.
- Pionono provides uncertainty estimations that indicate areas where the predictions are not conclusive.
- The proposed model can provide several coherent segmentation hypotheses, simulating different medical experts.

2 Related Work

In this section, we review existing methods of probabilistic deep learning and crowdsourcing for medical images and highlight the differences to our model.

Probabilistic Deep Learning. As already indicated, probabilistic approaches such as Bayesian neural networks [14, 19, 12, 11] and sparse Gaussian processes [13, 15, 3] have shown promising results in a multitude of tasks in the medical image domain, modeling different sources of uncertainties. Often, a general predictive uncertainty is addressed using probabilistic weight parameters [19]. This uncertainty can be bisected into model and data uncertainty which originate from model parameters or data ambiguities, respectively [12]. Other approaches have modeled the uncertainty of missing instance labels in multiple instance learning [20, 3] or uncertainty of out of distribution samples [11]. The uncertainty in annotations has previously been addressed by the Probabilistic U-Net [1] (*Prob U-Net*), which encodes the labeling behavior in a latent random variable. The model is trained as a variational autoencoder with an encoder network predicting the latent distribution. This approach models a general variability in annotations but lacks the explicit modeling of inter- and intra-observer variability. Therefore, it is not able to incorporate the rater information during training and cannot simulate expert opinions.

Crowdsourcing. While existing crowdsourcing methods aim to capture inter-observer variability in the training labels, this variability is often not reflected in the test predictions by probabilistic outputs [1]. The intra-observer variability is often not modeled at all, although it is often mentioned as a challenge in literature [7, 16, 17, 8, 5].

One way to handle multiple annotations is label fusion. With this method, the annotations of different raters are merged to a single set of labels. The "Simultaneous Truth and Performance Level Estimation" (*STAPLE*) mechanism performs label fusion with a probabilistic estimate of the true labels by weighting each segmentation depending upon the estimated performance level of each rater [21]. A supervised network can then be trained on these fused labels. More dedicated approaches incorporate different rater labels using confusion matrices (CM), for example for classification of image patches with Gaussian processes [18] or image segmentation with global confusion matrices [22] (*CM global*). In this direction, also pixel-wise confusion matrices were explored for semantic segmentation that are estimated by a dedicated deep neural network [23] (*CM pixel*). These models have shown promising results, but come with a conceptual problem: They assume that the pixels are statistically independent of each other, although neighboring pixels have a high correlation. Therefore, the output of different segmentation hypotheses is not coherent. Furthermore, the predictions of the mentioned approaches [22, 23] are not modeled by a predictive distribution, but by a deterministic point estimate. While the global confusion matrix approach [22] has a limited expressiveness, the pixel-wise calculation [23] is hard to scale for multiple raters, because for each rater, a complete deep neural network must be trained and stored.

Pionono unites the advantages of probabilistic and crowdsourcing methods. We summarize a comparison of different characteristics of Pionono and related methods for image segmentation in Table 1.

3 Methods

In this section, we outline the background of the proposed method. It is implemented in the Pytorch [24] framework and is publicly available at `https://github.com/arneschmidt/pionono_segmentation`.

3.1 Problem Definition

Let $X = \{x_i \in \mathbb{R}^{H\times W\times 3}\}_{i=1,..,N}$ be a set of images, $S^r = \{s_i^r \in \mathbb{R}^{H\times W\times C}\}_{i=1,..,N}$ the corresponding segmentation maps with image dimensions $H \times W$ and C the number of classes. The segmentation maps are provided by different raters $r \in R = \{1, 2, .., M\}$. Some or all images can be segmented by multiple raters, such that some segmentation maps s_i^r can be empty. The proposed model does not require any overlap of the sets of annotated images.

If there are images available with segmentations assigned by expert agreements (so-called gold labels), the model should be able to predict a gold distribution over outputs $p(S^{\text{gold}})$ with the mean estimating the segmentation and the variance estimating the uncertainty. In any case, the model should model different segmentation hypotheses for the raters $\{p(S^r); r = 1, 2, .., M\}$ for diagnostic decision support.

3.2 Proposed Model

First, we introduce the common segmentation backbone f_ω with trainable weights ω. We use the well-known U-Net architecture [25] with a Resnet34 feature extractor [26]. This model takes an image x_i and extracts a feature map $v_i \in \mathbb{R}^{H\times W\times L}$ with $H \times W$ being the image resolution and L the dimensions of the feature vectors ($L = 16$ in the case of U-Net). We denote the feature extraction as

$$v_i = f_\omega(x_i). \tag{1}$$

Based on these feature vectors, we could perform segmentation with a segmentation head f_θ:

$$s_i = f_\theta(v_i). \tag{2}$$

Now we extend this model to incorporate the *inter- and intra-observer variation*. The segmentation maps are influenced by the rater's experience, assessment, and personal choices. To encode the

labeling behavior, we use a random vector $z \in \mathbb{R}^D$. In practice, $D = 8$ are enough dimensions to reflect different labeling behaviors. We define a prior distribution $p(z) = \mathcal{N}(z|0, \sigma_{\text{prior}} * I)$ which encodes a generic labeling behavior without further information about the rater. It is possible to encode prior knowledge in this distribution, but we present a general model and leave this for future work. We set $\sigma^2_{\text{prior}} = 2.0$, because we observe a realistic variability in the output for this value. In section 4.5 we prove that the model is robust for different settings of hyperparameters D and σ^2_{prior}.

Now, the posterior distribution of $p(z|r)$ that depends on the rater r should be found. We approximate it with one multivariate Gaussian distribution for each rater:

$$q(z|r) = \mathcal{N}(z|\mu^r, \Sigma^r) \; \forall r = 1, .., M \tag{3}$$

where $\{\mu^r, \Sigma^r\}^{r=1,..,M}$ are trainable parameters. The variance of each distribution $q(z|r)$ models the $intra - observer$ variability. The differences between the distributions for different raters model the $inter - observer$ variability. To obtain the predictive gold distribution we add another 'rater' $r = M+1$ represented by an additional gold distribution q which is trained with the available gold segmentations. During prediction, this distribution provides the estimated agreement between experts.

The segmentation head f_θ, parametrized by weights θ, must be adapted to take the random vector z into account. The approximated predictive distribution is then obtained by:

$$q(s_i|x_i, r, \omega, \theta) = \int f_\theta(v_i, z) q(z|r) dz. \tag{4}$$

The closed-form calculation is not feasible and therefore we approximate it by Monte Carlo (MC) sampling:

$$\tilde{s}_{i,j}|x_i, r = f_\theta(v_i, \tilde{z}_j); \; \tilde{z}_j \sim q(z|r) \tag{5}$$

with $j = 1, ..., K$ indexing the MC samples. In practice, we concatenate the feature maps v_i and the latent vector $\tilde{z}_j$, which is broadcasted to the image size, leading to a feature map with dimensions $H \times W \times (L + D)$. The segmentation head consists of three layers with 1x1 convolutions and 16 filters in the first two layers and C filters in the last layer.

3.3 Training

First, all posterior distributions $q(z|r)$ are initialized randomly. Each initial value of the mean vectors μ^r is independently drawn from a distribution $\mathcal{N}(0, \sigma^2_{\text{post}})$. We set $\sigma^2_{\text{post}} = 8$, because this initializes the mean vectors sufficiently different for a good optimization. In section 4.5 we show, that the model is robust to other settings of this value. The covariance matrices Σ^r are initialized with $\sigma_{\text{prior}} * I$.

To optimize the parameters $\{\mu^r, \Sigma^r\}^{r=1,..,M}$ of the probability distribution $q(z|r)$, we maximize the ELBO:

$$\mathcal{L}_{ELBO} = \mathbb{E}_q log \; p(S^r|X, r, \omega, \theta) - \lambda KL(q(Z|r)|p(Z)). \tag{6}$$

with distribution q as defined in eq. 4. The first term defines a log-likelihood (LL) loss, making the model fit to the annotations of each rater. The second term defines the KL-divergence between the posterior distribution $q(Z|r)$ and the prior $p(Z)$ and works as a regularization of the latent distributions. The factor λ weights the regularization term and is set to 0.0005 to balance the magnitudes of the log-likelihood and the KL (we will check the robustness of this hyperparameter in Section 4.5). While the KL term can be optimized analytically, the log likelihood term must be approximated. We use the reparametrization trick [27] to split each probabilistic sample $\tilde{z}^r$ into its probabilistic component and deterministic parameters μ^r and Σ^r. These parameters can be optimized by backpropagation of gradients, together with the CNN parameters ω and θ. For numerical stability, we train the covariance matrix parameters by using the lower triangular matrix L of the Cholesky decomposition $\Sigma^r = L^r L^{r\top}$. The log-likelihood can be optimized with standard

methods like the categorical cross-entropy. We found that the general dice loss [28] leads to better results, so all final results are reported with this loss.

We use the Adam optimizer [29] for 100 epochs with a learning rate of 0.0001. The model parameters μ^r, Σ^r are optimized with a higher learning rate of $\nu = 0.02$, because else the gradient was not strong enough to properly learn the rater distributions. We tested $\nu = 0.01, 0.02, 0.04$ and include the results in section 4.5. Both learning rates are decreased after 40 epochs by dividing them by 1.1 in each epoch.

3.4 Predicting

For a test image x^*, the predictive gold distribution can again be obtained by drawing Monte-Carlo samples

$$\tilde{s}_j^* | x^*, r = f_\theta(v^*, \tilde{z}_j); \tilde{z}_j \sim q(z|r = M + 1) \tag{7}$$

with $j = 1, .., K$ indexing the MC samples and $q(z|r = M + 1)$ representing the gold distribution as described in section 3.2. The **mean** of these samples provides the segmentation hypothesis that approximates the expert agreements. The **variance** of the samples indicates uncertainties in the prediction.

Furthermore, the model is able to simulate *intra-observer variations* of rater r' by drawing multiple samples of the distribution $\tilde{z}'_j \sim q(z|r = r')$ for the final prediction. The *inter-observer variations* between rater r' and r'' can be simulated by using samples $\tilde{z}'_j \sim q(z|r = r')$ and $\tilde{z}''_k \sim q(z|r = r'')$ and finally taking the mean of both output distributions.

The model can therefore simulate **expert opinions** for a given test image. Other AI methods typically aggregate the expertise provided by all annotators to make predictions (e.g., using STAPLE, Prob U-Net). However, in such approaches, the knowledge of highly specialized experts can be diluted or lost among the less experienced annotators' knowledge. In our framework, we provide consistent predictions for each individual expert, thereby preserving their unique expertise and contributions.

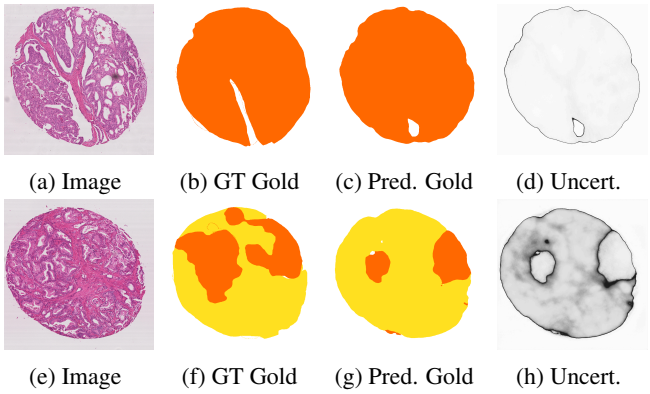
4 Experiments

In several experiments we demonstrate that the uncertainty estimation of the model indicates areas of false predictions (4.2), the model is able to capture the inter and intra-observer variations (4.3) and outperforms other related methods (4.4). Additionally, we analyze the robustness to hyperparameters (4.5), required resources (4.6), and limitations (4.7).

4.1 Datasets

For empirical validation, three public histopathological datasets were used. The first dataset, "Gleason 2019" [10] was published as a MICCAI grand challenge for pathology and includes 333 Tissue Micro Arrays (TMA) of prostate cancer, labeled by 6 different pathologists. The TMAs were scanned with a magnification of 40x and have a size of approximately 4000×4000 pixels. Of the 333 images, 244 are publicly available with labels (the test annotations of the challenge are not available). Each pathologist annotated between 61 and 241 TMAs with segmentation masks and the gold labels were obtained using the STAPLE algorithm [21], following the original work of the dataset [10]. We resize all images to 1024×1024 pixels and create 4 cross-validation splits.

The second dataset, which we will refer to as "Arvaniti TMA" was published in 2018 [30] and includes a total of 886 TMAs of prostate cancer of which 245 images were annotated by two pathologists (while the other images only have annotations of one pathologist and are therefore discarded in our study). The TMAs were scanned with a magnification of 40x but the scanned area is smaller than for the Gleason19 dataset. The images have a resolution of 3100×3100 pixels and



(a) Image (b) GT Gold (c) Pred. Gold (d) Uncert.

(e) Image (f) GT Gold (g) Pred. Gold (h) Uncert.

Figure 2: Gold prediction and uncertainty of the Pionono model. The first row shows a confident prediction as the uncertainty in 2d is low (white) for almost all the area. Indeed, the segmentation prediction 2c is very accurate, see the ground truth (GT) 2b. The second row shows an example of an uncertain prediction. Some parts of the area classified as G3 (yellow) in 2g are labeled as G4 in the ground truth 2f. These areas are estimated with a high uncertainty (black) in 2h, warning that these predictions are unreliable.

we resize them to 512×512 such that the magnification matches the resized images of the Gleason 2019 dataset. Again, we split the dataset into 4 cross-validation splits for the experimental setup.

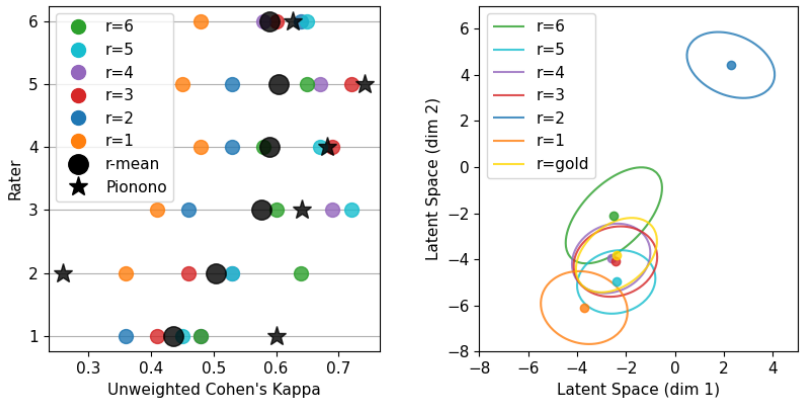
For the classification of prostate cancer, the tissue is segmented in the Gleason Grading (GG) scheme. The classes are 'Non-cancerous' (NC), 'Gleason 3' (G3), 'Gleason 4' (G4), and 'Gleason 5' (G5) depending on the architectural growth patterns of the tumor [17, 31]. To visualize the segmentations we use the colors: green for NC, yellow for G3, orange for G4, and red for G5. For the evaluation of algorithms for prostate cancer classification, previous works used the Cohen's kappa coefficient [10, 30, 17] which measures the agreement of two raters or a rater and an AI model. To compare to previously reported results for the two datasets, we use the unweighted Cohen's kappa κ for the Gleason 2019 dataset [10] and the quadratic weighted Cohen's kappa κ for the Arvaniti TMA dataset [30]. The main difference is that the quadratic kappa takes the class order into account and weighs the errors based on the quadratic distance of the predicted and the real class.

The third dataset contains 151 WSIs for breast cancer segmentation that were sliced into 11,836 patches of 512x512 pixels annotated by 25 raters [4, 32]. We will refer to this dataset as "bc segmentation". The tissue was segmented into "tumor", "inflammation", "necrosis", "stroma", and "other". Here, the gold labels were obtained by an actual discussion of experts. We use the predefined train/validation/test splits [32].

For all datasets we use image augmentation with the albumentations library [33] by applying random flip, rotation, shear, zoom, blur, and shifts in brightness, contrast, hue, and saturation. This leads to a broad range of realistic transformations of the image to avoid overfitting.

4.2 Uncertainty estimation

The proposed model provides probabilistic predictions that allow an accurate assessment of the predictive uncertainty. Fig. 2 shows the model predictions and uncertainties obtained with the gold distribution as described in section 3.2. For the first image (2a), the prediction of the model (2c) is accurate and matches the real gold annotation (2b) very well. The uncertainty (2d) for this predictions is low (white), which means that there is a low risk of a wrong prediction. Therefore, the model correctly indicates that this prediction is reliable. For the second image 2e, some areas that are predicted as G3 (yellow) in (2g) are actually G4 in the ground truth gold prediction (2f). The model's uncertainty estimation indicates that this prediction is not reliable: the misclassified areas are marked with a high uncertainty (dark) in the image (2h). Therefore, the probabilistic



(a) (b)

Figure 3: Analyzing the labeling behaviour. In Fig. (a) the agreement of each rater with all other raters is depicted, measured by the unweighted Cohen's Kappa of the true labels [10]. The mean agreement of each rater with all other raters is represented by a black dot and the agreement with Pionons test predictions, simulating the corresponding rater, by a star. This confirms that the model accurately models each rater, reaching even a higher agreement than the other raters average, except for rater 2. Fig. (b) shows the first two dimensions of the posterior distributions $q(z|r)$ with mean and covariance after training. The distributions of raters $3, 4, 5$ and 6 overlap significantly with the gold distribution and with each other, indicating a similar labeling behaviour. Indeed, these raters show the highest labeling agreement of true labels, as observed in (a).

output adds valuable information to the diagnostic process. It estimates if a prediction is reliable - or unreliable and should be double-checked.

4.3 Inter- and Intra-observer Variation

The Pionono model is able to capture the inter- and intra-observer variability. This accurate probabilistic modeling of the annotations does not only improve the predictive results (see section 4.4), but also allows to simulate specific experts at test time. In this section, we empirically show that the model learns the different label behaviors of the raters and is able to reproduce them.

In Fig. 3a we plot the *inter-observer variations* between the raters. The figure shows that there is indeed a high variability among the raters, with a Cohen's kappa ranging from 0.36 to 0.72. The simulated test predictions by Pionono show a higher agreement with each rater than the average agreement of the other raters, except for rater 2. For two raters (1 and 5), the simulated predictions of Pionono are even more than 15 percentage points higher than the average rater agreement. We also measured the IoU metric, which was $0.574, 0.540, 0.619, 0.649, 0.692, 0.507$, for the 6 raters respectively, compared to a mean inter-pathologist IoU of 0.361. The results confirm that most raters are modeled with high accuracy.

Fig. 3b shows the posterior distributions $q(z|r)$ of the proposed model, encoding the labeling behavior of each rater. The following observations confirm, that these learned distributions approximate well the real-world labeling behavior of the raters: (i) The four raters $3, 4, 5$ and 6 show a high overlap of the distributions and corresponding to a high labeling agreement shown in Fig. 3a. (ii) The gold distribution (simulating raters agreement) overlaps significantly with the distribution of these four raters. (iii) The distribution of rater 2 is far away from all other distributions. This rater shows a different labeling behavior due to frequent under-segmentation of images, assigning the 'background' class to areas that contain tissue. (iv) Raters 1 and 6 often deviate from the other raters, especially for the differentiation of classes G3 and G4. Their distribution accordingly has a smaller overlap with the gold distribution and the other raters. Fig. 4 shows some visual image

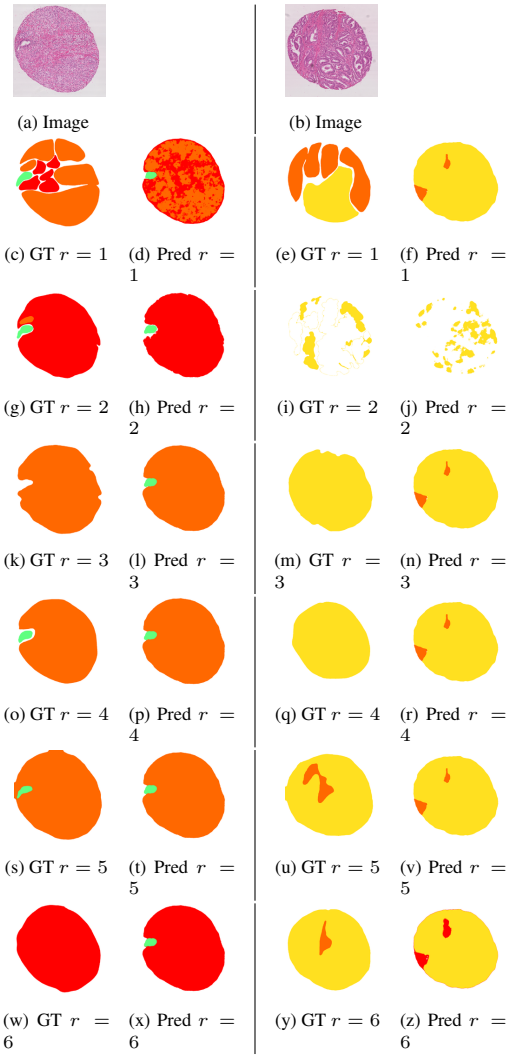


Figure 4: Inter-observer variations estimated by Pionono. For two test images we depict the ground truth (GT) segmentations of all raters and the predicted segmentations, simulating each rater. The proposed model is able to simulate certain labeling behaviour like the tendency of assigning class G5 (red) for raters 2 and 6 (see g and w) where other raters assigned G4 (orange). Furthermore, the model captures the under-segmention by rater 2 (see i).

examples of Pionono test predictions, simulating each rater r by drawing samples from the corresponding distribution $q(z|r)$ and then taking the mean of the output samples. The examples confirm that the rater differences are modeled well.

Next, we analyze the *intra-observer variations*. As the dataset does not contain multiple annotations of the same rater for the same image, the assessment of this quality is more difficult. Still, certain intra-observer variability can be assessed by observing the general labeling behavior of one annotator. For example, rater 6 tends to over-assign class G5 (red), and rater 2 tends to not segment all image parts that contain tissue. Interestingly, these intra-observer variations are present in the model predictions when multiple samples are drawn from their corresponding distribution. Fig. 5 shows visual examples of the simulated variations of raters 2 and 6.

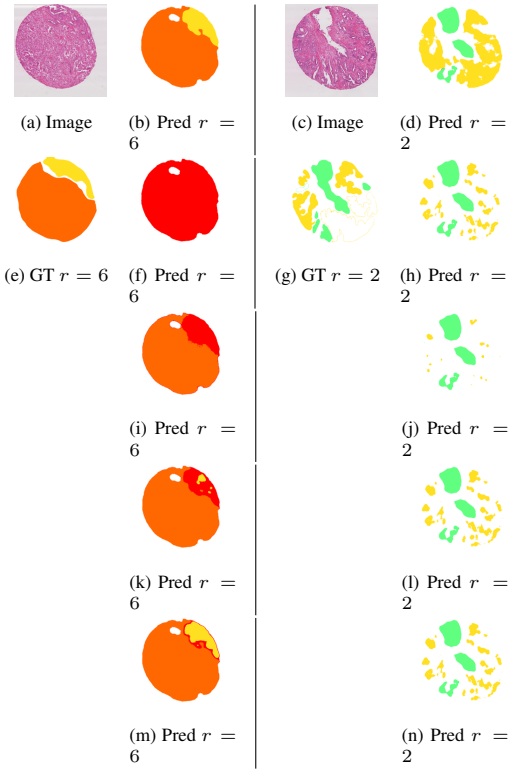


Figure 5: Intra-observer variations estimated by the Pionono model. For two example images we depict the true annotation of two different raters ($r = 2$ and $r = 6$). On the right side we show different coherent segmentation hypothesis for each rater estimated by our model. The differences in each column reflect possible intra-observer variations. The first example (a) shows that the segmentation of rater 6 might show some variations in the assigned classes. While the segmentation prediction (b) is composed by classes G3 (yellow) and G4 (orange), the segmentation sample (f) only consists of G5 (red). Indeed, this rater often assigns class G5 (red) in areas where other raters assign G4 (see Fig. 4) such that this is a plausible hypothesis. In the second example (c) we see variations due to the under-segmentation of rater 2 in some images. Our model captures this behaviour and provides different hypothesis of more (d) or less (j) segmentation of class G3 (yellow).

4.4 Model Comparison

The proposed model is compared to previously reported results and several state-of-the-art approaches (see section 2) for medical image segmentation with labels from multiple raters. For fair comparison, we use the same backbone architecture[2], epochs, learning rate, and optimizer for all experiments. We have tuned the model-specific hyperparameters to obtain the best possible results for each method.

First, we perform experiments with the Gleason 2019 dataset with a 4-fold crossvalidation. For comparison with previous works, we report the unweighted Cohen's kappa metric comparing gold predictions with gold ground truth. Additionally, we report the accuracy. As the results in Table 2 show, the proposed Pionono model outperforms the previously reported results [10, 34], including the winner of the Gleason 2019 challenge [34], by a large margin of over 20 percentage points. This accounts for the exact modeling of the raters by Pionono, but also for the different choices of backbone architecture and other training details. Compared to other state-of-the-art methods with the same architecture and training details, Pionono still shows a considerably better performance.

[2]Only the model *CM pixel* uses ResNet18 to fit on the GPU.

Method	Unweighted κ	Accuracy
Nir [10]	0.51	N.A.
Qiu [34]	0.524	N.A.
STAPLE	0.75 ± 0.006	0.834 ± 0.005
Prob U-Net	0.741 ± 0.002	0.83 ± 0.001
CM global	0.721 ± 0.018	0.814 ± 0.012
CM pixel	0.692 ± 0.019	0.791 ± 0.012
Pionono	$\mathbf{0.758} \pm 0.011$	$\mathbf{0.84} \pm 0.007$

Table 2: Cohens Kappa Comparison for the 4-fold crossvalidation experiment of the Gleason 2019 dataset, reported by mean and standard error.

Rater 1 Method	Quadratic κ	Accuracy
STAPLE	0.629 ± 0.002	0.718 ± 0.001
Prob U-Net	0.629 ± 0.005	0.73 ± 0.003
CM global	0.624 ± 0.003	0.728 ± 0.002
CM pixel	0.618 ± 0.007	0.72 ± 0.004
Pionono	$\mathbf{0.641} \pm 0.006$	$\mathbf{0.736} \pm 0.004$
Rater 2 Method	Quadratic κ	Accuracy
STAPLE	0.563 ± 0.002	0.621 ± 0.002
Prob U-Net	0.56 ± 0.003	0.626 ± 0.006
CM global	0.557 ± 0.003	$\mathbf{0.638} \pm 0.006$
CM pixel	0.551 ± 0.006	0.626 ± 0.008
Pionono	$\mathbf{0.569} \pm 0.005$	0.633 ± 0.005

Table 3: Cohens Kappa Comparison with the two raters of the Arvaniti TMA dataset trained on the Gleason 2019 dataset, reported by mean and standard error.

Next, we compare the generalization capabilities of the models by using the Arvaniti TMA dataset as an external test set, as reported in Table 3. This means, that the models are trained with all images from Gleason 2019 and tested with all images from Arvaniti TMA. As the Arvaniti TMA dataset does not contain gold labels, the model's gold predictions are compared to both raters independently, as previously done by Arvaniti [30]. We observe that the model generalizes better than all other methods, achieving a higher agreement in terms of Cohen's quadratic kappa with both raters. In terms of accuracy, only 'CM global' outperforms Pionono by a small margin.

In the third experiment, we use the Arvaniti TMA dataset for training and testing with the two rater annotations. Again, the proposed model is able to outperform previously reported results as well as other state-of-the-art methods in terms of quadratic Cohen's kappa. In terms of accuracy, only the "Prob U-Net" model obtains a better result for rater 1, while Pionono reaches the best accuracy for rater 2.

To validate the model on a different kind of data, we performed the fourth experiment on the "bc segmentation" dataset [4]. The results are reported in Table 5 and confirm the strong performance of the proposed Pionono model. The results support our hypothesis that explicitly modeling the inter- and intra-observer variations improves the model's performance. Pionono takes the different labeling behavior into account during training which leads to accurate predictions.

4.5 Robustness to Hyperparameter Settings

To measure the sensitivity of the model regarding different hyperparameters, we performed studies on the 4-fold cross-validation experiment of the Gleason 19 dataset. Table 6 shows that the model is robust to variations of all analyzed hyperparameters. We observe minor performance drops for different values of the regularization factor λ and the initialization variance σ^2_{post}. In both cases,

Rater 1 Method	Quadratic κ	Accuracy
Arvaniti [30]	0.55	N.A.
Silva-R. [17]	0.536	N.A.
supervised	0.658 ± 0.025	0.734 ± 0.008
Prob U-Net	0.697 ± 0.008	$\mathbf{0.762} \pm 0.004$
CM global	0.677 ± 0.028	0.745 ± 0.011
CM pixel	0.647 ± 0.016	0.731 ± 0.012
Pionono	$\mathbf{0.716} \pm 0.011$	0.751 ± 0.02
Rater 2 Method	Quadratic κ	Accuracy
Arvaniti	0.49	N.A.
supervised	0.521 ± 0.014	0.678 ± 0.014
Prob U-Net	0.534 ± 0.002	0.68 ± 0.005
CM global	0.533 ± 0.022	0.676 ± 0.011
CM pixel	0.508 ± 0.013	0.663 ± 0.008
Pionono	$\mathbf{0.548} \pm 0.008$	$\mathbf{0.697} \pm 0.012$

Table 4: Cohens Kappa Comparison with the two raters of the Arvaniti TMA dataset trained and validated by 4-fold crossvalidation of the Arvaniti data, reported by mean and standard error.

Method	Unweighted κ	Accuracy
STAPLE	0.647 ± 0.003	0.755 ± 0.002
Prob U-Net	0.685 ± 0.023	0.734 ± 0.004
CM global	0.654 ± 0.005	0.761 ± 0.004
CM pixel	0.689 ± 0.010	0.784 ± 0.007
Pionono	$\mathbf{0.711} \pm 0.002$	$\mathbf{0.799} \pm 0.001$

Table 5: Results for the breast cancer segmentation of WSIs, reported as mean and standard error of 4 runs.

wrong choices of the hyperparameters can hinder the correct optimization of the latent distributions. Furthermore, we tested different backbone architectures, indicating a limited performance with a VGG16 backbone. Overall the performance drops are minor and for all other settings, the model shows highly accurate results of $\kappa > 0.75$.

4.6 Required Resources

For the Gleason 2019 dataset with images of 1024×1024, the model can be trained with a batch size of 3 on a single NVIDIA GeForce RTX 3090 with 24Gb memory. The training takes less than 1.5h in total and test predictions less than 0.2s per image. The trained model occupies less than 350Mb when saved to the disk. As each additional annotator adds only one additional vector $\mu^r \in \mathbb{R}^8$ and one covariance matrix $\Sigma^r \in \mathbb{R}^{8\times8}$, it is scalable to a large number of annotators. The model's quick runtime and excellent scalability make it easily applicable in clinical practice.

4.7 Limitations

As semantic segmentation itself is a challenging task, some details of the annotator segmentations are not captured well by the model, such as the variations of class NC (green) in the GT of Fig. 4d - 4x or class G4 (orange) in the GT of Fig. 4f - 4z. Here, the model tends to predict similar shapes for the raters. A possible solution is to use more layers in the segmentation head f_θ with a wider kernel (e.g. 5×5 convolutions). This would increase the complexity of the model and might enable it to capture the different labeling behavior in even more detail.

Hyperp.	Value	Unweighted κ	Accuracy
D	4	0.752 ± 0.005	0.836 ± 0.003
	8	0.758 ± 0.011	0.84 ± 0.007
	16	0.752 ± 0.006	0.836 ± 0.004
σ^2_{prior}	1	0.758 ± 0.007	0.839 ± 0.004
	2	0.758 ± 0.011	0.84 ± 0.007
	4	0.757 ± 0.007	0.839 ± 0.004
σ^2_{post}	4	0.757 ± 0.007	0.839 ± 0.004
	8	0.758 ± 0.011	0.84 ± 0.007
	16	0.745 ± 0.009	0.83 ± 0.005
λ	0.0001	0.744 ± 0.003	0.829 ± 0.003
	0.0005	0.758 ± 0.011	0.84 ± 0.007
	0.001	0.745 ± 0.008	0.837 ± 0.005
ν	0.01	0.757 ± 0.004	0.839 ± 0.002
	0.02	0.758 ± 0.011	0.84 ± 0.007
	0.04	0.753 ± 0.01	0.836 ± 0.005
Backbone	VGG16	0.734 ± 0.01	0.823 ± 0.005
	Resnet34	0.758 ± 0.011	0.84 ± 0.007
	Eff.netB2	0.754 ± 0.01	0.836 ± 0.004

Table 6: Study of hyperparameter robustness using the Gleason 2019 dataset. The default hyperparameter value is marked with bold letters. While varying one hyperparameter, all other values are set to the default value. We observe consistent and robust performance across all settings of tested hyperparameters.

5 Conclusions

In this work we present "Pionono", a method for medical image segmentation that models the inter- and intra-observer variability explicitly with a probabilistic approximation. This is especially relevant for tasks where the labeling behavior of medical experts is known to vary widely, such as in the case of prostate cancer segmentation. Our experiments on real-world cancer segmentation data demonstrate that Pionono outperforms state-of-the-art models such as STAPLE, Probabilistic U-Net, and models based on confusion matrices. Apart from the improved predictive performance, it provides a probabilistic uncertainty estimation and the simulation of expert opinions for a given test image. This makes it a powerful tool for medical image analysis and has the potential to improve the diagnostic process considerably.

References

[1] S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, S. Eslami, D. Jimenez Rezende, and O. Ronneberger, "A probabilistic u-net for segmentation of ambiguous images," *Conference on Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.

[2] A. Schmidt, J. Silva-Rodriguez, R. Molina, and V. Naranjo, "Efficient cancer classification by coupling semi supervised and multiple instance learning," *IEEE Access*, vol. 10, pp. 9763–9773, 2022.

[3] A. Schmidt, P. Morales-Álvarez, and R. Molina, "Probabilistic attention based on gaussian processes for deep multiple instance learning," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2023.

[4] M. Amgad, H. Elfandy, H. Hussein, L. A. Atteya, M. A. T. Elsebaie, L. S. Abo Elnasr, R. A. Sakr, H. S. E. Salem, A. F. Ismail, A. M. Saad, J. Ahmed, M. A. T. Elsebaie, M. Rahman, I. A. Ruhban, N. M. Elgazar, Y. Alagha, M. H. Osman, A. M. Alhusseiny, M. M. Khalaf, A.-A. F. Younes, A. Abdulkarim, D. M. Younes, A. M. Gadallah, A. M. Elkashash, S. Y. Fala, B. M. Zaki, J. Beezley, D. R. Chittajallu, D. Manthey, D. A. Gutman, and L. A. D. Cooper, "Structured crowdsourcing enables convolutional segmentation of histology images," *Bioinformatics*, vol. 35, no. 18, pp. 3461–3467, 2019.

[5] E. C. Covert, K. Fitzpatrick, J. Mikell, R. K. Kaza, J. D. Millet, D. Barkmeier, J. Gemmete, J. Christensen, M. J. Schipper, and Y. K. Dewaraja, "Intra- and inter-operator variability in MRI-based manual segmentation of HCC lesions and its impact on dosimetry," *EJNMMI Physics*, vol. 9, no. 1, p. 90, 2022.

[6] N. S. Kulberg, R. V. Reshetnikov, V. P. Novik, A. B. Elizarov, M. A. Gusev, V. A. Gombolevskiy, A. V. Vladzymyrskyy, and S. P. Morozov, "Inter-observer variability between readers of CT images: all for one and one for all," *Digital Diagnostics*, vol. 2, no. 2, pp. 105–118, 2021.

[7] A. Mahbod, G. Schaefer, B. Bancher, C. Löw, G. Dorffner, R. Ecker, and I. Ellinger, "CryoNuSeg: A dataset for nuclei instance segmentation of cryosectioned h&e-stained histological images," *Computers in Biology and Medicine*, vol. 132, p. 104349, 2021.

[8] F. D. Allard, J. D. Goldsmith, G. Ayata, T. L. Challies, R. M. Najarian, I. A. Nasser, H. Wang, and E. U. Yee, "Intraobserver and interobserver variability in the assessment of dysplasia in ampullary mucosal biopsies," *American Journal of Surgical Pathology*, vol. 42, no. 8, pp. 1095–1100, 2018.

[9] L. Brochez, E. Verhaeghe, E. Grosshans, E. Haneke, G. Piérard, D. Ruiter, and J.-M. Naeyaert, "Inter-observer variation in the histopathological diagnosis of clinically suspicious pigmented skin lesions: Observer variation in pigmented lesion diagnosis," *The Journal of Pathology*, vol. 196, no. 4, pp. 459–466, 2002.

[10] G. Nir, S. Hor, D. Karimi, L. Fazli, B. F. Skinnider, P. Tavassoli, D. Turbin, C. F. Villamil, G. Wang, R. S. Wilson, K. A. Iczkowski, M. S. Lucia, P. C. Black, P. Abolmaesumi, S. L. Goldenberg, and S. E. Salcudean, "Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts," *Medical Image Analysis*, vol. 50, pp. 167–180, 2018, dataset link: https://gleason2019.grand-challenge.org/.

[11] J. Linmans, S. Elfwing, J. van der Laak, and G. Litjens, "Predictive uncertainty estimation for out-of-distribution detection in digital pathology," *Medical Image Analysis*, vol. 83, p. 102655, 2023.

[12] Y. Kwon, J.-H. Won, B. J. Kim, and M. C. Paik, "Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation," *Computational Statistics & Data Analysis*, vol. 142, p. 106816, 2020.

[13] M. Kandemir, M. Haussmann, F. Diego, K. Rajamani, J. Laak, and F. Hamprecht, "Variational weakly supervised gaussian processes," in *British Machine Vision Conference (BMVC)*, 2016, pp. 71.1–71.12.

[14] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning - ICML*, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48, 2016, pp. 1050–1059.

[15] Y. Wu, A. Schmidt, E. Hernández-Sánchez, R. Molina, and A. K. Katsaggelos, "Combining attention-based multiple instance learning and gaussian processes for CT hemorrhage detection," *Medical Image Computing and Computer Assisted Intervention – MICCAI*, vol. 12902, pp. 582–591, 2021.

[16] J. Li, W. Speier, K. C. Ho, K. V. Sarma, A. Gertych, B. S. Knudsen, and C. W. Arnold, "An EM-based semi-supervised deep learning approach for semantic segmentation of histopathological images from radical prostatectomies," *Computerized Medical Imaging and Graphics*, vol. 69, pp. 125–133, 2018.

[17] J. Silva-Rodríguez, A. Colomer, M. A. Sales, R. Molina, and V. Naranjo, "Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection," *Computer Methods and Programs in Biomedicine*, vol. 195, p. 105637, 2020.

[18] M. López-Pérez, M. Amgad, P. Morales-Álvarez, P. Ruiz, L. A. D. Cooper, R. Molina, and A. K. Katsaggelos, "Learning from crowds in digital pathology using scalable variational gaussian processes," *Scientific Reports*, vol. 11, no. 1, p. 11612, 2021.

[19] A. A. Abdullah, M. M. Hassan, and Y. T. Mustafa, "A review on bayesian deep learning in healthcare: Applications and challenges," *IEEE Access*, 2022.

[20] M. López-Pérez, A. Schmidt, Y. Wu, R. Molina, and A. K. Katsaggelos, "Deep gaussian processes for multiple instance learning: Application to CT intracranial hemorrhage detection," *Computer Methods and Programs in Biomedicine*, vol. 219, p. 106783, 2022.

[21] S. Warfield, K. Zou, and W. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Transactions on Medical Imaging*, vol. 23, no. 7, pp. 903–921, 2004.

[22] R. Tanno, A. Saeedi, S. Sankaranarayanan, D. C. Alexander, and N. Silberman, "Learning from noisy labels by regularized estimation of annotator confusion," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 11 244–11 253.

[23] L. Zhang, R. Tanno, M.-C. Xu, C. Jin, J. Jacob, O. Ciccarelli, F. Barkhof, and D. C. Alexander, "Disentangling human error from the ground truth in segmentation of medical images," in *International Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[24] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Conference on Neural Information Processing Systems (NeurIPS)*, ser. 32, 2019, pp. 8024–8035.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 9351, 2015, pp. 234–241.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[27] D. P. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," in *International Conference on Neural Information Processing Systems - NIPS*, 2015, pp. 2575–2583.

[28] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2017, pp. 240–248.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, 2015.

[30] E. Arvaniti, K. S. Fricker, M. Moret, N. Rupp, T. Hermanns, C. Fankhauser, N. Wey, P. J. Wild, J. H. Rüschoff, and M. Claassen, "Automated gleason grading of prostate cancer tissue microarrays via deep learning," *Scientific Reports*, vol. 8, no. 1, p. 12054, 2018, dataset link: https://doi.org/10.7910/DVN/OCYCMP.

[31] J. Silva-Rodríguez, A. Schmidt, M. A. Sales, R. Molina, and V. Naranjo, "Proportion constrained weakly supervised histopathology image classification," *Computers in Biology and Medicine*, vol. 147, p. 105714, 2022.

[32] M. López-Pérez, P. Morales-Álvarez, L. A. D. Cooper, R. Molina, and A. K. Katsaggelos, "Crowdsourcing segmentation of histopathological images using annotations provided by medical students," *Artificial Intelligence in Medicine - AIME*, vol. 13897, pp. 245–249, 2023.

[33] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, 2020.

[34] Y. Qiu, Y. Hu, P. Kong, H. Xie, X. Zhang, J. Cao, T. Wang, and B. Lei, "Automatic prostate gleason grading using pyramid semantic parsing network in digital histopathology," *Frontiers in Oncology*, vol. 12, p. 772403, 2022.

Chapter 7

Further Scientific Contributions

7.1 Probabilistic Multiple Instance Learning with CT Scans based on Gaussian Processes

7.1.1 Publication details

Authors: Yunan Wu, Arne Schmidt, Enrique Hernández-Sánchez, Rafael Molina, Aggelos K. Katsaggelos

Title: Combining Attention-based Multiple Instance Learning and Gaussian Processes for CT Hemorrhage Detection

Reference: International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), 2021, doi: 10.1007/978-3-030-87196-3_54

Status: Published

Quality indices:

- GGS Rating (2021): A
- GGS Class (2021): 2
- CORE: A

7.1.2 Main contributions

- We propose a two stage approach for multiple instance learning. First we train a feature extractor with an attention mechanism, extract the features, and apply a probabilistic model based on VGPMIL (Variational Gaussian Process for Multiple Instance Learning).
- The model is experimentally tested with two public datasets for binary classification of hemorrhage detection in computerized tomography (CT) scans. Each bag in this context consists of a scan of a patient's brain and the instances consist of multiple slices from different levels of the brain. If one or more slices show patterns of hemorrhage, the patient has a positive label, else a negative one.
- The developed method was the inspiration for the later developed end-to-end approach which includes Gaussian processes and attention mechanisms in a single model 2.

7.1.3 Abstract

Intracranial hemorrhage (ICH) is a life-threatening emergency with high rates of mortality and morbidity. Rapid and accurate detection of ICH is crucial for patients to get a timely treatment. In order to achieve the automatic diagnosis of ICH, most deep learning models rely on huge amounts of slice labels for training. Unfortunately, the manual annotation of CT slices by radiologists is time-consuming and costly. To diagnose ICH, in this work, we propose to use an attention-based multiple instance learning (Att-MIL) approach implemented through the combination of an attention-based convolutional neural network (Att-CNN) and a variational Gaussian process for multiple instance learning (VGPMIL). Only labels at scan-level are necessary for training. Our method (a) trains the model using scan labels and assigns each slice with an attention weight, which can be used to provide slice-level predictions, and (b) uses the VGPMIL model based on low-dimensional features extracted by the Att-CNN to obtain improved predictions both at slice and scan levels. To analyze the performance of the proposed approach, our model has been trained on 1150 scans from an RSNA dataset and evaluated on 490 scans from an external CQ500 dataset. Our method outperforms other methods using the same scan-level training and is able to achieve comparable or even better results than other methods relying on slice-level annotations.

7.2 Deep Gaussian Processes for Multiple Instance Learning with CT Scans

7.2.1 Publication details

Authors: Miguel López-Pérez, Arne Schmidt, Yunan Wu, Rafael Molina, Aggelos K. Katsaggelos

Title: Deep Gaussian Processes for Multiple Instance Learning: Application to CT Intracranial Hemorrhage Detection

Reference: Computer Methods and Programs in Biomedicine, vol. 219, 106783, 2022, doi: 10.1016/j.cmpb.2022.106783

Status: Published

Quality indices:

- Impact Factor (JCR 2022): 6.1
  - Rank 25/110 (Q1) in Computer Science, Theory and Methods
  - Rank 22/96 (Q1) in Engineering, Biomedical

7.2.2 Main contributions

- Extending the work with Gaussian processes for multiple instance learning (chapter 7.1), we propose a novel model with Deep Gaussian processes with several layers of Gaussian processes for more expressiveness. The output of one Gaussian process is the input for the next Gaussian process, such that a deep model is designed. It is able to capture more complex relationships in the features while maintaining its probabilistic properties.
- In a toy experiment with MNIST, we show that the model with more layers can process more complex data relationships. The deep model improved the results on CT hemorrhage detection previously obtained (7.1) and other state-of-the-art models.
- The extension of a Gaussian process-based model top deep Gaussian processes is very interesting for other applications, including the models that were presented in this thesis in Chapters 2 and 3.

7.2.3 Abstract

Background and objective:
Intracranial hemorrhage (ICH) is a life-threatening emergency that can lead to brain damage or death, with high rates of mortality and morbidity. The fast and accurate detection of ICH is important for the patient to get an early and efficient treatment. To improve this diagnostic process, the application of Deep Learning (DL) models on head CT scans is an active area of research. Although promising results have been obtained, many of the proposed models require slice-level annotations by radiologists, which are costly and time-consuming.
Methods:
We formulate the ICH detection as a problem of Multiple Instance Learning (MIL) that allows training with only scan-level annotations. We develop a new probabilistic method based on Deep Gaussian Processes (DGP) that is able to train with this MIL setting and accurately predict ICH at both slice- and scan-level. The proposed DGPMIL model is able to capture complex feature relations by using multiple Gaussian Process (GP) layers, as we show experimentally.
Results:
To highlight the advantages of DGPMIL in a general MIL setting, we first conduct several controlled experiments on the MNIST dataset. We show that multiple Gaussian process layers outperform one-layer Gaussian process models, especially for complex feature distributions. For ICH detection experiments, we use two public brain CT datasets (RSNA and CQ500). We first train a Convolutional Neural Network (CNN) with an attention mechanism to extract the image features, which are fed into our DGPMIL model to perform the final predictions. The results show that DGPMIL model outperforms VGPMIL as well as the attention-based CNN for MIL and other state-of-the-art methods for this problem. The best performing DGPMIL model reaches an AUC-ROC of 0.957 (resp. 0.909) and an AUC-PR of 0.961 (resp. 0.889) on the RSNA (resp. CQ500) dataset.
Conclusion:
The competitive performance at slice- and scan-level shows that DGPMIL model provides an accurate diagnosis on slices without the need for slice-level annotations by radiologists during training. As MIL is a common problem setting, our model can be applied to a broader range of other tasks, especially in medical image classification, where it can help the diagnostic process.

7.3 Multiple Instance Learning with Constrained Optimization

7.3.1 Publication details

Authors: Julio Silva-Rodríguez, Arne Schmidt, Maria A. Sales, Rafael Molina, Valery Naranjo

Title: Proportion constrained weakly supervised histopathology image classification

Reference: Computers in Biology and Medicine, Vol. 147, 105714, 2022

Status: Published

Quality indices:

- Impact Factor (JCR 2022): 7.7
  - Rank 18/110 (Q1) in Computer Science, Interdisciplinary Applications
  - Rank 7/92 (D1) in Biology

7.3.2 Main contributions

- Constrained optimization can be used to incorporate prior knowledge for multiple instance learning optimization. The proposed model translates known class proportions derived by the bag label into a constrained optimization problem to improve the multiple instance learning classifier. The model has a theoretical foundation in optimization with log-barrier extensions.
- The model was experimentally tested on a new dataset, SICAP-MIL, which was made publicly available. It shows a strong performance on instance, as well as on bag-level.
- Although the presented model is not probabilistic, it has a sound mathematical background and the application of multiple instance learning for prostate cancer classification is highly related to other articles presented in this thesis (see Chapters 2 - 4).

7.3.3 Abstract

Multiple instance learning (MIL) deals with data grouped into bags of instances, of which only the global information is known. In recent years, this weakly supervised learning paradigm has become very popular in histological image analysis because it alleviates the burden of labeling all cancerous regions of large Whole Slide Images (WSIs) in detail. However, these methods require large datasets to perform properly, and many approaches only focus on simple binary classification. This often does not match the real-world problems where multi-label settings are frequent and possible constraints must be taken into account. In this work, we propose a novel multi-label MIL formulation based on inequality constraints that is able to incorporate prior knowledge about instance proportions. Our method has a theoretical foundation in optimization with logbarrier extensions, applied to bag-level class proportions. This encourages the model to respect the proportion ordering during training. Extensive experiments on a new public dataset of prostate cancer WSIs analysis, SICAP-MIL, demonstrate that using the prior proportion information we can achieve instance-level results similar to supervised methods on datasets of similar size. In comparison with prior MIL settings, our method allows for $\sim 13\%$ improvements in instance-level accuracy, and $\sim 3\%$ in the multi-label mean area under the ROC curve at the bag-level.

7.4 Acquisition and Processing of Whole Slide Images

7.4.1 Publication details

Authors: Neel Kanwal, Fernando Pérez-Bueno, Arne Schmidt, Kjersti Engan, Rafael Molina

Title: The devil is in the details: Whole Slide Image Acquisition and Processing for Artifact Detection, Color Variation, and Data Augmentation. A Review.

Reference: IEEE Access, vol. 10, 58821-58844, 2022, doi: 10.1109/ACCESS.2022.3176091

Status: Published

Quality indices:

- Impact Factor (JCR 2022): 3.9
  - Rank 72/158 (Q2) in Computer Science, Engineering Systems
  - Rank 100/275 (Q2) in Engineering, Electrical and Electronic

7.4.2 Main contributions

- This article describes the process of preparing, digitizing and digesting biopsies for digital pathology. It informs about the several manual steps to obtain a glass slide which can be used for microscopes as well as the digital steps after scanning the biopsy, including the preparation for artificial intelligence algorithms.
- One focus lies on artifacts, such as air bubbles, pen markers or ink. They can have a high impact and played a major role in our article about probabilistic active learning (Chapter 2).
- Another topic is the color variation in the WSIs which can lead to problems for AI models when there are high differences between several datacenters.
- Data augmentation is one possibility to prevent problems due to color variations. Furthermore, it played a major role in our work about semi supervised and multiple instance learning (Chapter 4).

7.4.3 Abstract

Whole Slide Images (WSI) are widely used in histopathology for research and the diagnosis of different types of cancer. The preparation and digitization of histological tissues leads to the introduction of artifacts and variations that need to be addressed before the tissues are analyzed. WSI preprocessing can significantly improve the performance of computational pathology systems and is often used to facilitate human or machine analysis. Color preprocessing techniques are frequently mentioned in the literature, while other areas are usually ignored. In this paper, we present a detailed study of the state-of-the-art in three different areas of WSI preprocessing: Artifacts detection, color variation, and the emerging field of pathology-specific data augmentation. We include a summary of evaluation techniques along with a discussion of possible limitations and future research directions for new methods.

Chapter 8

Concluding remarks

The main conclusion of this thesis is that probabilistic deep learning methods can provide important contributions to overcome the labeling bottleneck for medical images. When limited or imperfect labels are available, probability theory can successfully address the uncertainties that can arise from different sources. The proposed models enable the application of AI to many new, future medical applications without the need for detailed, supervised annotations. Therefore, this work is one important step towards more accurate, faster and reproducible diagnostic processes supported by AI algorithms.

For multiple instance learning, we have shown that attention mechanisms can benefit substantially from probabilistic attention estimations by Gaussian processes. The exact attention weight regression of the Gaussian processes allows an overall better performance. Capturing the uncertainty of missing instance labels allows to provide uncertainty estimation which correlates with the risk of wrong predictions, as shown in Chapter 2. In a different work, presented in Chapter 3, we showed how correlations between instances can be modeled in a probabilistic framework with a Gaussian process classifier. For active learning, we adapted Bayesian neural networks to capture different uncertainties that are relevant to this task, as shown in Chapter 5. While the epistemic (model-related) uncertainty can be used to measure the informativeness of new image patches, the aleatoric (data-based) uncertainty and out-of-distribution detection avoid acquiring uninformative image patches like ambiguous images or artifacts. For crowdsourcing, investigated in Chapter 6, we explicitly modeled inter- and intra-observer variability in a novel proabilitistic generative model. This allows for accurate segmentation predictions, uncertainty estimations and the simulation of expert opinions.

Overall we observed that the proposed probabilistic models successfully tackle the challenge imposed by the labeling bottleneck. They are able to obtain a competitive performance, often outperforming the existing state-of-the-art. Additionally, they are able to perform probabilistic reasoning, which is of high importance. In the predictions, the output distribution incorporates uncertainties resulting from limited or imperfect labels during training.

Based on our research we consider the further development of probabilistic models very important for future investigation in medical image analysis. In the clinical context, each

source of uncertainty must be addressed in a principled way to be able to rely on AI models. Here, human reasoning can serve as an example: when a decision or opinion can not be made due to missing information, ambiguities, or incomplete knowledge, a trustworthy person would express this uncertainty together with the estimation of the situation.

We would like to highlight two specific challenges that should be addressed in future research. Firstly, the systematic evluation and benchmarks for uncertainty estimations on clinical datasets are essential. While existing methods are often evaluated based on performance metrics like accuracy or Cohen's kappa, additional metrics should be devised to compare systematically the estimated uncertainties of different models. For instance, the correlation of high uncertainties with a high risk of incorrect predictions or the analysis of uncertainties in controlled experiments with known uncertainties, such as artificial noise, could be explored. Secondly, the principles of probabilistic reasoning investigated in this thesis should be extended to the area of large language models. Recently, models on the GPT architecture [11] have gained considerable attention, also in the medical domain [21]. Despite their impressive performance, it is often criticized that these models present potentially incorrect facts or hallucinations with high confidence [22]. By using probabilistic models, future efforts should focus on ensuring that these models express uncertainties in their output when there is a risk of being wrong. Only with this extension, they could be reliably applied in the medical context. As attention mechanisms are a crucial component of transformers, leveraging the proposed attention mechanism with Gaussian processes could be a promising direction for achieving this goal.

In summary, we anticipate that probabilistic modeling will continue to play a crucial role in developing reliable AI models to overcome the labeling bottleneck and significantly improve the diagnostic processes of the future.

Bibliography

[1] M. L. Smith, L. N. Smith, and M. F. Hansen, "The quiet revolution in machine vision - a state-of-the-art survey paper, including historical review, perspectives, and future directions," *Computers in Industry*, vol. 130, p. 103472, 2021.

[2] A. Hekler, J. S. Utikal, A. H. Enk, W. Solass, M. Schmitt, J. Klode, D. Schadendorf, W. Sondermann, C. Franklin, F. Bestvater, M. J. Flaig, D. Krahl, C. von Kalle, S. Fröhling, and T. J. Brinker, "Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images," *European Journal of Cancer*, vol. 118, pp. 91–96, 2019.

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

[4] F. D. Allard, J. D. Goldsmith, G. Ayata, T. L. Challies, R. M. Najarian, I. A. Nasser, H. Wang, and E. U. Yee, "Intraobserver and interobserver variability in the assessment of dysplasia in ampullary mucosal biopsies," *American Journal of Surgical Pathology*, vol. 42, no. 8, pp. 1095–1100, 2018.

[5] W. C. Allsbrook, K. A. Mangold, M. H. Johnson, R. B. Lane, C. G. Lane, and J. I. Epstein, "Interobserver reproducibility of gleason grading of prostatic carcinoma: General pathologist," *Human Pathology*, vol. 32, no. 1, pp. 81–88, 2001.

[6] L. Brochez, E. Verhaeghe, E. Grosshans, E. Haneke, G. Piérard, D. Ruiter, and J.-M. Naeyaert, "Inter-observer variation in the histopathological diagnosis of clinically suspicious pigmented skin lesions: Observer variation in pigmented lesion diagnosis," *The Journal of Pathology*, vol. 196, no. 4, pp. 459–466, 2002.

[7] M. A. Makary and M. Daniel, "Medical error—the third leading cause of death in the US," *BMJ*, p. i2139, 2016.

[8] D. E. Newman-Toker, Z. Wang, Y. Zhu, N. Nassery, A. S. Saber Tehrani, A. C. Schaffer, C. W. Yu-Moe, G. D. Clemens, M. Fanai, and D. Siegal, "Rate of diagnostic errors

and serious misdiagnosis-related harms for major vascular events, infections, and cancers: toward a national incidence estimate using the "big three"," *Diagnosis*, vol. 8, no. 1, pp. 67–84, 2021.

[9] C. L. Hitchcock, "The future of telepathology for the developing world," *Archives of Pathology and Laboratory Medicine*, vol. 135, no. 2, pp. 211–214, 2011.

[10] N. Dimitriou, O. Arandjelović, and P. D. Caie, "Deep learning for whole slide image analysis: An overview," *Frontiers in Medicine*, vol. 6, p. 264, 2019.

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[12] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*, ser. Adaptive computation and machine learning. MIT Press, 2006, OCLC: ocm61285753.

[13] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning - ICML*, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. PMLR, 2016, pp. 1050–1059.

[14] S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, S. Eslami, D. Jimenez Rezende, and O. Ronneberger, "A probabilistic u-net for segmentation of ambiguous images," *Conference on Neural Information Processing Systems - NeurIPS*, vol. 31, 2018.

[15] N. Kanwal, F. Perez-Bueno, A. Schmidt, K. Engan, and R. Molina, "The devil is in the details: Whole slide image acquisition and processing for artifacts detection, color variation, and data augmentation: A review," *IEEE Access*, vol. 10, pp. 58 821–58 844, 2022.

[16] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329 – 353, 2018.

[17] S. Budd, E. C. Robinson, and B. Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis," *Medical Image Analysis*, vol. 71, p. 102062, 2021.

[18] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Neural Information Processing Systems - NeurIPS*, vol. 30, 2017.

[19] E. Snelson and Z. Ghahramani, "Sparse gaussian processes using pseudo-inputs," in *Advances in Neural Information Processing Systems*, vol. 18, 2006.

[20] M. Haussmann, F. A. Hamprecht, and M. Kandemir, "Variational bayesian multiple instance learning with gaussian processes," in *IEEE Conference on Computer Vision and Pattern Recognition - CVPR*, 2017, pp. 810–819.

[21] V. W. Xue, P. Lei, and W. C. Cho, "The potential impact of ChatGPT in clinical and translational medicine," *Clinical and Translational Medicine*, vol. 13, no. 3, p. e1216, 2023.

[22] M. Sallam, "ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns," *Healthcare*, vol. 11, no. 6, p. 887, 2023.