UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Learning from limited labelled data: contributions to weak, few-shot, and unsupervised learning

November, 2022

COMMUNICATIONS DEPARTMENT

Author: Julio José Silva Rodríguez

Supervisors: Valery Naranjo Ornedo
Ricardo Insa Franco
Pablo Salvador Zuriaga

The work of Julio Silva Rodríguez to carry out this research and to elaborate this dissertation has been supported by the Spanish Government under the FPI Grant PRE2018-083443.

Acknowledgements

La presente tesis es el resultado de un largo camino formativo en el cual muchas personas han dejado su huella. Por ello, creo conveniente dedicar unas palabras de agradecimiento a las mismas.

Me gustaría dedicar mis primeras palabras a mis directores, quienes me han otorgado la posibilidad de realizar esta tesis. A Ricardo y Pablo, por introducirme en el proyecto que ha hecho posible esta tesis. En especial, a Valery, por su consejo y ayuda en el día a día. Por ser la artífice de un grupo en el cual los jóvenes podamos crecer como investigadores, y aspirar a todas nuestras metas. Gracias.

A quienes han compartido su conocimiento conmigo. A Adri, por apoyar mis primeros pasos en el ámbito de la visión por computador y el *deep learning.* A Jose Dolz, por abrirme nuevos caminos durante mi estancia en Montreal. Asímismo, querría acordarme del AlgoTeam de ContinUse, por apostar por mi y darme una invaluable formación en mis primeros años de investigador.

Gracias a todos mis compañeros del CVBLAB por compartir ideas, bromas y amistad. A todos los que han compartido su camino con el mío: Gabri, Elena, Rocío, Cristian, Laëti y Claudio. A los que inician ahora el suyo: Alejandro, Pablo e Ilán. Mi eterna gratitud a quien más horas ha dedicado a esta tesis, día y noche, de forma desinteresada: gracias DGX, y gracias GPU-4.

Finalmente, querría tener un agradecimiento especial a mi familia. A mis padres y hermano, por educarme, cuidarme, y apoyarme en todos mis proyectos. A Anny, por aguantar mi pasión, en ocasiones obsesiva, en este trabajo; por escucharme y aconsejarme siempre que lo necesitaba; y por acompañarme en este camino, y en los que están por venir. Os quiero.

Abstract

In the last decade, deep learning (DL) has become the main tool for computer vision (CV) tasks. Under the standard supervised learnng paradigm, and thanks to the progressive collection of large datasets, DL has reached impressive results on different CV applications using convolutional neural networks (CNNs). Nevertheless, CNNs performance drops when sufficient data is unavailable, which creates challenging scenarios in CV applications where only few training samples are available, or when labeling images is a costly task, that require expert knowledge. Those scenarios motivate the research of not-so-supervised learning strategies to develop DL solutions on CV.

In this thesis, we have explored different less-supervised learning paradigms on different applications. Concretely, we first propose novel self-supervised learning strategies on weakly supervised classification of gigapixel histology images. Then, we study the use of contrastive learning on few-shot learning scenarios for automatic railway crossing surveying. Finally, brain lesion segmentation is studied in the context of unsupervised anomaly segmentation, using only healthy samples during training. Along this thesis, we pay special attention to the incorporation of tasks-specific prior knowledge during model training, which may be easily obtained, but which can substantially improve the results in less-supervised scenarios. In particular, we introduce relative class proportions in weakly supervised learning in the form of inequality constraints. Also, attention homogenization in VAEs for anomaly localization is incorporated using size and entropy regularization terms, to make the CNN to focus on all patterns for normal samples. The different methods are compared, when possible, with their supervised counterparts.

In short, different not-so-supervised DL methods for CV are presented along this thesis, with substantial contributions that promote the use of DL in data-limited scenarios. The obtained results are promising, and provide researchers with new tools that could avoid annotating massive amounts of data in a fully supervised manner.

Resumen

En la última década, el aprendizaje profundo (DL) se ha convertido en la principal herramienta para las tareas de visión por ordenador (CV). Bajo el paradigma de aprendizaje supervisado, y gracias a la recopilación de grandes conjuntos de datos, el DL ha alcanzado resultados impresionantes utilizando redes neuronales convolucionales (CNNs). Sin embargo, el rendimiento de las CNNs disminuye cuando no se dispone de suficientes datos, lo cual dificulta su uso en aplicaciones de CV en las que sólo se dispone de unas pocas muestras de entrenamiento, o cuando el etiquetado de imágenes es una tarea costosa. Estos escenarios motivan la investigación de estrategias de aprendizaje menos supervisadas.

En esta tesis, hemos explorado diferentes paradigmas de aprendizaje menos supervisados. Concretamente, proponemos novedosas estrategias de aprendizaje autosupervisado en la clasificación débilmente supervisada de imágenes histológicas gigapixel. Por otro lado, estudiamos el uso del aprendizaje por contraste en escenarios de aprendizaje de pocos disparos para la vigilancia automática de cruces de ferrocarril. Por último, se estudia la localización de lesiones cerebrales en el contexto de la segmentación no supervisada de anomalías. Asimismo, prestamos especial atención a la incorporación de conocimiento previo durante el entrenamiento que pueda mejorar los resultados en escenarios menos supervisados. En particular, introducimos proporciones de clase en el aprendizaje débilmente supervisado en forma de restricciones de desigualdad. Además, se incorpora la homogeneización de la atención para la localización de anomalías mediante términos de regularización de tamaño y entropía.

A lo largo de esta tesis se presentan diferentes métodos menos supervisados de DL para CV, con aportaciones sustanciales que promueven el uso de DL en escenarios con datos limitados. Los resultados obtenidos son prometedores y proporcionan a los investigadores nuevas herramientas que podrían evitar la anotación de cantidades masivas de datos de forma totalmente supervisada.

Resum

En l'última dècada, l'aprenentatge profund (DL) s'ha convertit en la principal eina per a les tasques de visió per ordinador (CV). Sota el paradigma d'aprenentatge supervisat, i gràcies a la recopilació de grans conjunts de dades, el DL ha aconseguit resultats impressionants utilitzant xarxes neuronals convolucionals (CNNs). No obstant això, el rendiment de les CNNs disminueix quan no es disposa de suficients dades, la qual cosa dificulta el seu ús en aplicacions de CV en les quals només es disposa d'unes poques mostres d'entrenament, o quan l'etiquetatge d'imatges és una tasca costosa. Aquests escenaris motiven la investigació d'estratègies d'aprenentatge menys supervisades.

En aquesta tesi, hem explorat diferents paradigmes d'aprenentatge menys supervisats. Concretament, proposem noves estratègies d'aprenentatge autosupervisat en la classificació feblement supervisada d'imatges histològiques gigapixel. D'altra banda, estudiem l'ús de l'aprenentatge per contrast en escenaris d'aprenentatge de pocs trets per a la vigilància automàtica d'encreuaments de ferrocarril. Finalment, s'estudia la localització de lesions cerebrals en el context de la segmentació no supervisada d'anomalies. Així mateix, prestem especial atenció a la incorporació de coneixement previ durant l'entrenament que puga millorar els resultats en escenaris menys supervisats. En particular, introduïm proporcions de classe en l'aprenentatge feblement supervisat en forma de restriccions de desigualtat. A més, s'incorpora l'homogeneïtzació de l'atenció per a la localització d'anomalies mitjançant termes de regularització de grandària i entropia.

Al llarg d'aquesta tesi es presenten diferents mètodes menys supervisats de DL per a CV, amb aportacions substancials que promouen l'ús de DL en escenaris amb dades limitades. Els resultats obtinguts són prometedors i proporcionen als investigadors noves eines que podrien evitar l'anotació de quantitats massives de dades de forma totalment supervisada.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter introduces the motivation and the objectives pursued in this thesis, as well as the main contributions. It also includes the thesis framework and the thesis outline.

Contents

1.1 Motivation

Computer vision and deep learning

Computer vision (CV) is a scientific field born in the early 1960s with the advent of the first digital image scanners, that aims to mimic the effect of human vision by electronically perceiving and understanding an image [1]. In other words, computer vision systems aim to develop visual perception: to describe the objects in the image, their categories and their context [2]. Image understanding includes, among others, different tasks such as image classification, semantic segmentation, object detection or scene description.

Classical computer vision systems included two stages. First, the information contained in the image was compressed in a set of features, usually obtained by image processing operations. Among others, outstanding approaches included texture analysis via local binary patterns (LBPs) [3], HoG [4], or SIFT [5] features. Then, those feature descriptors were passed through trained machine learning methods, that were in charge of providing predictions or decisions based on data patterns. Some of the typically used classifiers included simple linear regression, support vector machines (SVMs) [6], or random forest [7].

The definition of hand-crafted features using image processing techniques is a tedious and challenging process. There are tasks that are easy for people to perform, but hard to describe formally [8]. This is where deep learning (DL) comes to play. Deep learning is a field of machine learning that is not limited to pattern recognition, but also to representation learning. Representation learning is a set of methods that allow a machine to be fed with raw data and to automatically discover the representations (features) needed for detection or classification [9]. Deep learning methods are based on artificial neural networks (ANNs). Inspired on its biological counterpart, ANNs are algorithms that combine stacked layers of neurons, sequentially connected one to each other between, to learn patterns on input data given an objective task.

The foundations of deep learning in computer vision date back to late 1970s, with the Neocognitron architecture [10]. The maturity arrived thanks to the contributions of LeCun *et al.* [11, 12] on backpropagation architectures in the 1990s, and the refinement of pooling operations [13], the use of ReLU activation or dropout regularization [14]. These advances composed the modern convolutional neural networks (CNNs) architectures, a particular type of ANNs based on trainable convolutional filters able to integrate local visual patterns. Still, it is only in the recent 21st century that DL has become the main tool in computer vision systems thanks of software advances and CNNs training on graphic process units (GPUs) [15–17]. In particular, AlexNet network from Krizhevsky *et al.* [16], which ranked 1st position on ImageNet[1] 2012 challenge on image classification, is considered the milestone that started the CNNs era. Since then, a golden decade of computer vision using deep learning has led astonishing performance across different applications such as: natural image categorization [18], object detection [19], autonomous driving [20], medical imaging analysis [21], old image and video restoration [22], industrial predictive maintenance [23, 24], video captioning [25] or image generation [26], among others.

The data barrier

The successful application of CNNs in computer vision systems is closely connected to the ability to collect data. As discussed by Goodfellow *et al.* [8], deep learning has become more useful as the amount of available training data has increased. Indeed, the first outstanding results of deep learning temporally coincide with the release of public challenges with large amounts of data. Since early studies of modern CNN architectures [27, 28], its performance has shown a log-dependence on the amount of training data. This is, CNNs require a reasonable amount of data to perform properly.

When we refer to data, we are not only referring to the original images, but also to the expert knowledge that has to be provided to the model. Under the standard supervised scenario, images must be labelled at global label (i.e. categories in the image), object-level (bounding boxes ob objects), or at pixel level, on semantic segmentation tasks. The main core of deep learning architectures for image recognition task are benchmarked

[1] `https://www.image-net.org`

on natural images in datasets such as ImageNet, CIFAR[2], COCO[3], etc. Those datasets contain images of real-world animals and objects, which can be easily labeled with general knowledge. However, achieving large datasets in real-world applications can be a very challenging task. For instance, in medical imaging applications, expert knowledge is limited to physicians. Obtaining such curated labeled datasets is a cumbersome process prone to subjectivity, which makes access to sufficient training data difficult in practice. This problem can be magnified in the context of specific imaging applications that require pixel-level annotations, such as radiology, when volumetric data from magnetic resonance imaging comes to play, or histology, based on gigapixel images magnified under the microscope. Another challenging scenario is the application of CV in industry. In this case, it is often difficult to get a large number of examples, and the data domain, usually sensor-based, is very different from that of natural imagery. Also, in scenarios such as predictive maintenance, it is not possible to obtain a priori examples of possible defects in a heterogeneous way.

For all the above, there is a need to develop novel deep learning methods capable of performing well in data-poor scenarios, where the standard supervised learning scenario results impractical or infeasible. This includes algorithms capable of incorporating any type of knowledge into learning that is easily accessible, as opposed to the tedious process of annotation at the image or pixel level, and also models able to learn on scarce, and imbalanced datasets. In this thesis, we refer to those methods as *not-so-supervised* or *less-supervised* strategies, analogously to Cheplygina *et al.* [29]

Towards a less-supervised perspective

To alleviate the need of data to train well-performing deep learning models, different research lines are exploring the use of other *less-supervised* learning strategies. Popular strategies in this field using CNNs include transfer learning of knowledge from models trained on large datasets [30], or data augmentation [31]. Other hot research topics focus on *unsupervised learning*, which use only unlabeled images to feed deep

[2] https://www.cs.toronto.edu/~kriz/cifar.html
[3] https://cocodataset.org

learning models. In this setting, visual features are usually trained using pretext tasks in a self-supervised fashion. Promising methods in this line combine data augmentation with contrastive learning [32, 33], jigsaw puzzle reconstruction [34], or generative networks [35]. Then, the CNN backbone or classification layers are retrained for task-specific purposes. In a middle ground before the unsupervised and supervised scenarios, other learning strategies aim to leverage indirect, noisy, or inexact knowledge, so called *weakly supervised learning* scenarios. In a popular scenario, weakly supervised learning is used for object localization and segmentation tasks [36] using global labels instead of tedious pixel-level annotations. Still, those methods require large amounts of weak- or unlabeled data to get promising results. The scenario in which this data in simply unavailable is covered under the *few-shot learning* paradigm [37, 38], where learning is driven only by no more than tens of images.

Still, the performance gap between the supervised scenario and less-supervised ones is outstanding. For instance, the best performing method on weakly supervised segmentation [39] performs $\sim 25\%$ worse in mean intersection-over-union than its supervised counterpart on PASCAL VOC 2012 dataset. On unsupervised learning, the gap on image classification using SimCLR method [33] is of $\sim 8\%$ accuracy on ImageNet. Regarding the results obtained under the few-shot learning setting, they are still far of being competitive.

1.2 Objectives

The main objective of this thesis is the design, development and validation of innovative *not-so-supervised* methods to solve real-world computer vision challenges using deep learning. We aim to address a wide number of perspectives, that best fit in each particular application. However, we also seek to make the proposed methodologies largely generalizable. In particular, the specific objectives include:

- Contribute to the *weakly supervised learning* literature, using only global labels during training, in the context of multi-label gigapixel image segmentation.

- Study the capability of *few-shot learning* on real-world applications with scarce data, and improve the existing state-of-the-art approaches.
- Develop novel deep learning strategies on *unsupervised anomaly segmentation*, able to model normal data during training to locate out-of-distribution anomalies on inference.

Likewise, and transversally to the different methods, this thesis aims to propose novel strategies that can incorporate a priori knowledge easily accessible in each application, and which do not require a laborious annotation process. To do so, novel *constrained formulations* will be taken into account for the different learning settings.

1.3 Framework

This PhD thesis is framed within three different research projects on applied computer vision. These projects are introduced below:

- *SICAP* – Histopathological image interpretation system for the detection of prostate cancer. *SICAP* is a national project which objective is to develop computer vision algorithms to automatize the diagnosis and prognosis prediction in prostate histology biopsies analysis. This project was funded by the *Ministerio de Economía, Industria y Competitividad* (DPI2016-77869-C2-1-R). The automatic analysis of biopsies is a challenging task. Digitised biopsies under the microscope constitute gigapixel samples known as whole slide images (WSIs). The large size of these images makes computer vision systems work with small patches of the image, which require local annotations made by expert pathologists. In the case of prostate, the different tumor patterns must be delimited at the pixel level following the Gleason grading scale. This is a tedious process, prone tu interannotator variability, which makes it difficult to use large databases during training. An important part of this thesis is framed within this project. In particular, Chapter 2 develops a computer vision system that covers the different stages on biopsy analysis under standard supervised learning methods. Then, Chapter 3, Chapter 4 and Chapter 5 explore the use of weakly supervised strategies to alleviate the need of pixel-level annotations. In this case, training is driven by global labels, which indicate the presence of tumour patterns in the whole biopsy.

- *INTELVIA* – Smart track dynamic surveying approaches based on digital image processing. This national project aims to develop novel computer vision systems for the automatic analysis of railway deterioration using axle-box accelerations. *INTELVIA* project was funded by the *Ministerio de Economía, Industria y Competitividad* (TRA2017-84317-R-AR). Railway singularities are processed using time-frequency representations of the axle-box accelerations, that sequentially feed CNNs models for classification. Nevertheless, these predictive maintenance applications usually suffer from lack of samples for training the models. In the case of railway surveying, available singularities are limited to the extent of the railway system, and new examples cannot be generated in an straightforward way. The limitations on the amount of accesible data result on overffited models, that hardly generalize to unseen data. Chapter 6 is a key component of this project due to its contributions to deep-learning based railway crossing defects detection using an small number of training examples, under the few-shot learning paradigm, to alleviate the aforementioned limitations.

- *BraTS* – Brain tumor segmentation challenge. This competition seeks to promote the development of computational methods for the segmentation of gliomas in brain magnetic resonance imaging. This public challenge brings together one of the largest databases in the field, and has become very popular in the scientific community in recent years. The problem of lesion segmentation on MRI images on different organs is one of the main topics on medical image analysis. In this context, creating the dataset involves radiologists to assign a category to each voxel of the image. In addition, brain lesions experience large intraclass variations, which could not be captured during training but on very large datasets. This makes that, in a fully-supervised setting, deep models might have difficulties when learning from such class-imbalanced training sets. Thus, considering the scarcity and the diversity of target objects in these scenarios, lesion segmentation is typically modeled as an anomaly localization task, which is trained in an unsupervised manner. In particular, the training dataset contains only normal images and abnormal images (i.e. with lesions) are not accessible during training. Chapter 7 contributes to *BraTS* challenge under this paradigm, so called unsupervised anomaly segmentation, which has been a less explored solution than the standard supervised scenario.

1.4 Main contributions

This thesis incorporate outstanding contributions to the computer vision and deep learning community, which are detailed below.

1.4.1 Contributions to weakly supervised learning

Weakly supervised learning aims to leverage location information of objects using only global labels during training. This paradigm includes two related but substantively distinct tasks: weakly supervised semantic segmentation (WSSS) and multiple instance learning (MIL). MIL works group the data on bags of instances (images), and only bag-level labels are known during training. In this setting, instances are independent one to each other, and the global label is positive if one of the instances belongs to the given category. In the case of WSSS, the instances are the pixels of the image, whereby instances are correlated each other forming image patterns, and all are processed by a CNN together.

Weakly supervised semantic segmentation

Latest WSSS strategies using CNNs optimize the networks in an standard supervised way via global labels. In this setting, spatial features are extracted, and pooled into an uni-dimensional feature vector (embedding-based) using standard global average pooling, that serves as embedding for image classification. Then, segmentation maps are obtained using spatial intermediate activations [36] or gradient-weighted class-specific activation maps (Grad-CAMs) [40]. Recent works have focused on regularizing those attention maps during training to incorporate consistency losses [41], equivariant matching [42], or subcategory explorations [39]. Other less-popular choices include pixel-level classification layers into the CNNs (instance-based), which are lately pooled into a global classification using costume aggregation methods such as WILDCAT [43]. Finally, both kind of methods aggregate class-level segmentation masks into complementary semantic segmentation mask using costume post-processing pipelines, that include the background class for low-activated pixels [44]. In this thesis, we study the multi-label WSSS setting in the context of histology image segmentation in Chapter 3. Concretely, we propose an instance-based architecture that (i) does not require complex post-processing to aggregate class-wise attention maps, and (ii) uses log-sum-exponential pooling [45] to incorporate the concept of object size into training, which is optimized using only global labels.

Multiple instance learning

Regarding the MIL methods, as previously indicated, instances are composed of entire images belonging to the same group. In the embedding-based perspective, pooling operations such as mean, maximum, attention-weigthed [46], or RNN [47] produce a bag-level representation that serves to produce a global classification. Nevertheless, since instances do not present spatial dependence, activation-based methods are not applicable to leverage instance-level classifications. For this reason, we focus on this thesis in instance-based MIL methods, in the context of gigapixel histology WSI classification. In this application, each WSI is considered a bag, and extracted patches constitute instances. In Chapter 4 we propose a self-supervised Teacher-Student training framework to leverage instance-level labels. In particular, an instance-level MIL CNN using max-pooling is trained in a first stage as Teacher model. Thanks to the max-pooling properties, high precision hard pseudolabels are extracted from instances, which are further refined using the known global labels. Then, we propose to train a noisy Student [18] on pseudolabels in a standard supervised manner. Finally, in Chapter 5 we propose to include prior knowledge, in the form of class proportions, using constraint modifications of the standard MIL setting. In particular, we propose to use inequality constraints via log-barrier extensions [48] to (i) palliate the effect of max-pooling on Teacher model and promote the classification of instances in positive bags, and (ii) incorporate relative class proportion constraints, in the form of proportion ordering in the bags. Our formulation substantially differs from previous constraint formulations on weak supervision that incorporate the target size of the object in the image [49]. On the contrary, our inequality formulation only requires relative information (i.e. the primary and secondary categories), which is much more feasible to obtain.

Prostate histology diagnosis

In what refers to the prostate biopsy automatic analysis, this thesis also brings noticeable contributions. Concretely, in collaboration with pathologists of Hospital Clínico of Valencia, we have prepared and released a large public dataset containing both global biopsy-level labels and pixel-level annotations (see Chapter 2 and Chapter 5). Also, under the weakly supervised perspective, the proposed methods reach a performance around $\sim$ 0.80 of quadratic Cohen's kappa for tumor grading, which is similar to the inter-pathologist variability (see Chapter 4). The obtained results are consistent, and tested on different external datasets.

1.4.2 Contributions to few-shot learning

Few-shot learning

Few-shot learning (FSL) aims to train deep learning models able to generalize using only few samples from each category during training. In a tight formulation, the objective is to train a model capable of making predictions that can be generalized to new classes, of which few examples (K-shots) are given during inference. Nevertheless, on real-world applications, all classes are required to be used for training and testing, or they simply are binary scenarios. Still, methods proposed on the few-shot learning paradigm tend also to generalize best on standard supervised scenarios trained on very small data. Outstanding approaches on FSL are organized into deep distance metric learning using embedding matching [37] or relational networks [50], or memory-based methods via prototypical networks [38, 51]. A popular way to train these methods is to use the episodic training procedure, where training examples are divided between queries and support samples to simulate the inference setting. However, a recent work suggests that this form of training may not be optimal [52]. In Chapter 6 of this thesis, we propose (i) to use contrastive learning [53] to pull together samples belonging to the same class in an unity hyper-sphere hyper-plane, bypassing the episodic prototypical procure, in low-data scenario. Then, we propose to use the original latent representation to discern between classes using a prototypical l2-based distance.

Automatic railway crossing surveying

The proposed few-shot learning methods presented in Chapter 6 are applied in the context of automatic railway crossing surveying, in a binary classification scenario to detect deterioration patterns. As input to the model, spectrogramas from acceleration signals of railway crossings are used. With the proposed pipeline, we outperform previous literature on this field by accuracy gains of $\sim 8\%$. In addition, extensive ablation experiments for using CNNs in this application are presented, which will contribute to further research on this field.

1.4.3 Contributions to unsupervised anomaly segmentation

Unsupervised anomaly segmentation

Unsupervised anomaly segmentation aims to train deep learning models on normal data, able to identify abnormal pixels on test images, containing, for example, lesions on medical images [54], defects in industrial images [55] or abnormal events in videos [56]. The main core of literature on this field is focused on training constrained CNNs on normal data, under the hypothesis that anomalies will not fit the imposed constraint. In particular, generative methods such as autoencoders are a popular choice. In this case, the CNN is constrained to create a latent representation of the input image, such that the decoder is able to make a precise reconstruction of normal samples. Then, anomalies are located in the reconstructed image in the pixels that differ from the original input. This formulation is regularized in different ways, using variational latent space in VAEs [57] or context augmentation [58], or feature matching by incorporating a discriminator to the decoder output [59]. Still, this residual-based anomaly detection depends on the decoder performance to generate normal images, which is usually limited. For this reason, very recent methods have proposed to constrain intermediate attention maps of the encoder using Grad-CAMs [40]. In particular, [60] incorporates a size constraint into attention maps to force all pixels to be fully activated using an l1 penalty, and [61] uses disentanglement regularization [61]. In this thesis, Chapter 7 further along this line of research. In particular, we propose to relax the pixel-level constraint in [60] by (i) applying a size constraint at the image level, and (ii) to include a margin term, that allows using inequality constraints via log-barrier extensions [48], instead of penalty terms. Although this formulation brings substantial results, log-barrier extensions require several hyper-parameters to be optimized. For this reason, we also study an alternative formulation based on solely activation maps (instead of Grad-CAMs), and (iii) a regularization term that maximizes the Shannon entropy of attention maps distribution to force the CNN to be activated homogeneously in the whole image. Finally, anomalies are located into these attention maps on the pixels that differ from the homogeneous distribution.

Brain lesion segmentation

The proposed unsupervised anomaly segmentation methods are validated in the context of magnetic resonance imaging (MRI) brain tumour segmentation. In the popular BraTS dataset, our formulation brings outstanding improvements of nearly $\sim 25\%$ in terms of DICE compared to previous literature. In addition, as discussed in the ablation experiments in Chapter 7, the constraint formulation used offers good performance even without accessing to anomalous examples to set the threshold on anomaly scores, in contrast to previous literature. Finally, the method is satisfactorily validated on Physionet-ICH dataset for unsupervised intracranial hemorrhage (ICH) localization on CT scans, which shows its generalization capabilities to other image modalities and lesions.

1.5 Outline

This thesis is divided into 8 chapters. The current chapter introduces the motivation behind the research involved in this thesis, the proposed objectives and the main contributions. Subsequently, this chapter also details the framework and the thesis outline.

Chapter 2 corresponds to the paper: "Going Deeper through the Gleason Scoring Scale: An Automatic end-to-end System for Histology Prostate Grading and Cribriform Pattern Detection" [62]. It was published in the journal *Computer Methods and Prgorams in Biomedicine* (CMPB) belonging to the editorial *ELSEVIER*. CMPB journal had an impact factor of 5.428 when the article was published in 2020, and an h5-index of 79. The best rank was in the category *computer science, theory & methods* with a percentile of 88.64 (Q1).

Chapter 3 corresponds to the paper: "WeGleNet: A weakly-supervised convolutional neural network for the semantic segmentation of Gleason grades in prostate histology images" [63]. It was published in the journal *Computer Medical Imaging and Graphics* (CMIG) belonging to the editorial *ELSEVIER*. CMIG journal had an impact factor of 7.422 when the article was published in 2021, and an h5-index of 45. The best rank was in the category *radiology, nuclear medicine & medical imaging* with a percentile of 90.07 (Q1).

Chapter 4 corresponds to the paper: "Self-learning for weakly supervised Gleason grading of local patterns" [64]. It was published in the journal *IEEE Journal of Biomedical and Health Informatics* (JBHI) belonging to the editorial *IEEE*. JBHI journal had an impact factor of 7.021 when the article was published in 2021, and an h5-index of 80. The best rank was in the category *mathematical & computational biology* with a percentile of 93.86 (Q1).

Chapter 5 corresponds to the paper: "Proportion constrained weakly supervised histopathology image classification". It was published in the journal *Computers in Biology and Medicine* (CIBM) belonging to the editorial *ELSEVIER*. The paper was published in 2022, but the following publication details correspond to 2021, as the most recent journal indexes date from that year. CIBM journal had an impact factor of 6.698, and an h5-index of 76. The best rank was in the category *mathematical & computational biology* with a percentile of 90.35 (Q1).

Chapter 6 corresponds to the paper: "Supervised contrastive learning-guided prototypes on axle-box accelerations for railway crossing inspections". It was published in the journal *Expert Systems with Applications* (ESWA) belonging to the editorial *ELSEVIER*. The paper was published in 2022, but the following publication details correspond to 2021, as the most recent journal indexes date from that year. ESWA journal had an impact factor of 8.665, and an h5-index of 132. The best rank was in the category *engineering, electrical & electronic* with a percentile of 91.85 (Q1).

Chapter 7 corresponds to the paper: "Constrained unsupervised anomaly segmentation". It was published in the journal *Medical Image Analysis* (MedIA) belonging to the editorial *ELSEVIER*. The paper was published in 2022, but the following publication details correspond to 2021, as the most recent journal indexes date from that year. MedIA journal had an impact factor of 13.828, and an h5-index of 90. The best rank was in the category *radiology, nuclear medicine & medical imaging* with a percentile of 98.16 (Q1).

Note that Chapters 2, 3, 4, 5, 6 and 7 are based on the same structure. First, they present an abstract followed by an introduction containing the computer vision application and the motivation behind the conducted research. Next, the related works section contains a review of relevant previous literature in the field. Then, the proposed methods are detailed, followed by the experimental setting description (i.e. datasets, metrics, implementation details, and baselines). In that follows, the results, comparison with previous literature and ablation experiments are presented. Finally, the conclusions summarize the main findings in each chapter.

In Chapter 8, we relate the findings from each paper with the global aim of this PhD thesis. We also collect final remarks from a global perspective and suggest future research lines. Then, in Merits, we include journal publications, national and international conferences, as well as research awards derived from this thesis. Finally, we display the Bibliography.

Chapter 2

Going deeper through the Gleason scoring scale: An automatic end-to-end System for histology prostate grading and cribriform pattern detection

The content of this chapter corresponds to the author version of the following published paper: Silva-Rodríguez, J., Colomer, A., Sales, M.A, Molina, M., & Naranjo, V. 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, 195 (2020).

Contents

Going deeper through the Gleason scoring scale: An automatic end-to-end System for histology prostate grading and cribriform pattern detection

Julio Silva-Rodríguez[1], Adrián Colomer[2], Maria A. Sales[3], Rafael Molina[4] and Valery Naranjo[2]

[1]Institute of Transport and Territory, Universitat Politècnica de València, Valencia, Spain; [2] Institute for Research and Innovation in Bioengineering, Universitat Politècnica de València, Valencia, Spain; [3]Anatomical Pathology Service, University Clinical Hospital of Valencia, Valencia, Spain; [4]Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain

Abstract

Prostate cancer is one of the most common diseases affecting men worldwide. The Gleason scoring system is the primary diagnostic and prognostic tool for prostate cancer. Furthermore, recent reports indicate that the presence of patterns of the Gleason scale such as the cribriform pattern may also correlate with a worse prognosis compared to other patterns belonging to the Gleason grade 4. Current clinical guidelines have indicated the convenience of highlight its presence during the analysis of biopsies. All these requirements suppose a great workload for the pathologist during the analysis of each sample, which is based on the pathologist's visual analysis of the morphology and organisation of the glands in the tissue, a time-consuming and subjective task. In recent years, with the development of digitisation devices, the use of computer vision techniques for the analysis of biopsies has increased. However, to the best of the authors' knowledge, the development of algorithms to automatically detect individual cribriform patterns belonging to Gleason grade 4 has not yet been studied in the literature. The objective of the

work presented in this paper is to develop a deep-learning-based system able to support pathologists in the daily analysis of prostate biopsies. This analysis must include the Gleason grading of local structures, the detection of cribriform patterns, and the Gleason scoring of the whole biopsy. The methodological core of this work is a patch-wise predictive model based on convolutional neural networks able to determine the presence of cancerous patterns based on the Gleason grading system. In particular, we train from scratch a simple self-design architecture with three filters and a top model with global-max pooling. The cribriform pattern is detected by retraining the set of filters of the last convolutional layer in the network. Subsequently, a biopsy-level prediction map is reconstructed by bi-linear interpolation of the patch-level prediction of the Gleason grades. In addition, from the reconstructed prediction map, we compute the percentage of each Gleason grade in the tissue to feed a multi-layer perceptron which provides a biopsy-level score. In our SICAPv2 database, composed of 182 annotated whole slide images, we obtained a Cohen's quadratic kappa of 0.77 in the test set for the patch-level Gleason grading with the proposed architecture trained from scratch. Our results outperform previous ones reported in the literature. Furthermore, this model reaches the level of fine-tuned state-of-the-art architectures in a patient-based four groups cross validation. In the cribriform pattern detection task, we obtained an area under ROC curve of 0.82. Regarding the biopsy Gleason scoring, we achieved a quadratic Cohen's Kappa of 0.81 in the test subset. Shallow CNN architectures trained from scratch outperform current state-of-the-art methods for Gleason grades classification. Our proposed model is capable of characterising the different Gleason grades in prostate tissue by extracting low-level features through three basic blocks (i.e. convolutional layer + max pooling). The use of global-max pooling to reduce each activation map has shown to be a key factor for reducing complexity in the model and avoiding overfitting. Regarding the Gleason scoring of biopsies, a multi-layer perceptron has shown to better model the decision-making of pathologists than previous simpler models used in the literature.

2.1 Introduction

Worldwide, prostate cancer (PCa) is the second most common cancer in men, with 1.3 million new patients in 2018 [65]. According to the World Health Organisation, the yearly number of new cases will increase by more than 40% in this decade [66]. The main tool to diagnose PCa, once clinical explorations or blood test suggest its presence, is the prostate biopsy. Small portions of the tissue are extracted with a needle, laminated, stained with Hematoxylin and Eosin (H&E) and finally stored in crystal. Then, the sample is analysed under the microscope by the pathologist, determining the presence and grade of cancerous patterns depending on the morphology and organisation of the glands, nuclei and lumen using the Gleason grading system [67]. In this system, different cancer patterns in the tissue are grouped in different grades according to the prognosis of the cancer. In particular, for two-dimensional tissue slides, the Gleason grades (GG) range from 3 to 5, correlating inversely with the degree of gland differentiation of the tissue. The Gleason grade 3 (GG3) includes atrophic well differentiated and dense glandular regions. The GG4 contains cribriform, ill-formed, large-fused and papillary glandular patterns. Finally, GG5 includes isolated cells or file of cells, nests of cells without lumina formation and pseudo-roseting patterns. Examples of patterns belonging to different grades are presented in Figure 2.1.

Pathologists classify by visual inspection the tissue regions, detecting the presence of one or more Gleason patterns and, finally, diagnose the combined Gleason score according to the most prominent grades (e.g. the combined grade $5+4=9$ would be assigned to a sample in which the main cancerous Gleason grade is 5 followed by the grade 4). Therefore, the combined Gleason score ranges from 6 to 10, and it is assigned to the whole biopsy. This score is currently the best marker of prostate cancer prognosis and it defines the treatment to apply [68]. However, the Gleason scoring of histological prostate biopsies is a high time-consuming and repetitive task, which has intra and inter pathologist variability. Moreover, after the last International Society of Urological Pathology (ISUP) Consensus Conference in 2014 [69], new guidelines have been included that increase the pathologists' workload. In particular, it is

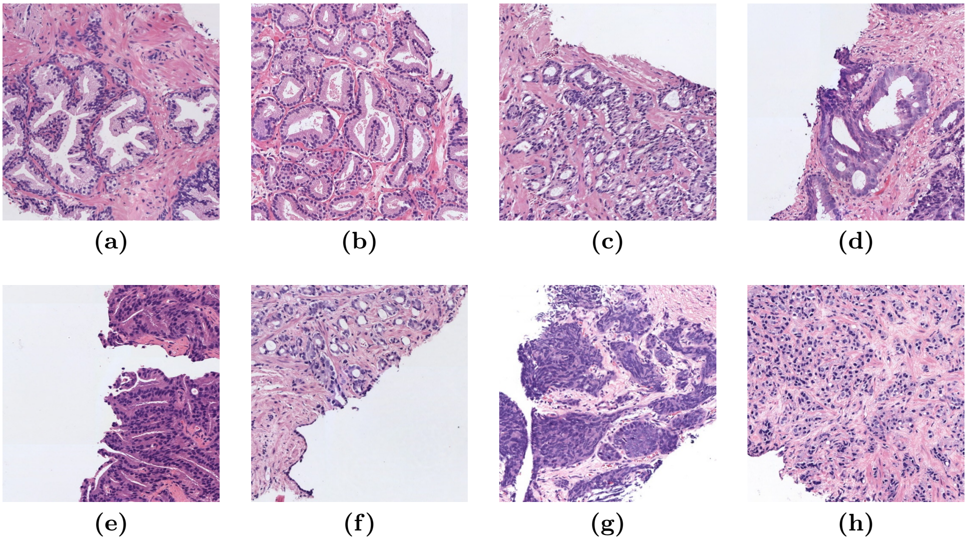


Figure 2.1: Patches of H&E histology samples presenting different Gleason patterns. (a): Non-cancerous well-differentiated glands; (b): Gleason grade 3 containing atrophic dense patterns; (c): Gleason grade 4 containing large fused glandular patterns; (d): Gleason grade 4 containing cribriform patterns; (e): Gleason grade 4 containing papillary structures; (f): Gleason grade 4 containing individual poorly-formed glands; (g): Gleason grade 5 including nests of cells without lumen formation; (g): Gleason grade 5 containing files of isolated cells.

recommended to also report the percentage of Gleason grade 4 in the sample, mainly for regions scored as $3+4=7$, where a higher percentage of Gleason grade 4 indicates the convenience of an earlier treatment [70], and the presence of cribriform glandular patterns, which indicate worse prognosis than the presence of other Gleason grade 4 patterns [71, 72]. Computer-Aided Diagnosis systems (CAD) support the work of pathologists and increase the objectivity in the this process. These are based on the digitisation of the histological crystals, obtaining whole slide images (WSIs) and developing computer vision algorithms to detect the cancerous regions inside the biopsy (or WSI).

The objective of this work is to develop an automatic Computer-Aided Diagnosis system working on WSIs and able to support pathologists in the analysis of the biopsy during the diagnosis process. The tasks of this analysis, to be included in the pathologists' report, are:

- Detection of the cancerous regions in the tissue according to the Gleason grading system.
- Detection of cribriform patterns.
- Calculation of the percentage of each Gleason grade in the biopsy.
- Gleason scoring of the whole biopsy, taking into account not only the grade proportion but also its severity.

This work is developed using our collected database SICAPv2, the largest public database of prostate biopsies with pixel-level annotations of Gleason grades, specifying the presence of cribriform patterns. In the following lines, we summarise the main contributions of this paper. The different blocks of our system are presented in Figure 2.2. First, we develop a patch-level predictor of Gleason grades with a carefully-designed CNN architecture trained from scratch. This architecture is based on three convolutional blocks and global-max pooling after the last block. With this model, we outperform, for the first time in the literature, the fine-tunning well-known state of the art architectures. Then, we discuss the model interpretability by means of the Class Activation Maps (CAMs) technique. Once the patches are classified, the trained architecture is fine-tuned to detect the presence of cribriform glandular structures for those images with Gleason grade 4. To the best of the authors' knowledge, no study has addressed this clinical need previously. Then, the WSIs are reconstructed in probability maps and the class (i.e. non cancerous, Gleason grade 3, 4 or 5) with the highest probability is assigned to each pixel. Once the percentages of each Gleason grade in the WSI are obtained, we developed a model, based on a multi-layer perceptron architecture, to predict the combined Gleason score to the whole biopsy. The obtained results show the good performance of this model which outperforms the previous state-of-the-art methods.

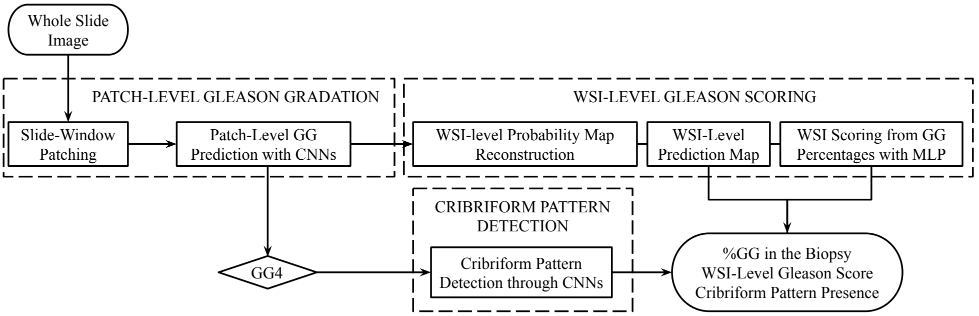

Whole Slide Image
PATCH-LEVEL GLEASON GRADATION
Slide-Window Patching
Patch-Level GG Prediction with CNNs
WSI-LEVEL GLEASON SCORING
WSI-level Probability Map Reconstruction
WSI-Level Prediction Map
WSI Scoring from GG Percentages with MLP
CRIBRIFORM PATTERN DETECTION
GG4
Cribriform Pattern Detection through CNNs
%GG in the Biopsy
WSI-Level Gleason Score
Cribriform Pattern Presence

Figure 2.2: Flowchart in which the different blocks of our system are presented. Taking as input a prostate whole slide image (WSI), the system performs a patch-level Gleason grade prediction through convolutional neural networks. If one patch is classified as Gleason grade 4 (GG4), a cribriform pattern detection is carried out by fine-tuning the model of the previous stage. Finally, the regions in the WSI are reconstructed and a pixel-level Gleason grade assignement is carried out. The WSI-level Gleason scoring is performed with a multi-layer perceptron taking as input the percentage of the Gleason grades in that region.

2.2 Related work

2.2.1 Computer vision in prostate cancer histology

Computer vision algorithms have been widely used to analyse histological PCa images. This section summarises the works previously presented in the CADs literature for prostate cancer detection, classifying them according to three factors: the kind of images included in the analysed database, the objectives addressed by CAD systems, and the techniques proposed to achieve them.

Regarding the images, mainly three types of histological images have been used: WSIs, prostactetomies and Tissue Micro Arrays (TMAs). TMAs are clusters of representative tumor areas extracted manually by pathologists [73]. TMAs are used for testing new techniques in a large number of different tumour samples. One of the main limitations of TMAs lies in the small amount of tissue that can be included in each samples, which may not be representative of the whole tumor region in epithelial tumors with heterogeneous patterns [74]. This is the case of prostate cancer, which has different patterns for each Gleason grade, as previously mentioned. Non-cancerous patterns that could confuse CAD systems, as the inflamed tissue or benign multi-nucleation, could be lost using TMAs. Thus, the strategy based on TMA analysis is not used in clinical practice [75] and it is more convenient to develop CAD systems based on raw WSI analysis. A model trained using large databases of WSIs could be used for both WSIs and prostactetomies. The works in [76–82] follow the strategy of WSI analysis, while in [83–85] the authors use TMAs to develop the CAD models.

With regard to the objectives to be addressed, some works focus just on the detection of prostate cancer against non-cancerous tissue [77, 81] or on the first-stage prostate cancer detection [86]. A full analysis of Gleason grades from 3 to 5 is usually limited by the size of the collected database, and the low prevalence of Gleason grade 5. Due to that, numerous researchers classify differentiating among non-cancerous samples, low grade (Gleason grade 3), and high Grade (Gleason grade ≥ 4) [82, 87, 88] or among non-cancerous, Gleason grade 3, and Gleason grade 4 [76, 79]. The most recent works tried to predict the full Gleason grading (Benign - Grade 3 - Grade 4 - Grade 5) in [83–85] but only using TMAs cores. To the best of the authors's knowledge, works analysing deeper the Gleason grades, this is, focusing on the automatic detection of individual patterns of a Gleason grade (i.e. cribriform pattern, which belongs to the Gleason grade 4 group) do not exist. This work represents an attempt in this direction.

Finally, concerning the techniques used to deal with the different mentioned objectives, the most common approach to analysed both is to perform a patch-based strategy (see Figure 2.3). The motivation for using this strategy is the large size of both TMAs and, especially WSIs, together with hardware limitations.

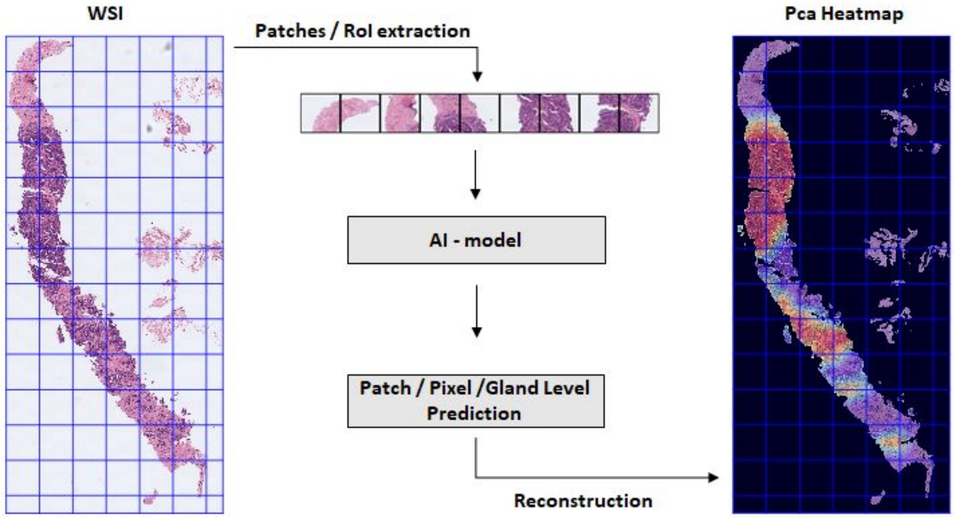


Figure 2.3: General workflow for high resolution histology slides processing.

2.2.2 Patch-level Gleason grading

Below, we will focus only on the description of the different techniques used, until now, for the patch-level Gleason grading. In the literature we can find approaches based on classic machine learning techniques with a hand-crafted feature extraction and deep learning algorithms (automatic feature extraction) by means of convolutional neural networks (CNN). In Nir et al. (2018) [84] a comparison between both approaches is carried out with a database of 333 cores of TMAs. Glands and nuclei are segmented to obtain features related to their size, intensity distributions and number of elements in each patch at different resolutions. Those are combined with full patch-level features related to the colour distribution and SURF descriptors to fit different machine learning models as linear discriminant analysis, linear regression, support vector machines, and random forests. Those models are compared with a U-Net CNN. The best result reported is a Cohen's quadratic kappa (κ) overall agreement measure of 0.51 obtained by the linear regression model. Nevertheless, in a later publication by Nir et al. (2019) [85] a κ of 0.60 was obtained by fine-tuning the CNN architecture MobileNet. In Arvaniti et al. (2018) [83] a larger database is used, with 886 cores. The patch-level grading is addressed through fine-tuning different CNN architectures such as VGG16, InceptionV3, ResNet50, DenseNet121, and MobileNet. The best results are reported with the last one, achieving a κ of 0.67 in the training set and 0.55 in the test one.

2.2.3 Biopsy scoring

Regarding the classification of the Gleason score for the whole biopsy (whole slide image), only a few works have addressed it, and only using TMAs. The common strategy used is to obtain the percentage of each grade in the analysed image and to assign the first and second components above a threshold as primary and secondary grades respectively. In Arvaniti et al. (2018) [83] the full Gleason scoring, using TMAs, is addressed, archiving κ of 0.75. Unfortunately, this simple model did not perform for extreme cases, for example $5 + 5 = 10$. In this case, a precision of 0.10 is reported in this work. In addition, the primary and secondary grades are not just related to the proportion of the different grades in the tissue, but also to the severity of each grade (e.g. GG5 could be diagnosed as secondary grade even having less proportion than GG4 or GG3 in the tissue).

2.3 Methods

2.3.1 Patch-level Gleason grading

The patch-level classification in the different Gleason grades is carried out by means of convolutional neural networks. We propose a self-designed base-model architecture (from now on called $FSConv$) which consists of a simple convolutional architecture with three convolutional layers and dimensional reduction operation employing max-pooling layers (Table 2.1).

After the automatic feature extraction blocks (base model), we introduce as top model a global-max-pooling layer. To show the superior performance of this architecture, different configurations already applied

Layer Name	**Filter Size**	**Stride**	**Activation**	**Output Shape**	**Connected to**
$Input$	$-$	$-$	$-$	$(224, 224, 3)$	$-$
$Conv_1$	$(3, 3, 32)$	1	$ReLU$	$(224, 224, 32)$	$Input$
$Max - Pooling_1$	$(2, 2)$	2	$-$	$(112, 112, 32)$	$Conv_1$
$Conv_2$	$(3, 3, 124)$	1	$ReLU$	$(112, 112, 124)$	$Max - Pooling_1$
$Max - Pooling_2$	$(2, 2)$	2	$-$	$(56, 56, 124)$	$Conv_2$
$Conv_3$	$(3, 3, 512)$	1	$ReLU$	$(56, 56, 512)$	$Max - Pooling_2$
$Max - Pooling_3$	$(2, 2)$	2	$-$	$(28, 28, 512)$	$Conv_3$

Table 2.1: *FSConv* architecture description. It consists of three blocks with convolutional filters, ReLU activation and max-pooling operation.

in the literature to the same problem, have been also tested as top models and are described next.

One of the main approaches is the flattening of the activation volume resulting from the final convolutional block and the class prediction through consecutive fully-connected layers. In this case, overfitting is addressed by means of a random dropout of a percentage of the neurons in each training iteration. Nevertheless, these top-model architectures include a large number of parameters to optimise, increasing the complexity of the model, and they are sensitive to the location of the structures in the image. This problem is usually dealt with data augmentation techniques, applying, for example, random rotations and translations to the images. Other approaches propose the convenience of using global-average pooling on the last feature maps as regulariser to make the model translation-invariant and decrease its complexity [89]. This technique is used in [83] for the prediction of prostate cancer Gleason degree with fine-tuned models. Due to the use of a patch-based strategy with sliding window, the location and amount of the cancerous structures in the image is not controlled. Thus, as shown in Figure 2.4, some patches could have small portions of cancerous tissue. The global-average pooling layer takes into account the information in the whole activation map, and in those cases, the output of the filter that detects this pattern could be diminished. To make the models robust to the amount and location of cancerous tissue, we propose in this work the use of the global-max-pooling layer to play the role of the global-average pooling. All different configurations, fully-connected layer with ReLU activation and dropout regularisation (FC), global-average-pooling (GAP) and global-max-pooling (GMP) layers and their combinations are implemented and their performance is discussed in this work.

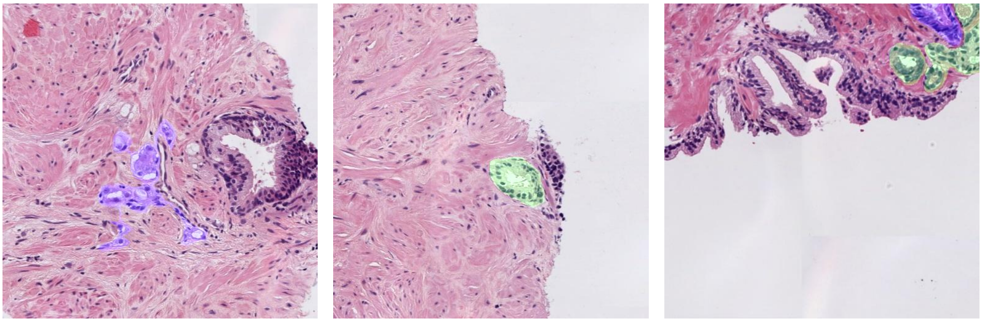


Figure 2.4: Patches with small amount of cancerous tissue. Green: GG3, Blue: GG4.

For comparison, together with the proposed architecture trained from scratch, we fine-tuned several well-known architectures: VGG19 [90], ResNet-50 [91], InceptionV3 [92] and MobileNetV2 [93]. All of them were pre-trained in the Imagenet data set [94]. For the feature extraction stage, the base model from those pre-trained models is extracted and partially retrained. This strategy is usually used to transfer the knowledge obtained in extracting features from a large database to specific domains where the amount of data is limited. Nevertheless, the patterns of the images used during the training are very different from the histology ones. To keep just the low-level features (contours, combination of basic colours, general shapes, etc.) from the pre-trained models, the weights of just the first convolutional blocks are frozen, while the rest are re-trained to adapt the model to the specific application. The layer from which the freezing strategy is applied is empirically optimised for each model, and is specified in the experimental part of the paper, in Section 2.5.1.

The output layer for all the different configurations is composed of one neuron per class with soft-max activation function to obtain the final probability per class. In the training process, we use categorical cross-entropy as loss function, modified to deal with the class imbalance in the training set as follows:

$$L(\widehat{y}, y) = -\frac{1}{C}\sum_{c=1}^{C} w_c(y_c log(\widehat{y}_c)) \tag{2.1}$$

where y and $\widehat{y}$ contain the one-hot-encoded reference labels and predicted probabilities, respectively, of each class c for a certain instance. $w_c = (C \times N)/N_c$ is the weight applied to each class, being N the total number of images, N_c the number of images belonging to class c and C the number of classes, $C = 4$ in our case (non-cancerous, GG3, GG4 or GG5).

Stochastic Gradient Descend is applied as optimiser and the training procedure is performed using mini-batches. The values of learning rate and batch size are fixed empirically for each configuration and experiment, and they are specified in Section 2.5.1. Data augmentation techniques are used on the training set applying random rotations and translations to the images.

2.3.2 Cribriform pattern detection

The detection of cribriform structures in GG4 patches is also carried out using convolutional neural networks. Due to the complexity of the task and the reduced number of samples, we address this problem by fine-tuning the model trained for the Gleason grades prediction. To take advantage of the specialised features extracted by the proposed architecture, the model is re-trained, optimising the layer from which the filter weights should be frozen to avoid over fitting. The top model used here is also proposed in the Gleason grading problem (global-max-pooling layer) followed by a last layer with one neuron and sigmoid activation function. The loss function used is the binary cross-entropy. Again, Stochastic Gradient Descent is used as optimiser applied on mini-batches and including data augmentation with random rotations, translations and brightness variations.

2.3.3 Whole slide image Gleason scoring

To predict the Gleason score of the WSI, it is necessary to compute the tissue percentage of each Gleason grade present in the WSI. For that purpose, the first step is to apply the patch-level classification (Section 2.3.1). Then, for each pixel, the predicted probabilities for each class is estimated by bilinearly interpolating the predicted probabilities of the closest patches in terms of euclidean distance to the center of the patches. Thus, a probability map per class (i.e NC, GG3, GG4, and GG5) is obtained per each WSI. Finally, the percentage of each Gleason grade is calculated after assigning each pixel the class, c, with the highest probability.

The pathologist's decision making while assigning a Gleason score to a WSI takes into account both the percentage of each Gleason grade and the severity of each grade. To model this process, we propose to train a Multi-Layer Perceptron (MLP) to automatically predict the combined Gleason scoring of a biopsy, by means of a multi-class classification task. This task requires the prediction of both primary and secondary Gleason grades. To address it, MLP is selected as a suitable classifier, due to its flexibility to adapt the architecture to perform a multi-output classification. The proposed MLP architecture consists of a branch with two outputs (see Figure 2.5). The branch is composed of two fully-connected layers with 16 and 8 neurons respectively, and ReLU as activation function. The branch is then divided into two output layers: one for the primary Gleason grade and one for the secondary grade. These output layers are composed of four neurons each, one neuron per target class (i.e. NC, GG3, GG4 or GG5) and soft-max as activation function. The loss function used is the categorical cross-entropy.

2.4 Experimental setting

2.4.1 Materials: SICAP database

The database presented in this paper, SICAPv2, is, to the best of the authors's knowledge, the largest public collection of prostate H&E

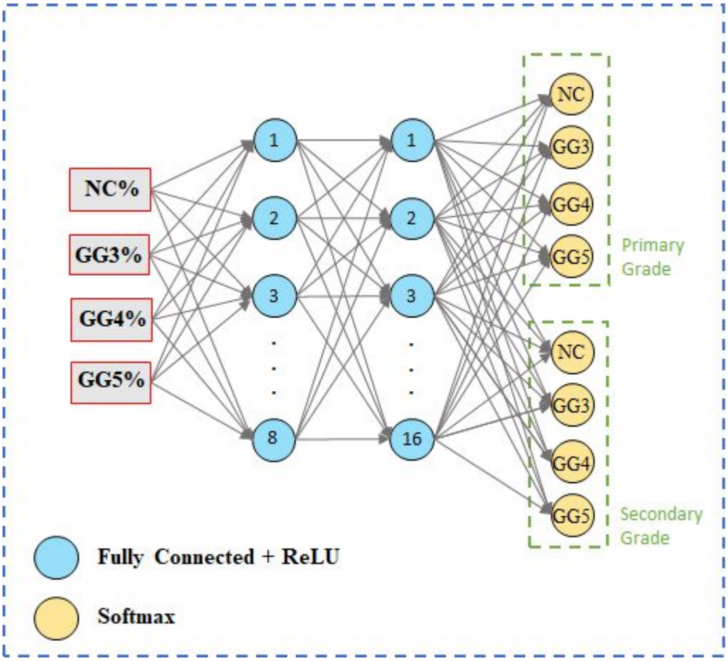


Figure 2.5: Proposed Multi-Layer Perceptron (MLP) for the whole slide image Gleason scoring. The model takes as input the percentage of each Gleason grade in the whole slide image, and is composed by a main branch with two fully-connected layers and two outputs. The intermediate layers consist of 8 and 16 neurons respectively and ReLU as activation function. The output layers present one neuron per target class and soft-max activation. NC: non cancerous, GG3: Gleason grade 3, GG4: Gleason grade 4, GG5: Gleason grade 5.

biopsies with local-level annotations of Gleason grades. SICAPv2 is an extension the database introduced in [81] and will be publicy available after the publication of this paper.

After analysing the literature, four main prostate cancer tissue image databases were found. The largest database with prostate biopsies was released by The Cancer Genome Atlas project[1] [95] with up to 720 prostate biopsy slides. Nevertheless, the lack of annotations at both the local and biopsy levels of the Gleason grades restricts the use of these data. The database shared by Arvaniti et al. [83] includes pixel-level

[1] https://portal.gdc.cancer.gov/

annotations of Gleason patterns from 886 small regions of slides (cores of TMAs). Unfortunately, as discussed earlier, those cores do not represent the heterogeneous patterns of local structures of prostate cancer and benign lesions, so they lack clinical relevance for the slide-level Gleason score diagnosis. Similar limitations are found in the recent database from the challenge Gleason19 in the MICCAI 2019 conference[2], with 331 cores annotated by different pathologists, and the dataset used in [80], composed by 625 isolated patches. Although those databases contribute to the validation of different algorithms, the lack of large databases with clinical reference of heterogeneous patterns has been a limiting factor for the scientific community to develop deep-learning-based methods which demand a large amount of data. One of the contributions of this work is the publication of a large database of WSIs containing biopsy-level labels (i.e. Gleason scores for each biopsy) and pixel-level Gleason grades annotations, in which for the first time, the presence of cribriform glandular regions is indicated.

SICAPv2 database includes 155 biopsies from 95 different patients who signed the pertinent informed consent. The tissue samples where sliced, stained and digitised using the Ventana iScan Coreo scanner at $40x$ magnification obtaining WSIs. The slides were analysed by a group of expert urogenital pathologists at Hospital Clínico of Valencia, and a combined Gleason score was assigned per biopsy. In cases where the grade was uncertain, the label was assigned by consensus of all expert pathologists to avoid inter-observer variability. The primary Gleason grade (GG) in each biopsy is distributed as follows: 36 non-cancerous regions, 40 samples with Gleason grade 3, 64 with Gleason grade 4 and 15 with Gleason grade 5 (henceforth NC, GG3, GG4, and GG5 respectively). Regarding the combined scores, the co-occurrence matrix of primary and secondary grades is shown in Figure 2.6.

The local cancerous patterns were annotated using an in-house software based on the OpenSeadragon libraries [96], following the Gleason scale and indicating the presence of cribriform glandular structures. In order to process the large WSIs, these were down-sampled to $10x$ resolution and divided into patches of size 512^2 and overlap of 50% between them. Those

[2] `https://gleason2019.grand-challenge.org/Home/`

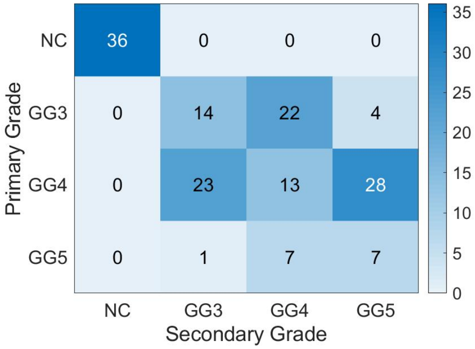


Figure 2.6: Description of the Gleason scores in the SICAPv2 database. Co-occurrence matrix of primary and secondary Gleason grades in each biopsy. NC: non cancerous, GG3: Gleason grade 3, GG4: Gleason grade 4 and GG5: Gleason grade 5.

values were previously optimised for the detection of cancerous patterns in [81]. A mask of the presence of tissue in the patches was obtained by applying the Otsu threshold method. To develop the model able to predict the main Gleason grade, patches with less than 20% of tissue were excluded. In addition, patches without cancerous patterns annotated by the pathologists belonging to cancerous biopsies where also discarded. After this procedure, the database contains 4417 non-cancerous patches, 1635 labelled as GG3, 3622 as GG4, and 665 as GG5. Note that if one patched contained more than one annotated grade, the majority grade was assigned as label. 763 GG4 patches also contain annotated cribriform glandular regions. A summary of the database description is presented in Table 2.2.

	Non cancerous	**Grade 3**	**Grade 4 (cribriform)**	**Grade 5**	**Total**
#WSIs	37	60	69 (36)	16	182
#Patches	4417	1636	3622 (763)	665	10340

Table 2.2: SICAPv2 database description. Amount of whole slide images and their respective biopsy-level primary label (first row) and number of patches of each one of the Gleason categories (second row).

In order to train the models and optimise the hyperparameters involved in this process, the database was divided following a cross-validation strategy. In particular, each patient was exclusively assigned to one fold with the aim of avoiding overestimation of the performance of the system [85] and ensuring its ability of generalisation. Thus, the database was divided into 5 groups containing approximately 20% of the patches each one. Notice that this process was carried out trying to guarantee the class balance character between sets. A summary of the resulting partition is presented in Table 2.3.

		Patients - Patches			
Group		Non Cancerous	GG3	GG4 (Cribriform)	GG5
Cross-validation	1	2 - 685	3 - 625	11 - 979 (237)	2 - 198
	2	1 - 717	4 - 346	10 - 950 (41)	2 - 153
	3	1 - 644	9 - 361	7 - 670 (126)	2 - 118
	4	1 - 1727	8 - 497	9 - 1042 (214)	2 - 247
Test		4 - 644	6 - 393	9 - 853 (145)	2 - 232

Table 2.3: Database partition description: number of patients-patches for each grade in each validation fold (4-fold cross-validation) and test subset.

Notice that four of the five sets were used to tune the hyper-parameters involved in the developed algorithms while the remaining partition was employed to test the final predictive system. For the evaluation of the patch-level Gleason grade prediction, a cross-validation strategy was used with the four validation cohorts, while for the WSI-level prediction of Gleason scores those sets were joined to apply a leave-one-out strategy per patient in training.

The data collected by Arvaniti et al. [83] was also utilised to validate the models produced in our study. The cores were resized to match the resolution used in our models and patched to the dimensions used in our database. By this approach, each one of these cores is approximately equivalent to one of our patches. Thus, 115 non-cancerous images, 274 patches labelled as GG3, 210 GG4, and 104 GG5 were used to validate our work in an external database. Also, the patches shared by Gerytch et al. [77] were used in our work for the validation of our proposed model. After normalisation of the images to match our methodology, 32 non-cancerous images, 95 patches labelled as GG3, 216 GG4, and 70 GG5 were obtained.

2.4.2 Metrics

In order to objectively evaluate the performance of the trained models the following metrics were used: accuracy, F1-score, and Cohen's quadratic kappa statistic. The accuracy (ACC) is defined as the percentage of samples correctly classified. Nevertheless, this metric does not provide information about the performance of the model for each class. This information was quantified by utilising the F1-score ($F1S$), a combination of precision and sensitivity per class computed as follows:

$$F1S_c = 2 \times \frac{precision_c \times sensitivity_c}{precision_c + sensitivity_c} \tag{2.2}$$

Cohewhere c indicates the predicted classes.

However, an automatic method should be less penalised when classifying a GG5 tissue as GG4 than as NC, even more so when taking into account the inter and intra-observer variability. In the literature, this fact is addressed using the Cohen's quadratic kappa (κ) metric [97]. The metric κ ranges from -1 to 1, being directly proportional to the level of agreement between observers (-1 no agreement, 1 total agreement). Although there is not objective interpretation of which are the reasonable values for κ in medical applications, recent proposals [98] define a moderate agreement if κ is higher than 0.6, while a strong agreement is stated when κ is higher than 0.8.

The patch-level Gleason grading models are evaluated using all the aforementioned figures of merit.

In order to evaluate the system for the detection of cribriform patterns, the area under the Receiver Operating Characteristic (ROC) curve (AUC) was used. In medical applications, a system is considered reliable if the AUC value exceeds 0.80 [99]. The predicted labels are obtained by thresholding the scores (cribriform if the probability is above 0.5), and then evaluated by means of ACC, sensitivity and specificity.

Regarding the evaluation of the WSI-level Gleason scoring, the Cohen's quadratic kappa was used.

2.5 Results

2.5.1 Patch-level Gleason grading

In the case of the patch-level Gleason grading model, in this section besides the obtained results using SICAPv2 database, we also discuss its performance in an external database.

After optimising the hyperparameters (learning rate, batch size, number of epochs, etc.), table 2.4 shows the obtained results in the validation sets for the proposed network *FSConv* with different top models: fully-connected layers (FC), global-max pooling (GMP), global-average pooling (GAP), or a combination of them (GAP+FC or GMP+FC). Table 2.4 also presents the results for the best tested fine-tuned architectures, VGG19 and RestNet, using the same top models as *FSConv*. The optimum hyperparameters were: learning rate of 0.01 for *FSConv* and 0.0001 for the finned-tunned networks, batch size of 32 images and 200 epochs in all cases. The base model of the fine-tuned networks were also optimised, being selected to freeze the first convolutional block for VGG19 and setting all layers as trainable for RestNet. Futhermore, Table 2.5 presents a comparison in terms of storage space (in kilobytes, KB) and number of trainable parameters of each architecture.

Experiment	**ACC**	**F1S**				**Avg-F1S**	κ
		NC	GG3	GG4	GG5		
VGG19+FC	0.721 ± 0.041	**0.887 ± 0.017**	0.663 ± 0.050	0.604 ± 0.169	0.520 ± 0.099	0.668 ± 0.065	0.734 ± 0.032
VGG19+GMP	0.721 ± 0.054	0.872 ± 0.020	0.648 ± 0.060	0.603 ± 0.167	**0.545 ± 0.094**	0.667 ± 0.076	0.717 ± 0.064
VGG19+GMP+FC	0.727 ± 0.042	0.886 ± 0.019	0.682 ± 0.063	0.609 ± 0.150	0.531 ± 0.082	0.677 ± 0.065	**0.747 ± 0.064**
VGG19+GAP	0.730 ± 0.046	0.881 ± 0.026	0.643 ± 0.096	0.653 ± 0.116	0.513 ± 0.084	0.672 ± 0.047	0.717 ± 0.073
VGG19+GAP+FC	0.724 ± 0.048	0.879 ± 0.013	0.690 ± 0.060	0.609 ± 0.154	0.521 ± 0.118	0.675 ± 0.072	0.717 ± 0.062
ResNet+FC	0.695 ± 0.031	0.838 ± 0.015	0.667 ± 0.075	0.572 ± 0.127	0.484 ± 0.053	0.640 ± 0.055	0.681 ± 0.046
ResNet+GMP	0.687 ± 0.038	0.836 ± 0.018	0.642 ± 0.072	0.556 ± 0.131	0.506 ± 0.073	0.635 ± 0.060	0.678 ± 0.033
ResNet+GMP+FC	0.699 ± 0.022	0.845 ± 0.013	0.674 ± 0.081	0.552 ± 0.123	0.492 ± 0.492	0.641 ± 0.044	0.689 ± 0.053
ResNet+GAP	0.696 ± 0.026	0.848 ± 0.013	0.677 ± 0.083	0.545 ± 0.124	0.501 ± 0.040	0.643 ± 0.055	0.692 ± 0.033
ResNet+GAP+FC	0.702 ± 0.028	0.847 ± 0.007	0.682 ± 0.089	0.555 ± 0.126	0.518 ± 0.052	0.650 ± 0.055	0.698 ± 0.042
FSConv+FC	0.733 ± 0.030	0.839 ± 0.043	0.650 ± 0.022	0.696 ± 0.060	0.544 ± 0.129	0.682 ± 0.020	0.680 ± 0.027
FSConv+GMP	**0.762 ± 0.007**	0.876 ± 0.016	**0.727 ± 0.022**	**0.709 ± 0.054**	0.536 ± 0.106	**0.712 ± 0.025**	0.732 ± 0.046
FSConv+GMP+FC	0.728 ± 0.061	0.872 ± 0.034	0.695 ± 0.037	0.631 ± 0.201	0.452 ± 0.037	0.663 ± 0.059	0.720 ± 0.040
FSConv+GAP	0.531 ± 0.088	0.683 ± 0.080	0.322 ± 0.240	0.441 ± 0.258	0.339 ± 0.183	0.446 ± 0.150	0.415 ± 0.237

Table 2.4: Results for patch-level Gleason grades prediction on the validation set. The performance of the different models ResNet, VGG19 and *FSConv* are presented with the different configurations of top models. The metrics presented are the accuracy (ACC), the F1-Score (FS1), computed per class and its average, and the Cohen's quadratic kappa (κ). GMP: global-max pooling, GAP: global-average pooling and FC: fully-connected layers.

Experiment	Storage (KB)	Parameters
VGG19+FC	180700	46203652
VGG19+GMP	78290	19987716
VGG19+GMP+FC	79832	20380676
VGG19+GAP	78289	19987716
VGG19+GAP+FC	79833	20380676
ResNet+FC	496022	126822916
ResNet+GMP	92579	23542788
ResNet+GMP+FC	97170	24716036
ResNet+GAP	92580	23542788
ResNet+GAP+FC	97179	24716036
FSConv+FC	104899	26846212
FSConv+GMP	2486	630276
FSConv+GMP+FC	4026	1023236
FSConv+GAP	2485	630276

Table 2.5: Number of parameters and memory usage of the different CNN architectures tested for the patch-level Gleason grading task. KB: kilobytes.

Regarding the results obtained in the fine-tuned models, the use of architectures with residual blocks provided slightly worse results than the sequential approach, similarly as the previous results reported in the literature where sequential models used to outperform residual ones [81, 83, 84]. In relation to the use of different top models, no differences were found in the accuracy of the fine-tuned architectures, observing similar results for all of them.

In relation to *FSConv* architecture, interesting results were obtained while testing the use of different top models. The best performing architecture to validate the system is the one with global-max pooling, *FSConv*+GMP. The outperforming of the global-max pooling compared to the fully-connected configuration could be explained by the reduction in the number of weights to be optimised (see Table 2.5), making the model simpler and more capable of generalising to new images, and by the invariance to the pattern location provided by the global-pooling operations. However, the *FSConv* model did no converge properly using global-average poling in the top model (*FSConv*+GAP), an effect non observed in the case of fine-tuned architectures. The explanation of this behaviour could be related to the receptive field of the model. The receptive field is defined as the region of the image involved in the cross-correlation operation resulting in one output element in the activation map. As *FSConv* is a shallow architecture, the final receptive field (i.e.

in the last convolution layer) is limited, and then the extracted features are more local than the obtained by deeper architectures. Then, if the pattern to be detected is just located in a small portion of the tissue, the activation could be masked in the global average. This effect is not present in deep networks with a large receptive field as the VGG19 or ResNet, and it could explain the similar behaviour of both top models for the pre-trained networks. Therefore, the use of top models based on global-max pooling in shallow architectures allows to extract relevant features to train models from scratch reducing the number of trainable parameters of the model and increasing its robustness against size and location variability of the region of interest.

Paying attention to Table 2.4 and taking into account all the figures of merit, we conclude that *FSConv*+GMP configuration is the best performing one for the patch-level Gleason grading. In the validation set used, this model outperforms the VGG19+GMP+FC architecture in terms of accuracy (0.7622 compared to 0.7273) and average F1-score (0.7125 against 0.6772). Furthermore, the *FSConv*+GMP model performs specially well when distinguishing between GG3 and GG4, the most difficult task in the pathologists' work, reaching F1-scores of 0.7277 and 0.7093 respectively (see Table 2.4). This is the first time in the literature that self-defined architectures trained from scratch outperform fine-tuned architectures from the state-of-the-art pre-trained in Imagenet for Gleason grading. Moreover, the reduced amount of parameters (2×10^7 in the VGG19+GMP+FC model against 6×10^5 in the *FSConv*+GMP model, see Table 2.5), makes more convenient the *FSConv* architecture for deployment. Thus, the model *FSConv*+GMP was trained using all the images in the cross-validation sets in order to evaluate its performance in the external test cohort.

The results of the proposed model for the test set and a comparison of them with previous state-of-the-art works are reported in Table 2.6. κ value increases up to 0.77 in the test subset for *FSConv*+GMP. In comparison with previous studies, our results outperform the state of the art, obtaining almost a strong agreement between our model and the pathologist, while just moderate agreement ($\kappa = 0.55$ [83]) was obtained previously in the test set. Figure 2.7 shows the performance evaluation of $FSConv$. In particular, the confusion matrix for validation and test subsets are presented. From this figure, it can be observed that most of the errors occur between adjacent classes.

Experiment		ACC	F1S				Avg-F1S	κ
			NC	GG3	GG4	GG5		
$FSConv$+GMP	Test	0.67	0.86	0.59	0.54	0.61	0.65	**0.77**
Arvaniti et al. [83]	Validation	-	-	-	-	-	-	0.67
	Test	-	-	-	-	-	-	0.55
Nir et al. [85]	Validation	-	-	-	-	-	-	0.61

Table 2.6: Results for the patch-level Gleason grading on the test set for the model $FSConv$+GMP and comparison with previous literature. The metrics presented are accuracy (ACC), F1-Score (1S), computed per class and its average, and Cohen's quadratic kappa (κ). Note that for the results reported in previous literature not all the metrics were reported. GMP: global-max pooling.

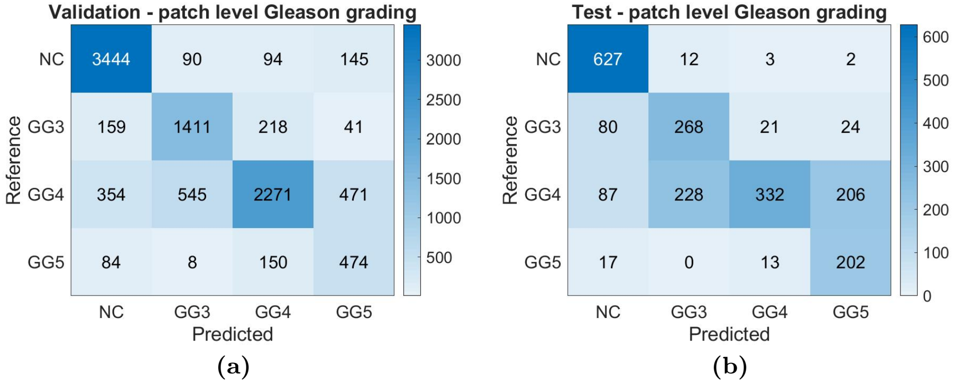


Figure 2.7: Confusion Matrix of the patch-level Gleason grades prediction done by $FSConv$ network in (a) validation set and (b) test set.

2.5.2 Model interpretation

One of the main drawbacks of deep learning models in medical practice is the lack of interpretability. This fact creates distrust in the clinicians, the final users of CAD systems. To deal with this problem, in this research we study the interpretability of the trained models by means of the Class Activation Maps technique (CAMs). Both VGG19+GMP+FC (the best fine-tuned model) and $FSConv$+GMP models are compared in this section using CAMs.

This technique was proposed in [100] as a procedure to obtain a heatmap indicating the regions of the input image to which the model is paying attention to predict certain class. CAMs for both models are obtained for images correctly classified (see Figure 2.8) and for images miss-classified by the VGG19 model (see Figure 2.9). These illustrations are organised as follows: the first row corresponds to the original patch, and the second and third rows show the CAMs for VGG19 and *FSConv* models, respectively. In Figure 2.8 each column shows an example per class: NC, GG3, GG4 and GG5 accordingly. The main difference in the results obtained by VGG19 and *FSConv* is the best differentiation between GG3 and GG4 by the second model (see Table 2.4), the most difficult task in the pathologists' work. In Figure 2.9 three of those cases are presented in each column: two cases predicted by the VGG19 as GG3 and one as GG5, respectively. Those cases were correctly classified as GG4 by *FSConv* model.

CAMs obtained for VGG19 in NC, GG3 and GG4 show that the model is basing the decision in glandular regions detected and classified correctly. In the case of GG5, the highlighted region presents a group of single cells and infiltrating cords without lumen formation, characteristic patterns of poor differentiate tissue in GG5. In the case of *FSConv* architecture, the CAM heatmap does not detect large regions, but small dots instead. Although the glandular regions are not detected, paying attention to the position where the dots are pointing at, we can extract interesting insights (see Figure 2.8). In the case of GG4, the map is activated in a small nest belonging to a fused-glands structure with irregular cribriform shape. Regarding the GG3 image, the dot indicates thick cytoplasm in different medium-sized tubular glands. In the image marked as GG5, the CAM highlights single isolated cells with hyperchromasia. Less interpretable is the CAM obtained in the NC image, where any gland is detected. We speculate that the model carries out this classification by dismissing the presence of cancerous patterns. Regarding the cases where VGG19 miss-classifies GG4 in Figure 2.9, a correct detection of the regions of interest is observed. However, these glandular regions are not correctly classified as GG4, while *FSConv* model does it just paying attention to closed lumens in small ill-formed glands. At this stage of understanding, we believe that this fact is the cause of the different performance by

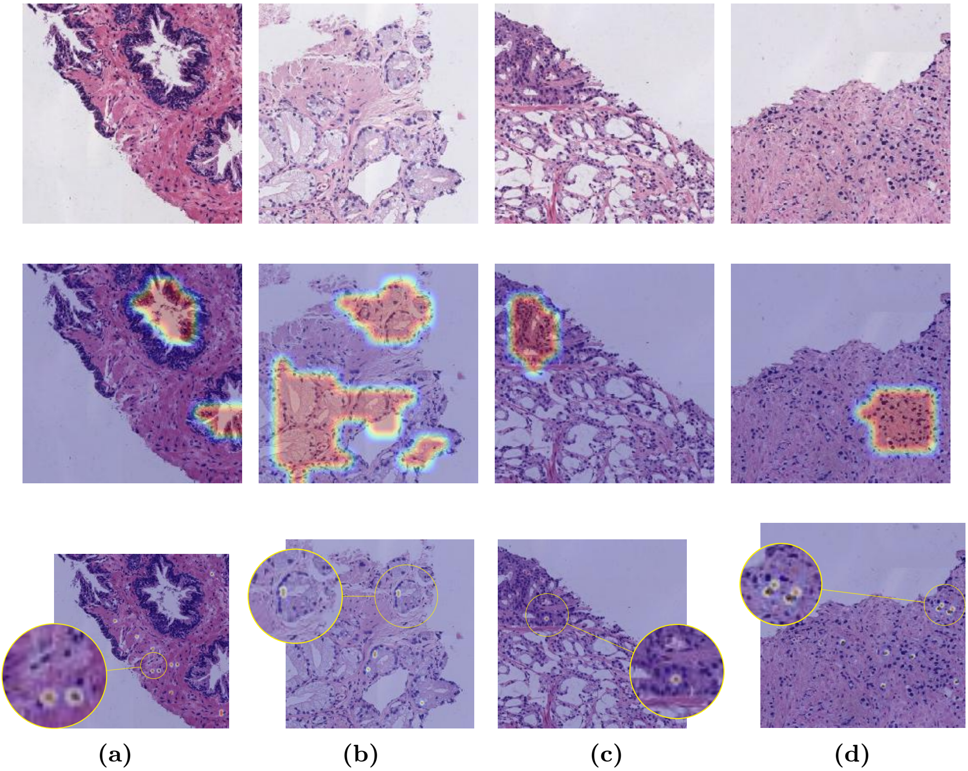


Figure 2.8: Original image (first row) and Class Activation Maps (CAMs) obtained by the VGG19 model (second row) and the *FSConv* network (third row) in four images correctly classified. Non-Cancerous (a), Gleason grade 3 (b), Gleason grade 4 (c) and Gleason grade 5 (d).

both models. VGG19 focuses the prostate cancer detection on detecting epithelial and glandular regions, and these structures present a larger heterogeneity than its basic components (colour and size of individual glands, diameter and opening degree of lumens in the glandular region, etc.). This could be the reason why the VGG19 generalises slightly worse than *FSConv*.

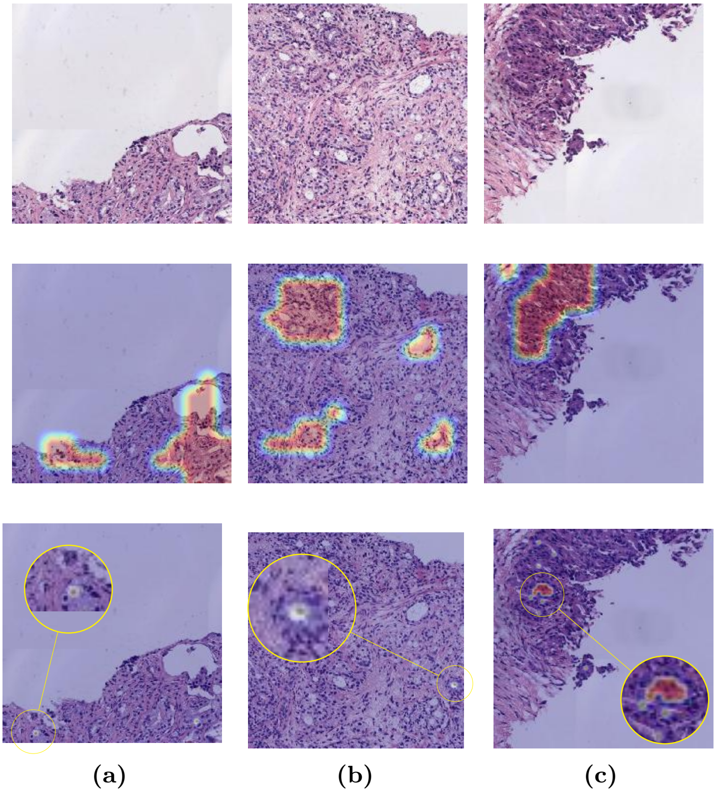
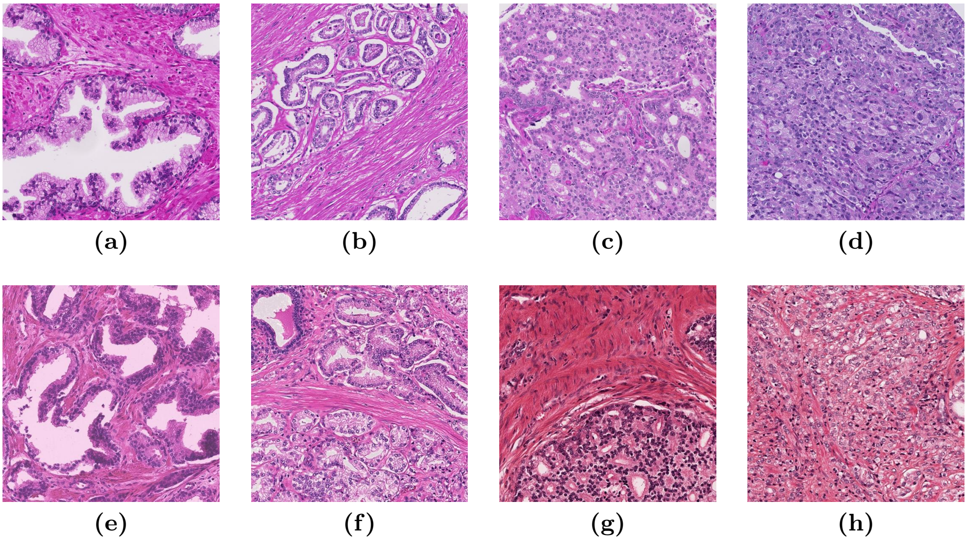


Figure 2.9: Original images (first row) and Class Activation Maps (CAMs) obtained on the VGG19 model (second row) and the *FSConv* network (third row) in images with GG4 correctly classified by the *FSConv*. The VGG19 model classification of those cases is GG3 in (a) and (b) and GG5 in (c).

2.5.3 Validation on external datasets

With the purpose of testing the generalization capability of the trianed model, *FSConv* net was validated on two external databases. The databases used were shared by Arvaniti et al. [83] and Gerytch et al. [77]. The first database is composed of 886 cores from Tissue-Micro Arrays digitised at 40× magnification, and the second has 625 patches of prostate histology images at 20× magnification. Each core was resized to 10× resolution and a central patch with dimensions 512^2 was extracted. For both databases, the ground truth was generated following the procedure in [83]. Non-cancerous patches were extracted from images with only benign structures annotated, labels GG3, GG4, and GG5 were assigned to patches with only the corresponding grade annotated. Examples of the obtained images from the Arvaniti et al. and Gerytch et al. databases are presented in the first and second rows of Figure 2.10, respectively. Note that the H&E stain color images are different from those appearing in the SICAPv2 database (see Figure 2.1 for examples of the images used to train the developed models). To normalise the colour distribution of the images in external databases, the method presented in [101] was used after applying a channel-wise histogram matching of the external images to a SICAPv2 database reference image. This image was selected by the expert pathologists involved in this work based on its structural and colour properties. Then, our best performing model, i.e. *FSConv*, was used to predict and evaluate our performance on the external databases. Table 2.7 and Figure 2.11 show the obtained figures of merit and confusion matrices, respectively.


(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 2.10: Examples of patches used from the external database from Arvaniti et al. (first row) and Gerytch et al. (second row). (a) and (e): Benign glands; (b) and (f): Patches containing GG3 patterns; (c) and (d): Patches containing GG4 patterns; (d) and (h): Patches containing GG5 patterns.

Database	ACC	F1S				Avg-F1S	κ
		NC	GG3	GG4	GG5		
Arvaniti et al. [83]	0.5861	0.5660	0.6858	0.4688	0.5603	0.5702	0.6410
Gerytch et al. [77]	0.5136	0.2901	0.6162	0.4990	0.4958	0.4753	0.5116

Table 2.7: Results of the patch-level Gleason grading in the Arvaniti and Gerytch databases by our proposed model, *FSConv*. The metrics presented are accuracy (ACC), F1-Score (F1S), computed per class and its average, and Cohen's quadratic kappa (κ).

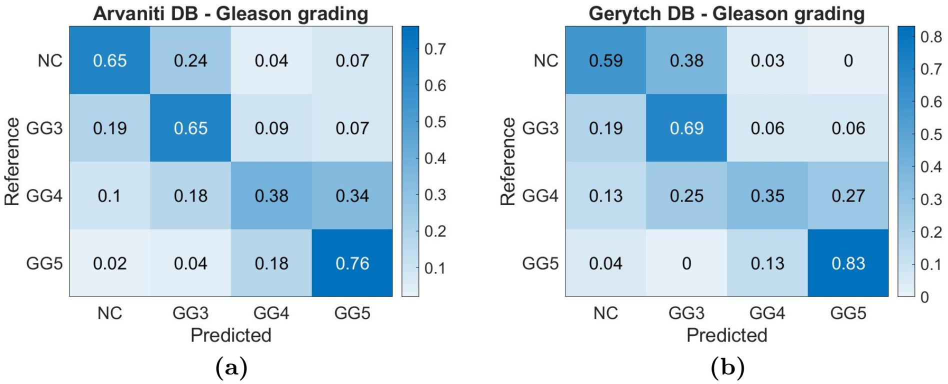


Figure 2.11: Confusion Matrix of the patch-level Gleason grades prediction in external databases using the proposed *FSConv* model. (a): Arvaniti database and (b): Gerytch database.

The obtained results in Arvaniti et al. database were slightly worse than the ones reached in our test cohort. The macro-averaged F1 score was 0.57, while 0.65 was obtained in the test cohort (see Table 2.6). To the best of the authors's knowledge, this is the first time in the literature that a model trained for patch-level Gleason grading in tested on an external database. This is a challenging task, taking into account the known inter-pathologist variability of the Gleason grading task and the differences in the histology sample preparation. Thus, the difference in the results could be explained by those factors. In comparison to the results obtained in [83] on this database, the reported κ in the test subset was 0.55 (see Table 2.6), while the κ obtained by our model was 0.64. Our proposed model outperforms the current state of the art on this set of images, even though we used the whole database for testing, and they reported the result on a specific test subset.

Regarding the obtained results on the Gerytch et al. database, a macro-averaged F1 score of 0.47, and a κ of 0.51 were obtained. Note that the small amount of non cancerous patches in this database (32 patches with only benign annotation, compared to 116 in Arvaniti et al. set) could be negatively affecting the figures of merit. Unfortunately, to the best of the authors' knowledge, no work has been reported on the use of the entire set of grades on this database, which makes the comparison impossible.

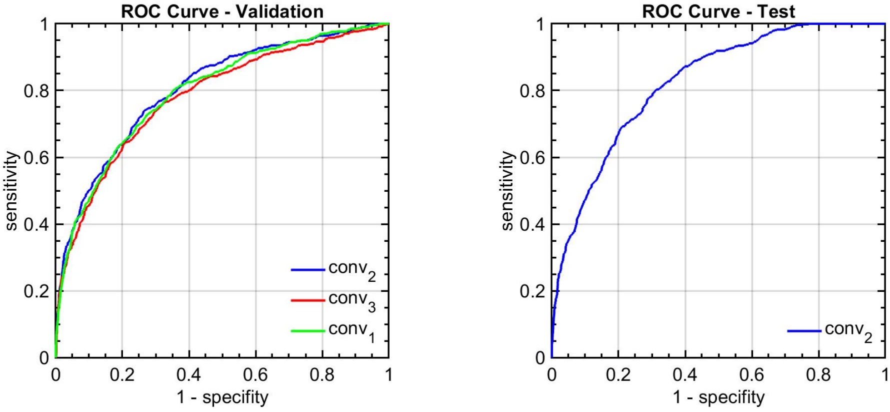
2.5.4 Cribriform pattern detection

To detect cribriform patterns in GG4 patches, $FSConv$ trained in the Gleason grading stage was re-trained as specified in Section 2.3.2 with a learning rate of 0.001 and a batch size of 32 samples during 200 epochs. The results were optimised freezing the weights of the convolutional filters at different depths. Concretely, at filters $conv_1$, $conv_2$ and $conv_3$ (see Table 2.1 for $FSConv$ architecture details). The output probability of each model was used to compute the Receiver Operative Curve (ROC) and evaluate the Area Under Curve (AUC). Then, probabilities were thresholded to output a positive classification when they are above 50%. The results obtained for the cross-validation set are presented in Table 2.8, and the Receiver-Operative-Curve in Figure 2.12 (a).

Experiment	**ACC**	**Sensitivity**	**Specificity**	**AUC**
$conv_1$	0.8218 ± 0.0541	0.8837 ± 0.0525	0.5263 ± 0.1159	0.8172 ± 0.0689
$conv_2$	**0.8350 ± 0.0599**	**0.8993 ± 0.0436**	0.5223 ± 0.1435	**0.8225 ± 0.0733**
$conv_3$	0.8103 ± 0.0712	0.8586 ± 0.0650	**0.5476 ± 0.2229**	0.7965 ± 0.1018

Table 2.8: Results in the detection of cribriform pattern in the validation set. The accuracy (ACC), Sensitivity, specificity and area under ROC curve (AUC) are presented for the fine-tuned $FSConv$ model freezing up to the convolutional layers $conv_1$, $conv_2$ or $conv_3$.

The best results were obtained for the validation set by the network whose weights were frozen up to the layer $conv_2$. Thus, just the last layer, $conv_3$ and the output neuron were trained. The accuracy obtained through this configuration was 0.8225, with a sensitivity and specificity of 0.8993 and 0.5223, respectively. The reached AUC was 0.8225. Slightly better results were obtained by this model in the test subset. The ROC computed in the test subset is presented in Figure 2.12 (b), and it encloses an AUC of 0.8240. This value is at the permissible confidence



(a)

(b)

Figure 2.12: ROC curves obtained for cribriform pattern detection in samples with Gleason grade 4.

level of systems for medical applications, above 0.80 [99]. Although the accuracy value decreases to 0.7239, the sensitivity and specificity are more balanced, with values 0.7168 and 0.7586, respectively. To the best of the authors's knowledge, this is the first time that the detection of cribriform patterns in histology prostate images is addressed and evaluated, so that it is not possible to establish comparison with previous works. Nevertheless, the studies comparing the inter-observer variability of the Gleason patterns classification show the challenging character of this task. In [102] the reproducibility in this problem was studied with 23 genitourinary pathologists. The consensus was achieved for cribriform glands in only 23% of the cases, and a consensus was not reached in how to classify the complex fused glands with cribriform shapes. We observed that the misclassified instances in our approach were mainly due to this kind of pattern. In Figure 2.13 few representative examples are presented, being (d), (e), and (f) images with complex fused glands that the model misclassified as cribriform pattern. Therefore, the results obtained by the model are auspicious, and its main limitation is the misclassification of patterns with large inter-pathologist variability.

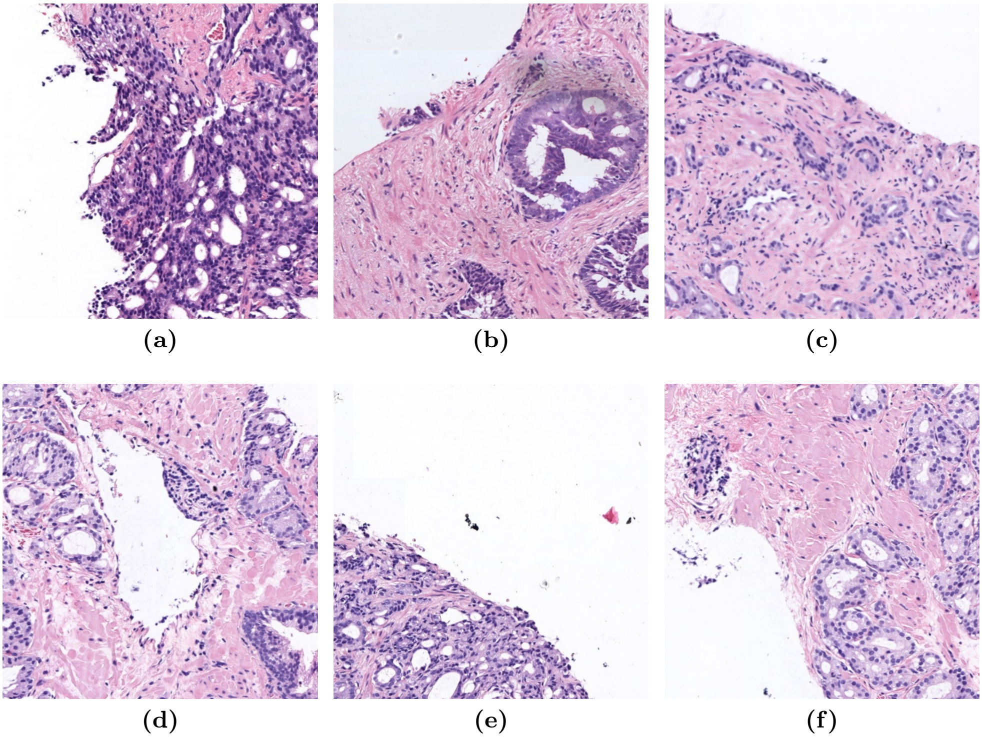


Figure 2.13: Examples of the system performance in the test subset for cribriform pattern detection. (a): True Positive, (b): True Positive, (c): True Negative, (d): False Positive, (e): False Positive, (f): False Positive.

2.5.5 WSI-level Gleason scoring

Once the patch-level prediction is performed with model $FSConv$, the probability maps for each Gleason grade are obtained, as specified in Section 2.3.3. The usability of these maps in the clinical practice were qualitatively validated by expert pathologists with satisfactory results.

Different examples of the test subset are presented in Figures 2.14, 2.15, and 2.16. These figures are organised as follows: in the first column, the WSI with pixel-level annotations (a) and pixel-level predictions (b) are presented, while in the second, the heatmaps of GG3 (c), GG4 (d) and GG5 (e) are shown from top to bottom, respectively. The regions of interest in the WSIs are highlighted with a higher resolution window to facilitate visualisation. The example in Figure 2.14 is a biopsy with Gleason score $3 + 4 = 7$, the biopsy in Figure 2.15 corresponds to a $3 + 3 = 6$ sample and the case in Figure 2.16, $5 + 5 = 10$. Finally, a non-cancerous case is presented in Figure 2.17.

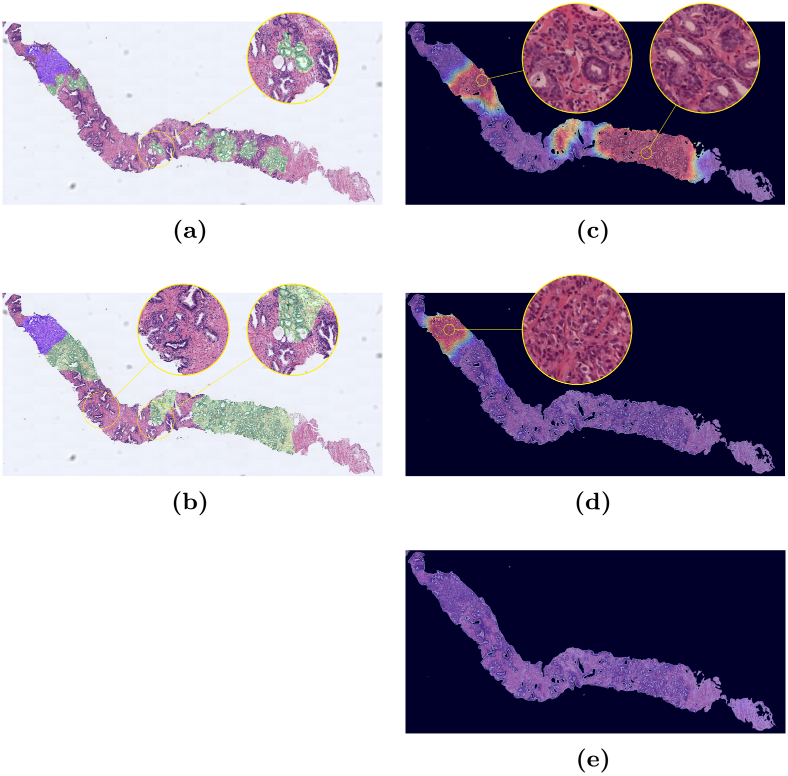
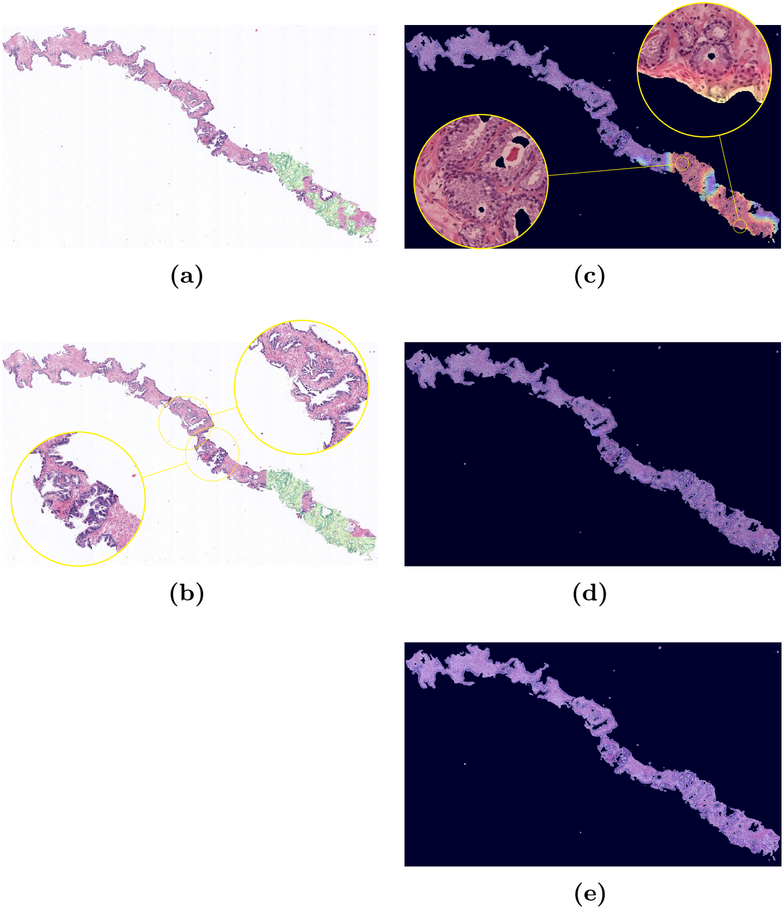


Figure 2.14: Whole slide image level prediction of a biopsy diagnosed as Gleason Score $3 + 4 = 7$. (a): manual annotations, (b): system predictions. Green: GG3, Blue: GG4, red: GG5. (c): GG3 heatmap, (d): GG4 heatmap, (e): GG5 heatmap.

In the case presented with Gleason score $3 + 4 = 7$ (see Figure 2.14), the GG3 and GG4 regions are correctly classified. In a subsequent review of this case, pathologists detected that some glands in the right region without pathologist's annotations in the ground truth and classified as GG3 by the model were actually cancerous patterns. Additionally, the



(a)

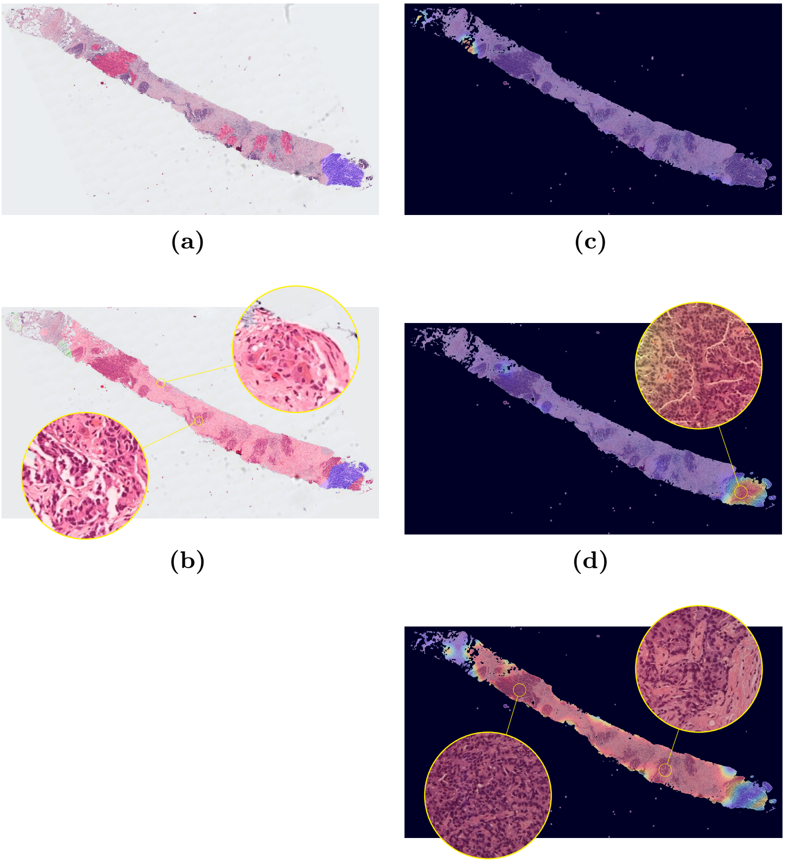
(c)

(b)

(d)

(e)

Figure 2.15: Whole slide image level prediction of a biopsy diagnosed as Gleason Score $3 + 3 = 6$. (a): manual annotations, (b): system predictions. Green: GG3, Blue: GG4, red: GG5. (c): GG3 heatmap, (d): GG4 heatmap, (e): GG5 heatmap.



(a)

(c)

(b)

(d)

(e)

Figure 2.16: Whole slide image level prediction of a biopsy diagnosed as Gleason Score $5+5=10$ (a): manual annotations, (b): system predictions. Green: GG3, Blue: GG4, red: GG5. (c): GG3 heatmap, (d): GG4 heatmap, (e): GG5 heatmap.

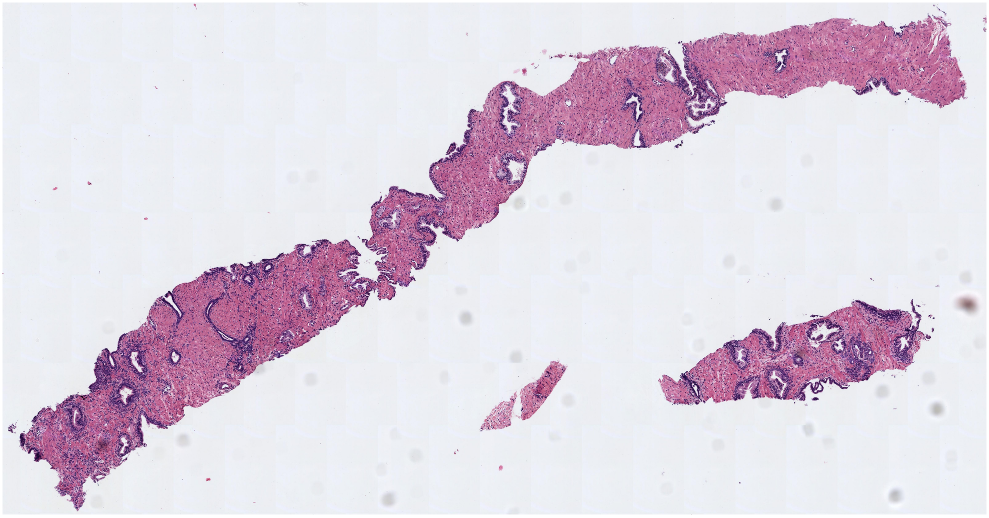


Figure 2.17: Non-cancerous biopsy without Gleason grades detected by the model.

few non-cancerous dilated and fusiform glands were correctly classified as non-cancerous (see Figure 2.14 (b), regions of interest highlighted). Regarding the biopsy with Gleason score of $3+4 = 7$, the model correctly detects the region with GG3 glands, but due to the patch resolution (512^2 pixels) some nearby stroma regions are highlighted as cancerous. Finally, analysing the case with a score of $5 + 5 = 10$, a papilar GG4 pattern is being correctly detected. The same occurs in the GG5 regions with isolated cells and pseudorosetting patterns. Nevertheless, in regions with a score of $GS \geq 9$ some stroma regions are frequently highlighted as GG5 by the model. This phenomenon does not occur in stroma of biopsies with $GS < 9$, as can be seen in the other cases. This fact suggests that the model could be detecting some hidden pattern of interest in the structure of the stroma in these regions.

Then, the percentages corresponding to each grade per WSI were obtained as specified in the methodology (Section 2.3.3). The proposed architecture MLP was then trained using as input the percentages obtained in the cross-validation subset. Adam optimiser was used, with a learning rate of 0.01, and a constant decay to zero over the 2000 epochs. The batch size was 32. The training strategy was leave-one-out.

This proposed approach is compared with the method proposed by Arvaniti [83] using $T = 10\%$ as minimum number of pixels with a certain label to be consider the corresponding grade in the WSI grading. The confusion matrix at biopsy level obtained for both methods is presented in Figure 2.18, and Cohen's quadratic kappa (κ) was calculated as a figure of merit.

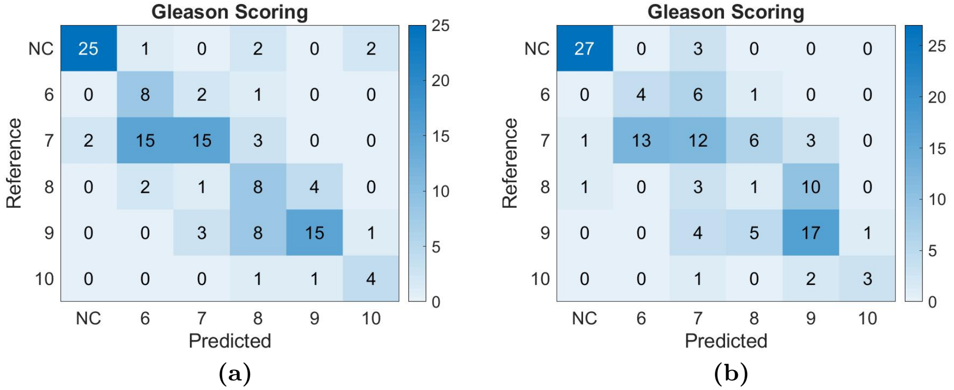


Figure 2.18: Confusion matrix of the whole slide image level Gleason scoring in the validation cohorts. (a): Method proposed in [83]; (b): *MLP* model.

The κ value obtained for Arvaniti's approach was 0.7693, in line with the results presented in [83] using their own database (using TMAs), where the obtained κ value was 0.75. Better results were obtained with the proposed model MLP (see Figure 2.18 (b)), obtaining a κ value of 0.8177. The main difference between methods was observed in few samples misclassified as Gleason score 8 and Gleason score 10 by Arvaniti's proposal which were correctly classified by our model. Therefore, our proposed strategy seems to model better the pathologist's decision to assign a Gleason score to the full image of the slide than the previous scoring methodology. The results obtained in the test subset by MLP model are similar to those obtained for the validation cohorts, with a κ value 0.8168.

2.6 Conclusions

In this work, we have proposed and validated end-to-end approaches to automatically support the pathologists analysis of prostate whole slide images. This support includes the pixel-level prediction of Gleason grades, cribriform patterns detection, calculation of the percentage of each grade in the tissue and finally the scoring of the entire biopsy.

We have compared fine-tuned state-of-the-art architectures and self-designed convolutional neural network architectures trained from scratch for the patch-level Gleason grades prediction. In addition, we have discussed the use of a global-max-pooling and global-average-pooling layers in the top model for this application. The use of global-max pooling has showed interesting properties in the model trained from scratch. It supports the use of shallow architectures with a small receptive field and a reduced amount of parameters, diminishing one of the main drawbacks of training from scratch: the over fitting to the training set. Thus, with a concise model composed of three convolutional layers, we have achieved the best results in our data set, reaching a Cohen‘s quadratic kappa of 0.77 in the test images. Furthermore, by just re-training the filter weights of the last convolutional layer, we have predicted the presence of cribriform regions in patches with Gleason grade 4, with an AUC value of 0.82 in the test subset. To the best of the authors’s knowledge, this is the first work contemplating the automatic detection of cribriform patterns in prostate histology images. We also have studied the interpretability of the developed deep-learning models by means of Class Activation Maps. Additionally, we have obtained probability heat maps indicating the presence of the different Gleason grades in the whole slide image. Finally, making use of the percentage of non-cancerous, Gleason grade 3, 4, and 5 tissues in the biopsy we have predicted its combined Gleason score through a multi-layer perceptron, reaching a Cohen’s quadratic kappa of 0.8168 in the test cohort. This model reproduces better the decision-making of the pathologist reporting the biopsy score than previous ones based on just assigning the two first grades with a higher percentage.

The limitations of the study naturally include the intra-observer variability of the annotator. This fact is not present on the trained algorithm, but it could affect the figures of merit obtained. Additionally, the large heterogeneity inside each Gleason grade makes difficult to balance the different folds, representing all the different patterns of the Gleason grades in all the training and testing groups.

It is important to note that this work brings an important contribution to the scientific community: the SICAPv2 database, the largest public database containing pixel-level annotations of prostate biopsies.

Further research will focus on developing convolutional-neural-network architectures that combine low and high-level features in the classification stage, as well as the inclusion in those models the prediction of all the individual cancerous patterns (i.e. ill-fused, papillary or large-fused) as the cribriform one, in an end-to-end training. Furthermore, the SICAPv2 database will be enlarged with additional annotated whole slide images.

Chapter 3

WeGleNet: A weakly-supervised convolutional neural network for the semantic segmentation of Gleason grades in prostate histology images.

The content of this chapter corresponds to the author version of the following published paper: Silva-Rodríguez, J., Colomer, A., & Naranjo, V. WeGleNet: A weakly-supervised convolutional neural network for the semantic segmentation of Gleason grades in prostate histology images. Computerized Medical Imaging and Graphics, 88, (2021).

Contents

WeGleNet: A weakly-supervised convolutional neural network for the semantic segmentation of Gleason grades in prostate histology images

Julio Silva-Rodríguez[1], Adrián Colomer[2] and Valery Naranjo[2]

[1]Institute of Transport and Territory, Universitat Politècnica de València, Valencia, Spain; [2] Institute for Research and Innovation in Bioengineering, Universitat Politècnica de València, Valencia, Spain

Abstract

Prostate cancer is one of the main diseases affecting men worldwide. The Gleason scoring system is the primary diagnostic tool for prostate cancer. This is obtained via the visual analysis of cancerous patterns in prostate biopsies performed by expert pathologists, and the aggregation of the main Gleason grades in a combined score. Computer-aided diagnosis systems allow to reduce the workload of pathologists and increase the objectivity. Nevertheless, those require a large number of labeled samples, with pixel-level annotations performed by expert pathologists, to be developed. Recently, efforts have been made in the literature to develop algorithms aiming the direct estimation of the global Gleason score at biopsy/core level with global labels. However, these algorithms do not cover the accurate localization of the Gleason patterns into the tissue. These location maps are the basis to provide a reliable computer-aided diagnosis system to the experts to be used in clinical practice by pathologists. In this work, we propose a deep-learning-based system able to detect local cancerous patterns in the prostate tissue using only the global-level Gleason score obtained from clinical records during training. The methodological core of this work is the proposed weakly-supervised-trained convolutional neural network, WeGleNet, based on a multi-class segmentation layer after the feature extraction module, a

global-aggregation, and the slicing of the background class for the model loss estimation during training. Using a public dataset of prostate tissue-micro arrays, we obtained a Cohen's quadratic kappa (κ) of 0.67 for the pixel-level prediction of cancerous patterns in the validation cohort. We compared the model performance for semantic segmentation of Gleason grades with supervised state-of-the-art architectures in the test cohort. We obtained a pixel-level κ of 0.61 and a macro-averaged f1-score of 0.58, at the same level as fully-supervised methods. Regarding the estimation of the core-level Gleason score, we obtained a κ of 0.76 and 0.67 between the model and two different pathologists. WeGleNet is capable of performing the semantic segmentation of Gleason grades similarly to fully-supervised methods without requiring pixel-level annotations. Moreover, the model reached a performance at the same level as inter-pathologist agreement for the global Gleason scoring of the cores.

3.1 Introduction

Prostate cancer is one of the most common diseases affecting men worldwide. It constitutes 14.5% of all cancers affecting men [65], and, according to the World Health Organization, the yearly number of new cases will increase by up to 1.8 million people in this decade [66]. The gold standard for prostate cancer diagnosis and prognosis prediction is the analysis of prostate biopsies under the Gleason grading system [67]. This system defines a series of cancerous patterns related to the morphology, distribution, and degree of differentiation of the glands in the tissue. Specifically, in histology slides, the observable Gleason grades (GG) range from 3 (GG3) to 5 (GG5). Examples of those patterns are presented in Figure 3.1.

In clinical practice, small portions of tissue are extracted, laminated, stained with Hematoxylin and Eosin, and finally analyzed under the microscope by expert pathologists using this system. Local cancerous regions of the sample are classified according to the Gleason grades, and finally, the two majority patterns are grouped to obtain a Gleason score as prognosis biomarker (e.g. the Gleason score $5 + 4 = 9$ would be assigned to a sample in which the main cancerous Gleason grade is GG5 and the

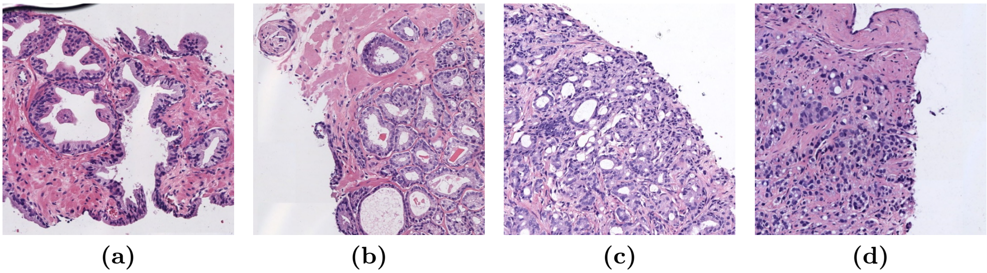


Figure 3.1: Histology regions of prostate biopsies. (a): region containing benign glands, (b): region containing GG3 glandular structures, (c): region containing GG4 patterns, (d): region containing GG5 patterns. GG: Gleason grade.

second is GG4). Due to the large size of the biopsies augmented under a microscope, this process results in a high time-consuming and repetitive task, and presents a large intra and inter pathologist variability [103].

In the last decades, the development of digitization devices has allowed the storage of biopsies at microscopic magnifications as digital images. Due to this advance, the field of Computer-Aided Diagnosis (CAD) systems to support pathologists based on computer-vision techniques has experienced a great growth. However, the development of those applications is limited due to the high data-demanding character of deep learning algorithms, and the difficulty in obtaining pixel-level labeled histology images [104]. Normally, pathologists store in the clinical history the global-level diagnosis of the biopsy (e.g. the Gleason score per prostate biopsy). In order to train/build models or develop algorithms able to detect and grade local cancerous patterns, a laborious manual annotation process is required, which must be performed by expert pathologists due to the complexity of the task. In the case of prostate cancer, the different tumor patterns have to be accurately delimited at the pixel level to avoid noisy annotations. Even though multi-resolution graphical user interfaces are provided to clinicians for performing this task, it is a tedious process prone to error. These limitations encourage the development of weakly-supervised deep-learning techniques able to utilize global labels during the training process to accurately identify local cancerous patterns in the images. The main benefit of those methods is that they are not limited to the annotated samples. They can work using histology images labeled only in the global-level patient diagnosis. Recent advances in the literature have proposed the use of the global Gleason score (obtained from the clinical record) to develop CAD systems for biopsy scoring (these works are detailed in Section 3.2.2). Nevertheless, these methods focus on predicting only global biopsy-level markers, while the location of the cancerous structures in the tissue is qualitatively evaluated or not addressed. The classification of local Gleason grades in prostate biopsies is the basis of CAD systems during its use in clinical practice. Accurate heat-maps provide confidence to the pathologists in the daily use of the CAD system, and support the biopsy-level markers provided by the system.

In this work we propose a deep-learning architecture based on convolutional neural networks able to perform a semantic segmentation of the Gleason grades (i.e. non-cancerous tissue, GG3, GG4 or GG5 classes) in prostate histology images, trained via weak supervision using the diagnosed Gleason score of the sample. To the best of the authors' knowledge, this is the first time in the literature that weakly-supervised methods are explored and quantitatively assessed for the local segmentation of cancerous Gleason grades. The main contributions of this research are the following: (i): a weakly-supervised framework based on a convolutional neural network (CNN) architecture able to obtain complementary semantic segmentation maps based on a novel configuration of multi-class activation maps, aggregation layers and the slicing of the background class prediction during training; (ii) the validation of different aggregation layers and regularization techniques to optimize the model; and (iii) the comparison of the proposed weakly-supervised model with fully-supervised state-of-the-art methods.

3.2 Related work

3.2.1 Weakly-supervised semantic segmentation

Weakly-supervised learning deals with the challenge of using incomplete, scarce, inexact, inaccurate, or noisy information. The problem addressed in this work, image segmentation using just global labels during training, is covered within the Multiple Instance Learning (MIL) scope. MIL works with data clustered on bags of instances, under the assumption that bags labeled as a certain class present, at least, one instance belonging to that class. For one image X composed by the instances (pixels) x_{ij}, the bag-level label (Y) for a class (c) could be interpreted as:

$$Y_c = \begin{cases} 1, & \text{if } \exists\ x_{ij} : y_c = 1 \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

where y_c is the instance-level label for certain class c.

In this topic, two different kinds of classification problems are defined: the prediction of bag-level (global) labels, or the classification of individual instances. In this work, both problems are addressed. A recent extensive review of MIL and its characteristics can be found in [105]. Regarding MIL in image classification, convolutional neural networks (CNNs) are the most used technique, since they have demonstrated promising properties for locating objects while performing image-level classification tasks [30, 36].

The approaches to obtain segmentation maps from global-level image classification using CNNs can be divided into aggregation and gradient-based methodologies. Aggregation methods build segmentation maps into the CNN architecture. They are composed of three main blocks: a feature-extraction stage (or base model), an adaptation layer that constructs segmentation maps per class, and a global aggregation layer that resumes each map to one representative value. Then, a multi-label loss function is used to optimize the network weights. The main proposed architectures in this field are WILDCAT [43] and Attention-MIL [46]. WILDCAT constructs the adaptation layer by pooling activation maps after the last convolutional block of the base model and then applies a global-pooling operator to obtain the bag-level probabilities. Attention-MIL joins the adaptation and global aggregation layer by using an attention mechanism that combines all the features obtained in each instance by fully-connected layers. Regarding the gradient methods, the segmentation maps are obtained by post-processing the network output.

In this line, the most relevant technique in the literature is the gradient-based class activation maps (Grad-CAMs) [40]. In this technique, the activation maps of the last convolutional block are linearly combined. Each map is weighted by back-propagating gradients in the network from the classification layer, and a ReLU activation is applied to the weights to keep just the features with a positive influence on the classification. Recently, the efforts on weakly supervised semantic segmentation have focused on self-supervised learning. In this methodology, CAMs obtained from gradient-based methods are used as pseudo labels to feed a pixel-level semantic segmentation network. Although these methods have reached promising results, they are still limited by the CAMs used, and the propensity of CNNs to look only at specific and discriminatory patterns. In this line, Ficklenet [106] and IRNet [107] have proposed the use of center-fixed spatial dropout and class propagation respectively to alleviate this limitation. In all the strategies, the aggregation of the different class-level maps (or CAMs) in a semantic segmentation mask is not straightforward. This process is usually carried out by hand-crafted post-processing. Some methods are based on simply assigning to each pixel the label with the highest probability and let as background those instances with probabilities below certain threshold [106]. Other works apply complex fully-connected conditional random fields (CRF) to combine the different class-level maps into one combined mask [43, 108–110]. In our work, we take steps forward in order to solve this limitation, and propose a CNN architecture that obtains complementary multi-class semantic segmentation maps without requiring any post-processing (see Section 3.3.1 for further explanation). An extensive survey regarding the application of weakly-supervised learning across different image domains and its current limitations was recently presented in [111].

3.2.2 Weakly-supervised segmentation in histology images

Weakly-Supervised learning is a field of increasing interest for histology images, due to the difficulty of preparing large datasets labeled by expert pathologists. While some works just focus on the prediction of bag-level labels in biopsy slides [47, 112–114] carrying out a qualitative evaluation of instance-level (local) classifications, others quantitatively evaluate their proposed models for the local-level classification task [44, 49, 115, 116]. Nevertheless, most of the works only focus on binary classification cancer/no cancer. Early work in [116] proposes a MIL model based on hand-crafted feature extraction (SIFT, color histogram, Local Binary Patterns, etc.), machine learning classifiers and aggregation of the instance-level probabilities for colon cancer detection. Lately, semi-supervised CNNs were used for gland segmentation in prostate images in [115]. However, the proposed UNet required to incorporate some instance-level annotations during training to perform properly. Finally, recent work in [49] included previous knowledge by applying constraints in the training stage of a weakly-supervised CNN to control the size of positive instances in the image for colon cancer detection. Recent works have used weakly-supervised CNNs approaches for multi-class semantic segmentation. Concretely, HistoSegNet, introduced in [44], performs a weakly-supervised segmentation of different tissue types in histology images based on CNNs and Grad-CAM gradient method. Then, a complex hand-crafted post-processing is proposed to join the class-level segmentation maps and to include the background class.

3.2.3 Prostate Gleason grading

In the analysis of prostate histology samples, as mentioned previously, there are two main tasks: the grading of local structures using the Gleason system, and the global scoring.

First works in this field focused on fine-tuning well-known CNN architectures in a supervised patch-level classification, with the requirement of pixel-wise expert annotations. In this line, Nir et al. [84, 85] obtained a patch-level Cohen's quadratic kappa (κ) of 0.60 in the validation set, while 0.55 and 0.49 was reached by Arvaniti et al. in [83] in the test cohort referenced to two different pathologists. Then, the percentage of each cancerous tissue in the sample was calculated from the patch-level probabilities to predict the Gleason score of the sample. Arvaniti et al. [83] obtained with this method a κ of 0.76 and 0.71 against the annotations of two different pathologist, at the level of the inter-pathologist agreement ($\kappa = 0.71$).

Latest works in the literature have started to develop weakly-supervised techniques to avoid the tedious process of pixel-level labeling of Gleason grades. These techniques are based on assigning the global labels (i.e. the primary and secondary grades obtained from the Gleason score) to patch-level regions of interest (i.e. glandular or nuclei structures). Then, convolutional neural networks are trained to perform a patch-level classification with the obtained pseudo-ground truth. The selection of regions of interest in the tissue are based on different approaches, detailed in the following lines. The work in [117] developed a semi-supervised pipeline detecting the glandular tissue via a UNet trained with manual annotations. A few works works focus on selecting these regions with larger amounts of nuclei, based on color [78, 118] or Laplacian filters [119]. Finally, the work in [47] directly assigns the global label (cancerous against non cancerous) to all the patches in the tissue. All previous methods train patch-level convolutional neural networks with the obtained pseudo-ground truth, and finally they combine the patch-level predictions to obtain the global score. The first works aggregate the predictions using the percentage of each Gleason grade in the sample and then they train different machine learning models to predict the global Gleason score [78, 117, 119]. Also, novel approaches combine the patches using the features extracted by the CNN through recurrent neural networks [47]. Although the aforementioned methods provide promising results for Gleason scoring of prostate biopsies, the assumptions made to develop their weakly-supervised pipeline could be affecting the local grading of cancerous patterns. To the best of the authors' knowledge, none of previous works in the literature focus on locating the Gleason grades in the tissue using weakly-supervised learning. They only perform a qualitative evaluation of the heat-maps obtained by their models.

3.3 Methods

3.3.1 WeGleNet: weakly-supervised Gleason grading network

The methodological core of this work consists of a convolutional neural network able to predict semantic segmentation maps of non-cancerous, Gleason grade 3 (GG3), GG4, and GG5 tissue in prostate histology images, trained using global labels of the grades present in the tissue during training. The proposed weak-supervised Gleason grading network (WeGleNet) is presented in Figure 3.2.

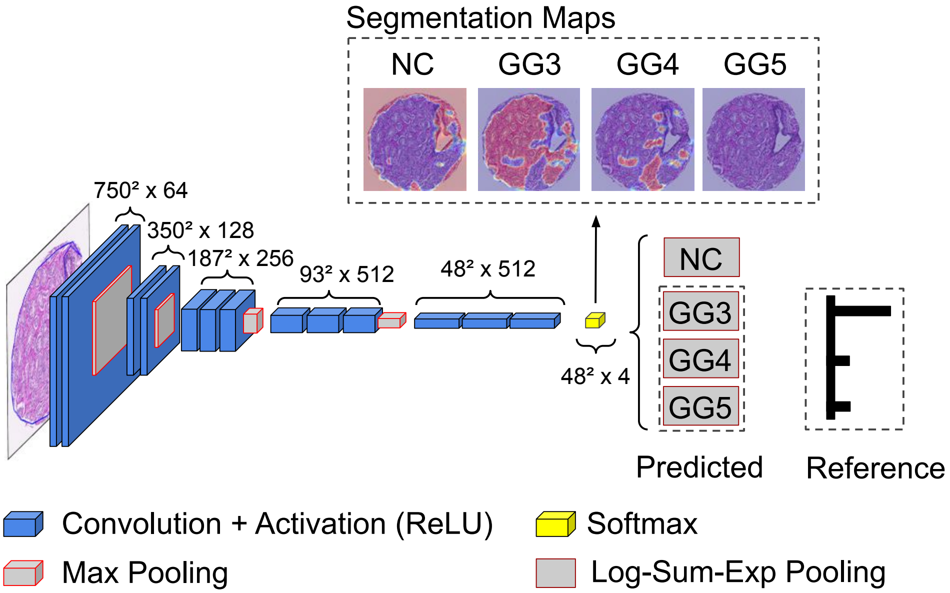


Figure 3.2: WeGleNet, weakly-supervised framework for semantic segmentation of local cancerous patterns via Gleason grading using the Gleason score of the global sample during the training stage. NC: non cancerous; GG3: Gleason grade 3; GG4: Gleason grade 4; GG5: Gleason grade 5.

The architecture is composed of three main components: the base model, the segmentation (also called adaptation) layer, and the global-aggregation operation, and it takes as input the prostate core image, which is resized to 750^2 pixels due to computational limitations. First, the base model is in charge of extracting automatic-learned features from the input image. Concretely, the VGG19 architecture [90] is used. This is based on convolutional blocks with an increasing number of filters with 3×3 kernels with ReLU activation and dimensional reduction via max pooling of size 2×2. In order to reduce the over-fitting during the training stage, weights are initialized using the VGG19 model pretrained in the ImageNet dataset [94]. Secondly, the segmentation layer applies to the output convolutional feature volume of the base model as many convolutional filters of size 1 × 1 as classes to be predicted. This layer also computes a softmax activation along the class dimension generating a multi-class segmentation volume of activation maps, in which each value represents the probability of that pixel of belonging to a class. During the inference stage, this layer will be the model output, and each segmentation map will be resized to the original core dimensions (3100^2 pixels). During the training stage, the pixel-level probabilities in the activation maps are aggregated in order to output one global probability per class ranging between 0 and 1. This operation is performed by a global-aggregation layer, which is detailed in Section 3.3.2. This aggregation of instance-level predictions embedded in the training stage of the model avoids previous assumptions in the literature to locate the regions of interest in the tissue. Then, binary cross-entropy is used as a loss function. As all cores contain non-cancerous regions, the loss function is only calculated using the Gleason grade classes (i.e. GG3, GG4, and GG5). Thus, the NC class segmentation map gathers those patterns not related to cancer but does not contribute to the calculation of the loss function. This strategy allows obtaining complementary segmentation maps including the background class (in our case non-cancerous class). This is a step forward compared to previous methods, which were based on the individual prediction of segmentation maps per class, and complex post-processing to join them including the background class (see Section 3.2.2 for a more detailed explanation of these methods).

During the training stage, two techniques are carried out to regularize the model and avoid over-fitting: data augmentation and hide-and-seek [120]. Data augmentation is performed by transforming the input images with random translations, rotations and mirroring in each iteration. Hide-and-seek (HS) is a method that regularizes weakly-supervised-trained architectures by replacing random patches of the images with the average intensity level of the input. In each iteration, the hidden patches vary, and thus the network is forced to focus on heterogeneous patterns during training. The input image is divided into patches of 75^2 pixels, which have a 25% probability of being hidden in each iteration.

3.3.2 Global-aggregation layers

Global-aggregation layers summarize information from all spatial locations in the activation maps (x_{ij}) to one representative value (p). For this task, we propose the use of the log-sum-exponential (LSE) layer [45] in WeGleNet, which is defined as:

$$p_{LSE} = \frac{1}{r} \cdot log \left[\frac{1}{S} \cdot \sum_{(i,j) \in S} exp(r \cdot x_{ij}) \right] \tag{3.2}$$

where S constitutes the number of pixels in the activation map x_{ij} and r is a parameter to be optimized.

The LSE operation permits us to obtain a domain-specific representation of the activation map via the parameter r, with large values of r ($r \rightarrow \infty$) similar to a global-max pooling operation (GMP) [30] and small values ($r \rightarrow 0$) equivalent to a global-average operation (GAP) [89]. The r parameter is empirically fixed by optimizing the model performance in the validation cohort (see Section 3.5.1). By this procedure, the training stage overcomes the limitations of the other global-aggregation layers (i.e. GAP assumes that the pattern is uniformly distributed across with the activation map, and GMP could produce over-fitting to small, specific patterns).

3.3.3 Global Gleason scoring

Once the probability maps per class are obtained, the Gleason score of the sample is inferred from the percentage of each class k in the tissue, w^k. In [83], the Gleason score is obtained assigning the majority and secondary grades in terms of percentage, considering only the classes above certain threshold c. In this work, we introduce another term, d, which models the tendency of pathologists to focus on the majority cancerous pattern if it is widespread in the tissue. Thus, the final percentage weights are assigned to each class such that:

$$w^k = \begin{cases} 0, & \text{if} \quad \max_{k'} \; w^{k'} > d \quad \text{and} \quad k \neq argmax_{k'} \; w^{k'} \\ w^k, & \text{otherwise} \end{cases} \tag{3.3}$$

where k denotes the different classes, i.e. Gleason grade 3, 4 and 5.

The operator d adapts the weakly-supervised framework to the global scoring procedure in clinical practice. Pathologists annotate regions focusing on primary patterns, while the weakly supervised model performs a more fine-grained segmentation, that increases the percentage of secondary patterns. Thus, d allows to suppress the system's confidence on these patterns for the global scoring task. The values of the parameters c and d are empirically fixed in the validation set to optimize the results.

3.4 Experimental setting

3.4.1 Datasets

The experiments described in this work were carried out using the public dataset presented by Arvaniti et al. in [83][1]. This dataset consists of 886 prostate Tissue Micro-Arrays (TMAs, samples of representative regions of cancerous biopsies known as cores), digitized at 40× magnification in images of size 3100^2 pixels. The cores include pixel-level annotations of Gleason grades and benign structures, and global labels of Gleason scores (primary and secondary Gleason grades in the sample). The distribution of the Gleason grades (GG) in the cores is distributed as follows: 421, 387 and 148 cores with GG3, GG4 and GG5, respectively. Regarding the pixel-level annotations, the dataset includes five different classes: benign tissue, GG3, GG4, GG5, and background. In order to evaluate our proposed methodology, the benign and background classes are joined in the non-cancerous (NC) class. To establish fair comparisons with previous literature, the partition of the dataset proposed by Arvaniti et al. was used for training, validating, and testing. Note that the

[1] We contacted the corresponding authors to obtain the dataset.

test cohort contains pixel-level annotations made by two different expert pathologists.

3.4.2 Experimental strategy and metrics

In order to validate the proposed WeGleNet model, two types of figures of merit are extracted from the model output: global-level (bag-level in the MIL framework) and local-level (instance-level) metrics. Figure 3.3 illustrates the evaluation strategy.

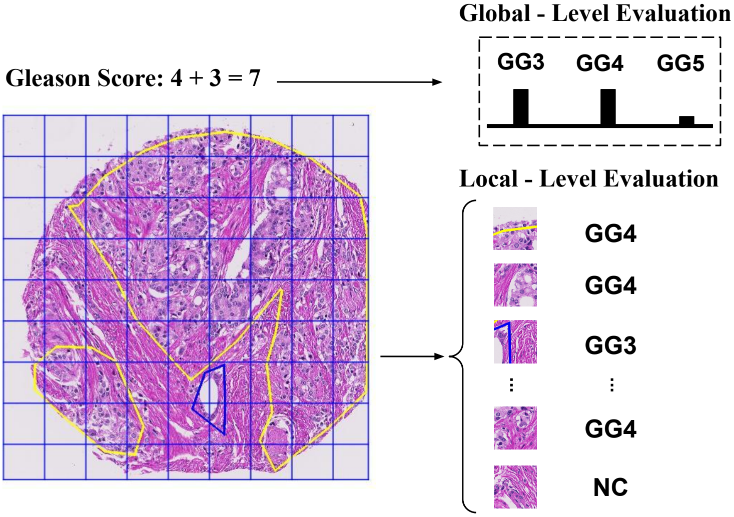


Figure 3.3: Strategies for the evaluation of the model performance. NC: non cancerous; GG: Gleason grade. The core-level (global) predictions are evaluated using the Gleason score. The local-level predictions are evaluated at pixel-level or using small patches extracted from the core.

Global-level metrics are obtained comparing the multi-label prediction of the WeGleNet in the global-aggregation layer and the Gleason grades observed in the core using the reference Gleason score. This evaluation is used to optimize the weakly-supervised model using the Area Under ROC curve (AUC) as a figure of merit. The decision of using this metric during the optimization stage is related to being closer to the output probabilities of the model. Finally, during the comparison of the model performance with previous literature (Section 3.5.4) the Cohen's quadratic kappa (κ) [97] is obtained for the Gleason score prediction. This agreement statistic takes into account that in a set of ordered classes, errors between adjacent classes should be less penalized.

Regarding the local-level evaluation, this is performed to analyze the capability of the trained model for segmenting the Gleason grades in the tissue. During WeGleNet optimization and its comparison with fully-supervised methods for semantic segmentation, metrics are obtained at pixel level. The obtained figures of merit are the accuracy (ACC), f1-score per class, the macro-average (F1), mean intersection over union (mIoU) and Cohen's quadratic kappa (κ). Usually, in the Gleason grading literature, the local grading of cancerous patterns is evaluated at patch level to avoid underestimation of the model performance due to an inaccurate pixel-level annotation in the ground truth. Therefore, WegleNet is evaluated at patch level for the comparison of its performance with previous state-of-the-art works in this field. In order to establish fair comparisons with previous results reported in the literature in the used dataset, patch-level labels are obtained as proposed by Arvaniti et al. [83]. Concretely, patches are extracted using a moving-window of size 750^2 and a step of 350 pixels. Patches with multiple or no annotations were discarded, and the remaining were labeled by majority voting according to the annotations in the central region of the patch (i.e. benign, GG3, GG4 or GG5).

3.4.3 Baselines

To compare our proposed weakly-supervised framework, two state-of-the-art supervised architectures for semantic segmentation of Gleason grades are implemented. To take advantage of the pixel-level annotations, patches are extracted from the cores with a size of 750^2 pixels and a step of 350. Due to hardware limitations during training, patches are resized to 224^2 pixels. Then, a UNet architecture and a classifier based on a patch-level VGG19 fine-tuned network (VGG19Sup) are selected as supervised architectures to be compared to the WeGleNet model. It is important to highlight that these methods require an accurate pixel-level labeling of the images. The implementation of the models is detailed in the following lines.

VGG19Sup is based on training a patch-level multi-class classifier and then modifying the architecture to obtain segmentation maps. VGG19Sup is composed of a feature-extraction stage using VGG19 backbone pre-trained in Imagenet dataset, a global-average pooling (GAP) to aggregate the activation maps, and a fully-connected layer with as many neurons as classes to predict and soft-max activation as output. In this method, each patch is labeled as the majority grade annotated. If none Gleason grade is annotated, the patch is labeled as non-cancerous. Training is performed by optimizing the categorical cross-entropy as loss function. For the inference of segmentation maps, the output fully-connected layer is converted in a convolutional layer with kernel 1×1, which is applied over the activation volume previous to the GAP layer to obtain a segmentation map per class. This approach is equivalent to using a class activation map (CAM) post-processing, but the segmentation maps are obtained directly from the CNN in an end-to-end manner. This method was previously used by Arvaniti et al. in [83] to obtain the probability maps in prostate samples.

Regarding the UNet architecture [121], it is based on a symmetric encoder-decoder path. In the encoder, feature extraction is carried out based on convolutional blocks and dimensional reduction through max-pooling layers. Each convolutional block increases the number of filters by a factor of 2×, starting from 64 filters up to 1024. After each block, the max-pooling operation reduces the dimensions of the activation maps in a factor of 2×. Then, the decoder path builds the segmentation maps, recovering the original dimensions of the image. The reconstruction process is based on deconvolutional layers with filters of size 3×3 and ReLU activation. These layers increase the spatial dimensions of the activation volume in a factor of 2× while reducing the number of filters by a half. Then, the encoder features from a specific level are joined with the resulting activation maps of the same decoder level by a concatenation operation, feeding a convolutional block that combines them. The convolutional block used during both encoder and decoder paths includes residual connections [91] to improve the model optimization. This residual UNet configuration was proposed in [122], and showed to outperform other configurations for Gleason grading in [123]. It consists of three convolutional layers with 3×3 kernels and ReLU activation. The output of the last convolutional layer of the block in connected via a shortcut residual operation with the output of the first layer. Finally, after the decoder, a 1×1 convolutional layer creates the segmentation probability maps. The loss function used during the training process is the categorical *Dice* used in [123].

During the inference stage, the supervised models are used to predict the entire core instead of local patches. Cores are resized to match the resolution used during training, and then the output segmentation maps are resized to the original dimensions of the cores (3100^2 pixels).

3.5 Results

The following section describes the experiments carried out to optimize the WeGleNet architecture (Section 3.5.1), and its comparison on the local-level segmentation of Gleason grades with supervised methods (Section 3.5.3) and with previous works using the same dataset (Section 3.5.4).

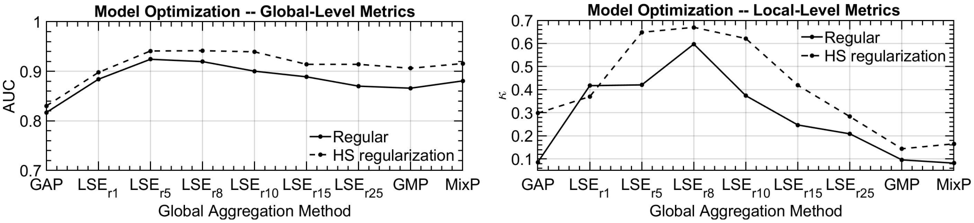
3.5.1 Model optimization

In the first experiments, the objective was to optimize the WeGleNet architecture for semantic segmentation using global-level labels (i.e. the presence of certain Gleason grade in the core). The model performance was studied under the different regularization techniques and global-aggregation layers. WeGleNet model was trained using the proposed log-sum-exponential (LSE), global-max (GMP) and global-average (GAP) pooling. In LSE layer, different values of the r parameter, $r = \{1, 5, 8, 10, 15, 25\}$, were used. In addition, to compare the performance of the LSE with respect to an automatic-learned combination of GMP and GAP, a mixed-pooling (MixP) aggregation layer is implemented such that:

$$p_{MixP} = \alpha \cdot p_{GMP} + (1 - \alpha) \cdot p_{GAP} \tag{3.4}$$

where α is a parameter learned during training.

The use of hide-and-seek (HS) regularization was validated by training the models with and without it. The training was performed in mini-batches of 8 images, and Stochastic Gradient Descent (SGD) was used as the optimizer with a learning rate of $1 \cdot 10^{-3}$. Exponential decay in the learning rate was applied in the last 20 epochs to stabilize the model weights such that: $\eta = 1 \cdot 10^{-3} \cdot e^{-0.1 \cdot t}$, where η is the applied learning rate and t is the epoch. The training was carried out during 120 epochs, which were increased to 400 when applying HS regularization. WeGleNet was trained using the training cohort, and early stopping was applied by keeping the weights of the model obtaining the best performance in the validation set (in terms of the obtained losses). After each experiment, segmentation maps were obtained from the segmentation layer, and core-level predictions were obtained from the global-aggregation layer using the images of the validation cohort. The scripts to reproduce the experiments reported in this work are publicly available on (`https://github.com/cvblab/prostate_wsss_weglenet`). Figures of merit related to global-level predictions and pixel-level segmentation are presented in Figure 3.4 (a) and (b) respectively.



(a) **(b)**

Figure 3.4: Model performance using different global-aggregation methods and regularization techniques. (a): global prediction performance; (b): pixel-level segmentation performance. HS: hide-and-seek; GAP: global-average pooling; LSE: log-sum-exponential pooling; GMP: global-max pooling; MixP: mixed pooling.

Regarding the obtained results, LSE pooling showed superior performance compared to other global-aggregation techniques. In particular, the best results were obtained using $r = 8$ (LSE_{r8}), with an AUC of 0.9243 for the core-level detection of Gleason grades and a κ of 0.5973 for the pixel-level segmentation. Hide-and-Seek regularization (HS) showed to improve the results in all the experiments, forcing the model to focus on all the patterns of the images. Thus, results improved to an AUC of 0.9416 and a κ of 0.6699 in the best-performing model, WeGleNet - LSE_{r8}. Finally, a high correlation was observed between the global-level and the local-level performance of the model. A Pearson correlation coefficient of 0.5462 was obtained between κ and AUC when using HS regularization. Then, improvements in the global-level predictions produced a better segmentation of the Gleason grades. This promising behavior indicates that the model can be optimized without any pixel-level annotations.

3.5.2 Qualitative evaluation

Once WeGleNet was optimized using the validation cohort, the best performing configuration, WeGleNet - LSE_{r8} with HS regularization, was used to predict the segmentation maps from the images of the test cohort. Representative examples of the obtained results are presented in Figure 3.5. This figure is organized as follows: each row is a different core and each column represents the ground truth of the Pathologist 1, and the predicted heatmaps for GG3, GG4 and GG5 classes, respectively. Finally, the last column presents the discrete-valued semantic segmentation maps, assigning to each pixel the class with the highest probability. In this figure, green, blue and red color indicate GG3, GG4 and GG5 patterns, respectively.

3.5.3 Weak supervision vs. strong supervision

Then, we carried out experiments to compare our proposed weakly-supervised model with respect to the state-of-the-art supervised methods. UNet model was trained using Nadam as optimizer, with a learning rate of $1 \cdot 10^{-4}$ during 60 epochs. In each iteration, a mini-batch of 16 images was used to update the models weights. Regarding the VGG19Sup

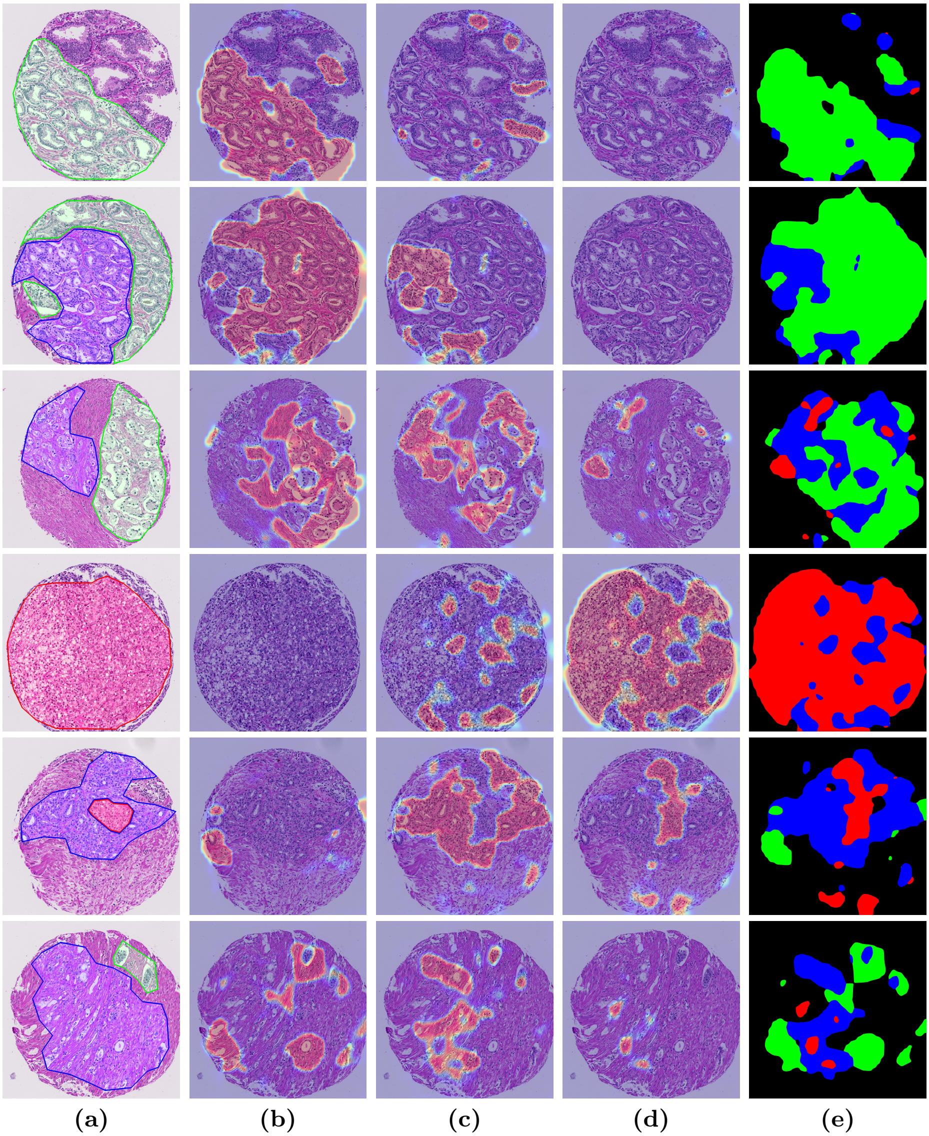


Figure 3.5: Examples WeGleNet segmentation performance in the test set. The reference annotations are obtained from Pathologist 1. In green: Gleason grade 3; blue: Gleason grade 4 and red: Gleason grade 5. (a): Reference; (b): Gleason grade 3; (c): Gleason grade 4; (d): Gleason grade 5; (e): Semantic segmentation mask. The reference and predicted Gleason scores (Pathologist 1 - Pathologist 2 - Predicted), from top to bottom, are: (6 - 6 - 6); (6 - 7 - 7); (7 - 7 - 8); (10 - 10 - 10); (8 - 9 - 8) and (7 - 7 - 7).

model, a learning rate of $1 \cdot 10^{-3}$ with SGD as optimizer was used. Training was performed in mini-batches of 64 images during 120 epochs. For both models, early stopping was applied to keep the best-performing model in the validation cohort (in terms of the obtained loss). From the trained models, the segmentation maps of the images in the test cohort were predicted. The figures of merit obtained by our proposed WeGleNet - LSE_{r8} and the supervised models are presented in Table 3.1, using as reference the annotations carried out by the pathologist 1 (the same pathologist that annotated the training and validation images). In order to perform a detailed comparison, accuracy (ACC), class-level f1-score (F1), average intersection over union (mIoU) and quadratic Cohen's kappa (κ) were obtained as detailed in Section 3.4.2.

Experiment	**ACC**	**F1**					**mIoU**	κ
		NC	GG3	GG4	GG5	Avg.		
WeGleNet - LSE_{r8}	0.685	0.815	0.588	0.562	**0.353**	**0.579**	**0.436**	0.610
SupVGG19	0.542	0.674	0.519	0.495	0.155	0.461	0.349	0.263
UNet	**0.696**	**0.838**	**0.593**	**0.573**	0.241	0.561	0.417	**0.638**

Table 3.1: Results of the Gleason grades semantic segmentation using the proposed weakly-supervised model, WeGleNet, and two supervised approaches, SupVGG19 and UNet. The metrics presented are the accuracy (ACC), the F1-Score (F1), computed per class and its average, the mean intersection over union ($mIoU$) and the Cohen's quadratic kappa (κ).

WeGleNet - LSE_{r8} model reached a κ value of 0.6105, a $mIoU$ f 0.4368 and an average F1 of 0.5798 in the semantic segmentation of Gleason grades in the test cohort. Our proposed model outperformed the supervised SupVGG19 model segmentation ($\kappa = 0.2630$, $mIoU = 0.3497$ and $F1 = 0.4613$), and it performs similarly to the UNet model ($\kappa = 0.6387$, $mIoU = 0.4178$ and $F1 = 0.5618$). Although the UNet model reached better results in the non-cancerous class ($F1 = 0.8383$), WeGleNet - LSE_{r8} differentiated better the Gleason grades, reaching an F1 of 0.3531 for the GG5 class, a challenging task due to the low prevalence of these patterns. Thus, our proposed WeGleNet model performed at a level equivalent to supervised methods in the segmentation of Gleason grades, without requiring pixel-level annotations.

3.5.4 State-of-the-art comparison

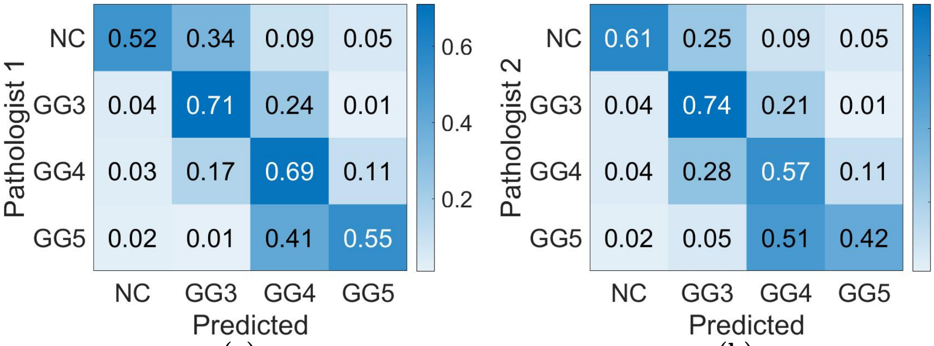
Finally, predictions were obtained at patch-level (which extraction is specified in Section 3.4.2) to compare WeGleNet against previous works in the used dataset. In the test cohort, patch-level classifications were obtained by majority voting of pixel-level predictions. Only fully non-cancerous patches were predicted as benign. The Cohen's quadratic kappa (κ) was obtained using the annotations of both pathologists. The figures of merit are presented in Table 3.2 and confusion matrices are presented in Figure 3.6.

Then, the global Gleason scoring of the cores was performed as described in Section 3.3.3. The parameters $c = 0.03$ and $d = 0.70$ were empirically fixed using the validation set. The κ and confusion matrices were obtained using as reference both pathologists, and the results are reported in Table 3.2 and Figure 3.7, respectively. Moreover, the obtained Gleason Score and references of representative cores are indicated in Figure 3.5.

In order to compare the obtained figures of merit with previous literature, the reported results for the patch-level grading and global scoring obtained using fully-supervised models with pixel-level annotations by Arvaniti et al. [83] are indicated in Table 3.2. Also, the results obtained in this test set by Bulten et al. [117] using semi-supervised models trained in a large set of biopsies (see Section 3.2.3 for a more detailed description) are pointed out in that table.

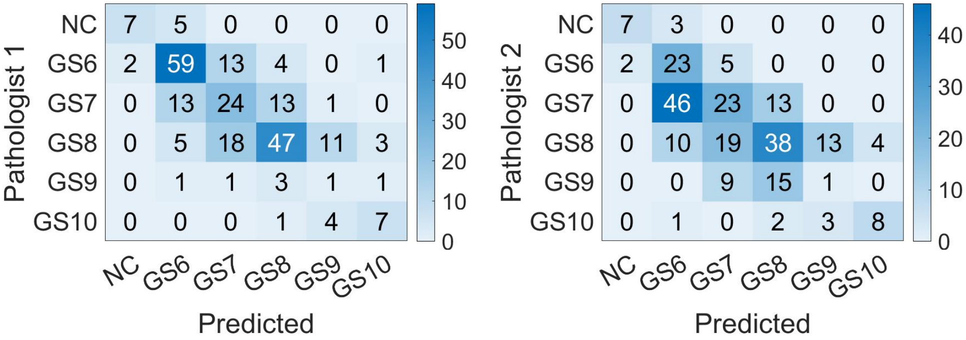
Approach	κ	
	Pathologist 1	**Pathologist 2**
Patch-Level Grading		
WeGleNet	0.59	0.50
Arvaniti et. al (2018) [83]	0.55	0.49
Pathologist 2	0.65	–
Core-Level Scoring		
WeGleNet	0.76	0.67
Arvaniti et. al (2018) [83]	0.75	0.71
Bulten et al. (2020) [117]	0.72	0.70
Pathologist 2	0.71	–

Table 3.2: Results of the patch-level Gleason grading and core-level scoring of the proposed model and comparison with previous literature. The metric presented is the Cohen's quadratic kappa (κ).



(a) **(b)**

Figure 3.6: Confusion Matrix of the patch-level Gleason grades prediction done by WeGleNet - LSE_{r8} network in the test subset. The reference labels in each matrix are obtained from: (a) pathologist 1, and (b) pathologist 2. GG: Gleason grade; NC: non cancerous.



(a) **(b)**

Figure 3.7: Confusion Matrix of the global-level Gleason scores prediction done by WeGleNet network in the test subset. The reference labels in each matrix are obtained from: (a) pathologist 1, and (b) pathologist 2.

The obtained results are in line with our previous experiments, and WeGleNet performed comparably to the fully-supervised approach used by Arvaniti et. al [83]. We reached a better κ value ($\kappa = 0.59$ against $\kappa = 0.53$) with the first pathologist, and similar performance was observed using the annotations from the second pathologist ($\kappa = 0.50$ against $\kappa = 0.49$). In addition, Figure 3.6 showed that most of the errors were conducted between adjacent classes.

Regarding the core-level Gleason scoring, the performance was also similar to previous works in the test set. A κ of 0.76 and 0.67 was obtained with each pathologist, respectively. In average, the obtained κ (0.715) is similar to the one obtained by Arvaniti et al. (0.730) and Bulten et al. (0.719). These results are at the same level of inter-pathologist agreement ($k = 0.710$). In addition, our approach obtained accurate localization heat-maps validated in Section 3.5.3 without using pixel-level annotations during training.

3.6 Conclusions

In this work, we have presented WeGleNet, a weakly-supervised trained architecture able to obtain semantic segmentation maps of Gleason grades in prostate histology images. The model is trained using just global-level labels, the Gleason score obtained from medical history, and it is capable of locating the local cancerous patterns in the tissue according to its grade.

Our proposed architecture makes use of multi-class segmentation layers after the feature-extraction stage, and a global-aggregation of the pixel-level probabilities into one representative value per class. Then, the output of the non-cancerous class (background) was sliced to obtain the loss of the model during training. This strategy allows us to obtain complementary maps in the architecture, without requiring complex post-processing of the output. In the experimental stage, we compared different global-aggregation layers and regularization techniques to optimize the model performance in the validation cohort. The log-sum-exponential pooling (LSE) showed superior performance than other layers, thanks to its ability to adapt the model to the specific domain via the adjustable parameter r. Thus, we have achieved a Cohen's quadratic kappa (κ) of 0.67 for the Gleason grading of local patterns in the validation cohort at the pixel level. During this optimization stage, we have observed a high correlation between global and local-level figures of merit. Thus, optimizing the proposed architecture using just global-labels involves improving the local-level localization of cancerous patterns. Additionally, we have compared the model performance with state-of-the-art supervised methods for semantic segmentation of Gleason grades in the test cohort. The proposed WeGleNet architecture performed similarly to supervised methods, without requiring any kind of pixel-level annotations during the training stage, reaching a pixel-level k of 0.61 and an average f1-score of 0.58. The performance for the core-level Gleason scoring was similar to previous works, and comparable to inter-pathologist agreement in the test cohort, reaching an average κ of 0.715. These promising results constitute a step forward in the literature of the analysis of prostate histology images and could avoid the tedious process of pixel-level generation of ground truth by expert pathologists.

Further research will focus on generalizing the proposed method to be trained using entire slices of biopsies digitized as whole slide images, whose larger size presents an added challenge in developing weakly-supervised methods for locating local cancerous patterns.

Chapter 4

Self-learning for weakly supervised Gleason grading of local patterns

The content of this chapter corresponds to the author version of the following published paper: Silva-Rodríguez, J., Colomer, A., Dolz, J., & Naranjo, V. Self-learning for weakly supervised Gleason grading of local patterns. IEEE Journal of Biomedical and Health Informatics, 25, (2021).

Contents

Self-learning for weakly supervised Gleason grading of local patterns

Julio Silva-Rodríguez[1], Adrián Colomer[2], Jose Dolz[3] and Valery Naranjo[2]

[1]Institute of Transport and Territory, Universitat Politècnica de València, Valencia, Spain; [2] Institute for Research and Innovation in Bioengineering, Universitat Politècnica de València, Valencia, Spain; [3]École de Technologie Supérieure, Montreal, Canada

Abstract

Prostate cancer is one of the main diseases affecting men worldwide. The gold standard for diagnosis and prognosis is the Gleason grading system. In this process, pathologists manually analyze prostate histology slides under microscope, in a high time-consuming and subjective task. In the last years, computer-aided-diagnosis (CAD) systems have emerged as a promising tool that could support pathologists in the daily clinical practice. Nevertheless, these systems are usually trained using tedious and prone-to-error pixel-level annotations of Gleason grades in the tissue. To alleviate the need of manual pixel-wise labeling, just a handful of works have been presented in the literature. Furthermore, despite the promising results achieved on global scoring the location of cancerous patterns in the tissue is only qualitatively addressed. These heatmaps of tumor regions, however, are crucial to the reliability of CAD systems as they provide explainability to the system's output and give confidence to pathologists that the model is focusing on medical relevant features. Motivated by this, we propose a novel weakly-supervised deep-learning model, based on self-learning CNNs, that leverages only the global Gleason score of gigapixel whole slide images during training to accurately perform both, grading of patch-level patterns and biopsy-level scoring. To evaluate the performance of the proposed method, we

perform extensive experiments on three different external datasets for the patch-level Gleason grading, and on two different test sets for global Grade Group prediction. We empirically demonstrate that our approach outperforms its supervised counterpart on patch-level Gleason grading by a large margin, as well as state-of-the-art methods on global biopsy-level scoring. Particularly, the proposed model brings an average improvement on the Cohen's quadratic kappa (κ) score of nearly 18% compared to full-supervision for the patch-level Gleason grading task. This suggests that the absence of the annotator's bias in our approach and the capability of using large weakly labeled datasets during training leads to higher performing and more robust models. Furthermore, raw features obtained from the patch-level classifier showed to generalize better than previous approaches in the literature to the subjective global biopsy-level scoring.

4.1 Introduction

Prostate cancer is one of the major diseases affecting men worldwide. It accounts for 14.5% of all cancers in men [65] and, according to the World Health Organization, its yearly incidence will increase to 1.8 million cases this decade [66]. The Gleason grading system [67] is the main tool for its diagnosis and prognosis. This system describes different stages of cancer based on the morphology and distribution of glands in prostate biopsies. Specifically, the Gleason grades (GG) observable in histology samples range from 3 (GG3) to 5 (GG5). Fig. 4.1 shows representative patterns of each grade.

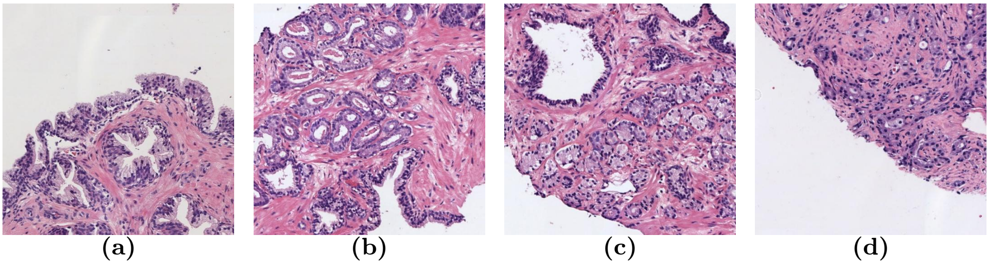


Figure 4.1: Histology regions of prostate biopsies. (a): region containing benign glands, (b): region containing GG3 glandular structures, (c): region containing GG4 patterns, (d): region containing GG5 patterns. GG: Gleason grade.

In order to make a diagnosis of prostate cancer, pathologists extract small portions of tissue, which are laminated and stained using Hematoxylin and Eosin. Then, the slides are carefully analyzed under the microscope to grade local glandular patterns according to the Gleason grading system. Finally, the two most prominent grades in terms of proportion and severity are used to obtain a global Gleason score as prognosis marker. For instance, the Gleason score $3 + 5 = 8$ would be assigned to a sample in which the main cancerous Gleason grade is GG3 and the second is GG5. Recently, after the 2014 conference of the International Society of Urological Pathology, a new grading system referred to as Grade Group [124] has been adopted. This systems takes into account the different prognosis between patients with Gleason score $3 + 4 = 7$ and $4+3 = 7$, including them to different groups (Grade Group 2 and 3, respectively). The whole diagnostic process is highly time-consuming, and is characterized by a large variability among pathologists [103]. These limitations have motivated the development of automatic tools to analyze whole slide images in recent years.

Computer-aided diagnosis (CAD) systems based on computer vision algorithms are able to support pathologists in the daily analysis of prostate biopsies. However, the development of these applications is limited, mainly due to the high data-demanding nature of deep learning algorithms, the large size of digitized biopsies (known as whole slide images (WSIs)) and the difficulty in obtaining pixel-level labeled histology images [104]. Current CAD systems are usually developed to classify local cancerous regions, which are finally combined into a global score. In the case of prostate cancer, this requires manual annotation using multi-resolution graphical user interfaces to accurately delimit the cancerous structures using the Gleason grading system. This is a laborious process, prone to error due to pathologists discouragement, which could incorporate the annotator's bias for certain patterns. Moreover, heterogeneous epithelial cancer such as prostate cancer requires a large number of samples to cover the wide range of possible patterns, which is difficult to reach on annotated datasets.

The limitations of using large annotated datasets encourages the development of weakly-supervised methods able to leverage global labels

–easily accessible from the clinical record via the Gleason score– during the training process. Nevertheless, the literature on employing the global Gleason score to develop CAD systems for prostate biopsy grading remains scarce. The main limitation of the proposed approaches is that they focus on the global-level scoring, while the challenge of local grading cancerous patterns is only qualitatively validated, or simply not addressed. It is noteworthy to mention that the classification of local Gleason grades in prostate biopsies is the basis of an explainable prostate CAD system. The resultant heatmaps support the biopsy-level scoring provided by the system, and they demonstrate that the model relies on relevant medical markers. Thus, the accuracy of the proposed methods in this task must be validated, in order to provide confidence to pathologists in the daily-use of CAD systems.

In this work, we propose a novel weakly-supervised learning strategy to perform both, the global scoring of biopsies and the local grading of cancerous structures in the tissue, where learning is driven only by the global Gleason score. To the best of our knowledge, this is the first attempt to accurately grade local cancerous patterns in prostate whole slide images using biopsy-level labels during training. In the following lines, we summarize the main contributions of this paper. First, we propose an end-to-end CNN architecture, based on patch-level inference aggregation, that is able to detect high-confidence cancerous instances in a weakly-supervised multiple-instance learning (MIL) scenario. Then, we propose a self-learning framework that converts the MIL dataset into a pseudo-supervised task, employing the patches predicted by the previous model and a subsequent post-processing label refinement. We empirically demonstrate that weakly-supervised models trained on large datasets are able to generalize better on the patch-level Gleason grading task than supervised models trained in smaller databases with pixel-level annotations. Finally, we predict the global biopsy-level score based on the aggregation of local features by using the models trained for the patch-level Gleason grading.

4.2 Related work

4.2.1 Self-learning

In the context of this work, we refer to self-learning (a.k.a. self-paced learning or self-training) as the training procedure introduced in [125], which aims to use the knowledge of a firstly trained model (usually called teacher) into a second model (known as student). Interest in this technique has grown in recent years due to the promising results obtained in semi- and weakly-supervised learning scenarios. For instance, in semi-supervised learning approaches, the teacher is used to obtain pseudo-labels from non-labeled data, after it has been trained on annotated examples [18, 126, 127]. Afterwards, the student model is trained by integrating the pseudo-labels in the augmented training dataset. To train more robust students, which are also consistent with the teacher, [18] introduces noise to the samples, as well as model noise, while [126] selects the top–K images based on the corresponding classification scores by the teacher. These works also exploit knowledge distillation by transferring the teacher knowledge to either larger [18] or smaller [126] students. In the context of weakly-supervised object localization, several works employ a teacher model to select regions of interest from the image to train student model in a simplified dataset [128–130]. We want to emphasize that the techniques presented here differ from the similarly so-called self-learning methods, which pre-train networks on pretext tasks where both the inputs and labels are derived from an unlabeled dataset [131, 132][1]. Even though both techniques aim at leveraging unlabeled, or weakly unlabeled data, they present fundamental and methodological differences.

Inspired by these previous works, we adopt a self-learning strategy to accurately classify instances using image-level labels from WSIs. Nevertheless, our work differs from these in that our strategy: (1) trains a teacher model on global image labels, (2) uses the teacher predictions to generate pseudo-labels on unlabeled instances at patch-level, and (3)

[1] The terminology used in this work is proposed by analogy to the previous work on self-learning applied to semi-supervised learning in [18].

trains a student model, of the same complexity, on the pseudo-labels generated by the teacher.

4.2.2 Multiple instance learning

The multiple instance learning (MIL) paradigm falls under the umbrella of weakly supervised learning. In this setting, the training instances are grouped in sets, referred to as bags, X, where only the label for an entire bag, Y, is known. Thus, a bag is considered positive for certain class if at least one instance is positive such that:

$$Y_k = \begin{cases} 1, & \text{if } \exists \; x : y_k = 1 \\ 0, & \text{otherwise} \end{cases} \tag{4.1}$$

With the advent of deep learning, recent efforts on this field have focused towards training a feature extractor under the MIL framework using CNNs. Then, a bag-level representation is obtained by aggregation of either the instance-level features (*embedding-based*) or predictions (*instance-based*). Typical aggregation functions include the maximum [36], average, and log-sum-exponential [45] pooling, more advanced min-max mechanisms recently proposed, such as WILDCAT [43], and trainable functions such as AttentionMIL [46]. Most recent works on MIL adress the problem of weakly-supervised segmentation. In this scenario, embedding-based methods are employed to obtain pixel-level predictions via gradient methods (e.g., grad-CAMs [40]), which are later refined via self-training iterative strategies [42, 106]. Nevertheless, it is noteworthy to highlight that weakly-supervised segmentation works with co-dependent instances (pixels), which are merged on combined features in the CNN. Thus, the generalization of these methods to MIL scenarios which use images as instances is not straightforward.

Nonetheless, despite the wide adoption of MIL in computer vision, its use in prostate histology images still remains scarce. There have been only few attempts to resort to the MIL paradigm in this scenario, which are detailed on the following section.

4.2.3 Gleason grading in prostate histology images

A reliable Computer-Aided Diagnosis system in prostate cancer using histology biopsies aim two main tasks: the global scoring of slides and the quantification and localization of cancerous tissue, both using the Gleason grading system. Due to the large dimension of WSIs and the computational limitations of CNNs, the basis of these systems is the use of small patches extracted from the slide. The proposed methods in the literature can be divided into two categories: bottom-up frameworks, which perform a patch-level classification of Gleason grades using pathologists annotation, and top-down methods, which perform a pseudo-labeling of the patches based on the global Gleason score of the sample.

First works in this field focused on bottom-up frameworks. They usually fine-tune well-known CNN architectures in a supervised patch-level classification [62, 83–85]. Note that these methods require pixel-level expert annotations to obtain the ground truth. Recently, different approaches have been proposed in the literature to overcome the need of pixel-level annotations of Gleason grades. These methods are based in a top-down strategy, where global labels (easily accessible from the patient clinical record) are assigned to local regions of interest. In this way, a weakly-supervised patch-level classification model is trained using the pseudo-labels obtained from global images. In this vein, Campanella et al. [47], under the MIL formulation, assigned the global label (cancerous against non-cancerous) to all the patches of the slide, resulting in a considerable amount of noisy labels. In [78] and [118], color-based filtering was employed to select only patches with high presence of nuclei in cancerous slides, and Ström et al. [119] followed a similar strategy using Laplacian filters. Bulten et al. [117] proposed a semi-supervised pipeline and discarded patches that presented low amount of epithelial tissue. In that work, glandular tissue was previously segmented using an UNet trained using pixel-level annotations. Although these works provided promising results for the global biopsy-level scoring, the patch-level classification was not quantitatively validated. Few works only performed a qualitative evaluation of the produced heatmaps. In these methods, the localization of Gleason grades in the tissue could be affected by the assumptions made to obtain the patch-level pseudo-labels. Accurate localization of Gleason grades in the tissue is a major task that CAD systems should address. The produced heatmaps provide explainability to the system, and ensures that the output is based on medical factors to entrust pathologists in their daily use. Contrary to these works, we propose a teacher model based on instance-level MIL to infer patch-level Gleason grades from bag (biopsy)-level Gleason scores.

Regarding the global scoring of biopsies, the main approach is based on aggregating the patch-level predictions of Gleason grades via the percentage of each grade in the tissue. Particularly, the different models proposed to predict the global Gleason score or ISUP group include: threshold strategies [83, 117], a k-nearest neighbor model [133], a multilayer perceptron [62] or random forests [119].

4.3 Methods

The methodological core of the proposed approach is a self-supervised CNN classifier able to grade prostate histology patches using only the biopsy-level Gleason score during training. The proposed workflow, which is composed by a teacher (θ^t) and a student (θ^s) model, is presented in Fig. 4.2. The first model, i.e., (θ^t), classifies high-confidence patches under a noisy multiple instance learning (MIL) paradigm. In this context, a prostate biopsy is considered as a bag X_b containing instances $x_{b,i}$, and the goal is to predict the instance-level labels $y_{b,i}$ when only the biopsy-level labels Y_b are known. Y_b are obtained using the primary and secondary Gleason grades of the biopsy indicated in the Gleason score. Concretely, the non-cancerous (NC), Gleason grade (GG) 3, 4 and 5 classes are included in Y_b as a multi-label one-hot-encoding ground truth. Then, during the second step, the student model (θ^s) resorts to the instance-level pseudo-labels predicted by the teacher model for training on a pseudo-supervised dataset. The details of these steps are given below.

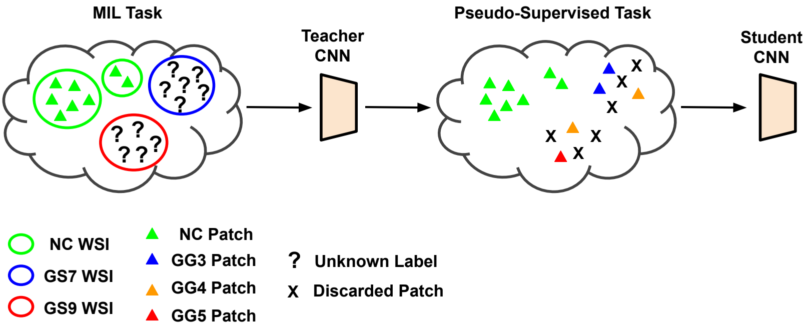


Figure 4.2: Self-learning CNNs pipeline for weakly supervised Gleason grading of local cancerous patterns in whole slide images. MIL: multiple-instance learning; WSI: whole slide image; NC: non-cancerous; GS: Gleason score; GG: Gleason grade.

4.3.1 Teacher model

The teacher model aims to grade high-confidence patches using biopsy-level labels for learning. Formally, let us denote each individual bag as $X_b^t = \{x_{b,1}, ..., x_{b,I}\}$, where $x_{b,i}$ is the i-*th* instance and I denotes the total number of patches, i.e., instances, in the slide. Hence, the objective becomes to predict the global Gleason grade $(\hat{Y}_b^t)$ from the instances $(x_{b,i})$, which can be defined as follows:

$$\hat{Y}_b^t = f(\{x_{b,1}, ..., x_{b,i}, ..., x_{b,I}\}, \theta^t) \tag{4.2}$$

where θ denotes the teacher model weights.

To accomplish the inference of instance-level predictions, the learning process is based on the aggregation of patch-level predictions, i.e., $\hat{y}_{b,i}^t$. Thus, for each instance $x_{b,i}$ in the bag, the teacher model predicts the Gleason grade as follows:

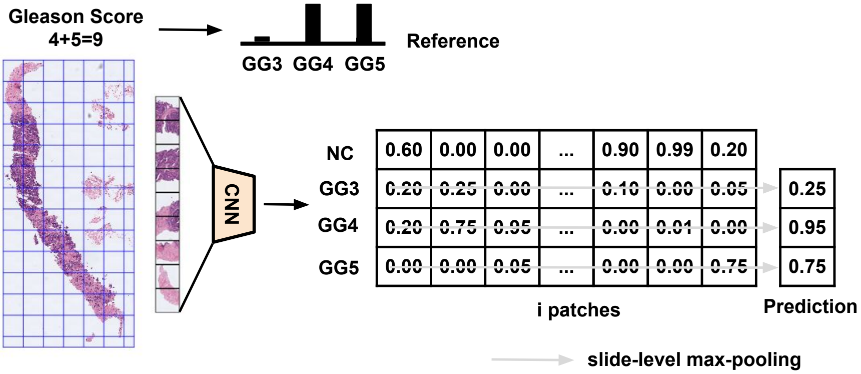
$$\hat{y}_{b,i}^t = f(x_{b,i}, \theta^t) \tag{4.3}$$

Then, we employ an aggregation function $p(\cdot)$ to resume all the instance-level predictions into one representative value that serves as global-level inference. Following this, eq (4.2) can be rewritten as:

$$\hat{Y}_b^t = p(\{\hat{y}_{b,1}^t, ..., \hat{y}_{b,i}^t, ..., \hat{y}_{b,I}^t\}) \tag{4.4}$$

In the context of this work we employ pooling as aggregation function. It is important to mention that the pooling function should be robust to the MIL characteristics. A bag-level class could be positive if just one of the instances is positive for that class. For instance, the use of average pooling would diminish the global cancerous classes activation if the slide contains a large number of non-cancerous patches. Inspired by the properties observed in the max-pooling operation in weakly-supervised segmentation tasks [36], we propose the use of a slide-level max-pooling. Using this operation, the global probability per class is the maximum of the patch-level inferences. This architecture ensures the classification only of high-confidence instances, since gradients in the network are only back-propagated on the instance with largest entropy.

Finally, multi-label binary-cross entropy loss is used during training for gradient estimation. Concretely, the loss is obtained using only the cancerous grades under the assumption that all slides could contain non-cancerous regions, but only cancerous slides contain patches with Gleason patterns. A summary of the Teacher model training is illustrated in Fig. 4.3.



NC	0.60	0.00	0.00	...	0.90	0.99	0.20	
GG3	0.20	0.25	0.00	...	0.10	0.00	0.05	0.25
GG4	0.20	0.75	0.95	...	0.00	0.01	0.00	0.95
GG5	0.00	0.00	0.05	...	0.00	0.00	0.75	0.75

Figure 4.3: Teacher CNN for the prediction of local Gleason grades in a multiple-instance learning framework. GG: Gleason grade.

4.3.2 Student model

The student model aims to perform a patch-level Gleason grade prediction based on a pseudo-supervised data set of images, using the teacher model predictions as pseudo-labels. First, all instances from the dataset are predicted using the teacher model. Then, a label refinement process is carried out based on patch-level teacher model predictions ($\hat{y}^t_{b,i}$) and the known global slide-level labels (Y_b). During this process, labels are modified and patches are discarded under the following premises:

$$\hat{y}^t{}_{b,i} = \begin{cases} NC, & \text{if } G \not\subset Y_b \\ G, & \text{if } G \subset \hat{y}^t{}_{b,i} \wedge G \subset Y_b \\ \text{Discarded}, & \text{Otherwise} \end{cases} \tag{4.5}$$

where NC denotes the non-cancerous class, G an undefined Gleason grade, and $\hat{y}^t_{b,i}$ is the hard one-hot encoding of the Teacher model prediction for certain patch i, belonging to the slide b. By doing this, only the patches classified as certain Gleason grade that belong to a slide actually containing that grade are kept for the subsequent learning of the student model. Regarding the non-cancerous patches, these are obtained only from known benign slides. This label refinement post-processing, together with the simplification of the problem from a MIL to a pseudo-supervised framework, allows the student model to better learning feature representations of the patches. Finally, the student model, which has the same architecture as the teacher, is trained minimizing the categorical-cross entropy between predictions and pseudo-labels. To account for class imbalance, class-specific weights are integrated into the loss function:

$$L(\hat{y}^s_{b,i}, \hat{y}^t_{b,i}) = -\frac{1}{C}\sum_{c=1}^{C} w_c(\hat{y}^t_{b,i,c} log(\hat{y}^s_{b,i,c})) \tag{4.6}$$

where C is the total number of classes (i.e. non-cancerous, Gleason grade 3, 4 and 5), $w_c = (C \times N)/N_c$ is the weight corresponding to each class, being N the total number of images and N_c the number of images belonging to class c.

4.3.3 Biopsy-level Gleason scoring

Once the Gleason grades are located in the tissue, we propose to use the specialized features extracted by student model to predict the global Gleason score. Thus, if we denote ${z^s}_{b,i}$ the features extracted by the student model for each patch, the slide-level feature representation is obtained by global averaging the instance-level features as follows:

$$z_b^s = \frac{1}{I} \sum_i z_{b,i}^s \tag{4.7}$$

Then, two different models are used to predict both the global Gleason score and Grade Group. First, a simple multi-layer perceptron (MLP) composed of one hidden layer with 64 neurons followed by a ReLU activation is used to predict the one-hot-encoding of the global labels. In this case, soft-max activation is used in the output layer and the weights are optimized using the categorical cross-entropy loss. Regarding the second model, k-Nearest Neighbors (kNN) is employed to compare the generalization capability of neural networks and non-parametric models for this task.

4.4 Experimental setting

4.4.1 Dataset

The experiments described in this paper are conducted using several public datasets, which are well known in the prostate cancer histology literature. Concretely, two different datasets of prostate WSIs are used to validate the global biopsy-level methodology, while three databases with pixel-level annotations are used to test the local Gleason grading capability of the proposed methods.

Regarding the datasets used to validate the biopsy-level classifier performance, the recently released dataset from the MICCAI 2020 PANDA challenge [134] is used to evaluate the proposed algorithms. This dataset consists of $10,415$ prostate WSIs whose primary and secondary

Gleason grades have been labeled by expert pathologists. The gigapixel images were resampled to 10× resolution, and randomly clustered into three groups for training, validating and testing. Further, the external SICAP[2] database presented in [62] is used for testing. This dataset consists of 155 prostate WSIs with both global primary and secondary Gleason grades annotated by expert pathologists. The obtained splits for the PANDA dataset, as well as the Gleason score distribution across both datasets are presented in Table 4.1.

Partition	**NC**	**GS6**	**GS7**	**GS8**	**GS9**	**GS10**
PANDA						
Train	2,297	2,122	2,075	1,002	874	99
Validation	98	89	85	42	33	6
Test	497	455	425	205	190	22
Total	2,892	2,666	2,585	1,249	1,097	127
SICAP						
Test	36	14	45	18	35	7

Table 4.1: Datasets of prostate biopsies used. Whole slide images partition and Gleason scores (GS) distribution.

In order to validate the capability of the proposed methods to grade local cancerous patterns, three different external datasets containing pixel-level annotations of Gleason grades are used. Concretely, the test cohort from SICAP dataset and the ARVANITI [83] and GERTYCH [77] databases[3] are used. For these sets, patches of size 512^2 pixels are extracted at 10× resolution. This choice is motivated by prior literature, which determined this configuration as the most optimum for the binary cancer vs. no cancerous supervised classification task [81]. Furthermore, the main study on supervised learning used for comparison, i.e., [62], employs the same patch size, which makes direct comparison easier. Even though this image size might be considered large, benign fusiform or dilated glands, and cribriform GG4 structures may come to have sizes in this range, and smaller patch size could impede the visualization of complete glandular structures. For SICAP dataset, the label was assigned by majority voting of the pixel-level annotations, and for ARVANITI and GERTYCH datasets, only patches containing one

[2] SICAPv2 dataset is accessible at: http://dx.doi.org/10.17632/9xxm58dvs3.2.

[3] ARVANITI and GERTYCH datasets were obtained upon request of corresponding authors of [83] and [77], accordingly.

Gleason grade were used. The non-cancerous class is assigned to patches containing only benign annotations. The data source and the number of patches from each dataset, as well as the Gleason grade distribution are presented in Table 4.2.

Database	Source	NC	GG3	GG4	GG5
SICAP	Biopsies	644	393	853	232
ARVANITI	Tissue Micro-Arrays	115	274	210	104
GERTYCH	Prostatectomies	32	95	216	70

Table 4.2: Datasets with patch-level Gleason grade annotations used for testing. Distribution of the patches among non-cancerous (NC) and the different Gleason grades (GG).

The three external databases were normalized to homogenize the color distribution to the PANDA database. More concretely, the method presented in [101] was used after applying a channel-wise histogram matching to the images from the external databases to a PANDA reference image. This image was selected by the pathologists involved in this work based on its structural and stain properties.

4.4.2 Metrics

In order to evaluate the different approaches, we resort to accuracy (ACC), f1-score (F1S) per class and its average, and Cohen's quadratic kappa (κ). The last metric, κ, is the main figure of merit typically used in prostate Gleason grading. It takes into account that the Gleason system consists on a set of ordered classes, and errors between adjacent classes should be less penalized. In addition, precision and sensitivity are obtained for the non-cancerous class for better understanding of the Teacher-Student pair behavior.

4.4.3 Implementation details

The patch-level classification of Gleason grades was obtained using the self-learning pipeline detailed in Sections 4.3.1 and 4.3.2. The teacher model takes tiles of size 256 × 256 as input, and uses VGG16 architecture as backbone with weights pre-trained on Imagenet for the feature extraction stage. Then, a global-average pooling and a dense layer with soft-max activation is used as top model. The proposed Teacher model is based on the aggregation of patch-level outputs (Gleason grade predictions). Due to the variable amount of instances in a biopsy I, the architecture can not be trained using a mini-batch strategy. Thus, the learning process was carried out using a batch size of 1 slide. The number of patches per slide varies from 40 up to 300 in the training set. In order to avoid computational limitations, up to 200 random patches were used in each iteration. During the learning stage, the teacher model was trained during 30 epochs by using SGD optimizer. The learning rate (η) was initialized to $1 \cdot 10^{-2}$, whose value is decreased by 10 after half the iterations. Then, an exponential decay was applied during the last 5 epochs to stabilize the weights such that $\eta = 1 \cdot 10^{-3} \cdot e^{-0.1 \cdot t}$, where t is the epoch number. The student model was trained following the same procedure than the teacher model, i.e., same number of epochs, optimizer, and learning rate schedule. The global scoring of biopsies was carried out using the features extracted by the student model as detailed in Section 4.3.3. For the multi-layer perceptron model (MLP), Adam was applied as optimizer, using a learning rate of $1 \cdot 10^{-2}$ during 20 epochs. Also, a k-Nearest Neighbors (kNN) model was fitted using a $k = 20$, optimized on the validation set. The proposed methods were implemented in Python 3.6 using Pytorch. The scripts to reproduce the results reported in this work are publicly available on (`https://github.com/cvblab/self_learning_wsi_prostate`).

4.4.4 Baselines

In order to compare our proposed approach with previous state-of-the-art models, baseline methods are implemented for both the local grading and the biopsy-level scoring.

First, regarding the weakly-supervised patch-level Gleason grading, a model similar to [47] for MIL classification was implemented. This method, hereafter referred to as Global-Assignment, consists in assigning the global label to each patch of the WSI, and training a CNN using this pseudo-supervised dataset. In order to reduce noise on the pseudo-labels, only slides with one unique Gleason grade were used during training. Regarding the network employed, we trained a model using the same architecture, optimizer and learning rate schedule than we used in our student model. Furthermore, state-of-the-art methods for MIL aggregation were used in the teacher model to compare to the proposed max-pooling operation. Concretely, recent methods focus on using attention-based mechanisms for embedding-based aggregation in binary MIL classification tasks [46]. The gated attention aggregation, referred to as AttMIL, was adapted to the instance-based multi-class aggregation use-case to obtain global predictions per each class k, Y_k, from the instance-level predictions, $y_{i,k}$ such that: $Y_k = \sum_i a_{i,k} y_{i,k}$. Thus, the attention weights, $a_{i,k}$, determine the contribution of each patch i in the global prediction for each class via the features extracted by the CNN, $\mathbf{z}_i \in \mathbb{R}^M$, and the trainable parameters $\mathbf{V} \in \mathbb{R}^{L \times M}$, $\mathbf{U} \in \mathbb{R}^{L \times M}$ and $\mathbf{W} = [\mathbf{w}_0, ..., \mathbf{w}_{K-1}] \in \mathbb{R}^{L \times K}$ as follows:

$$a_{i,k} = \frac{exp\{\mathbf{w}_k^\top (\tanh(\mathbf{V}\mathbf{z}_i) \odot sigm(\mathbf{U}\mathbf{z}_i))\}}{\sum_{i,k} exp\{\mathbf{w}_k^\top (\tanh(\mathbf{V}\mathbf{z}_i) \odot sigm(\mathbf{U}\mathbf{z}_i))\}} \tag{4.8}$$

where $\tanh(\cdot)$ and $sigm(\cdot)$ are non-linearity functions, and $\odot$ an element-wise multiplication. The number of features extracted by the CNN is $M = 512$ per instance, which are reduced to $L = 128$ during the attention mechanism.

The global Gleason score and Grade Group was predicted also using previously proposed methods based on the percentage of each cancerous grade in the tissue (GG%) using a kNN model as in [133] and a MLP used in [62]. It is noteworthy to mention that other learnable aggregation functions were tested to obtain the embedding of instance-level features instead of the proposed average pooling. In particular, AttentionMIL [46] and miGraph [135] were evaluated. Nevertheless, these methods did not perform properly. Obtaining the Gleason score, by its very definition, involves obtaining the percentage of cancerous patterns in the biopsy, which does not match the formulation of the MIL methods (Equation 4.1). It is noteworthy to mention that, in the weakly-supervised patch-level grading, we solve this limitation by using the presence of Gleason grades as global label, which fits the MIL formulation. We also performed an extensive comparison with previous results obtained in the same test subsets in [62]. In that work, referred to as Supervised, a supervised CNN is trained using pixel-level annotations of Gleason grades performed by expert pathologists on WSIs from SICAP database.

4.5 Results

4.5.1 Grading of local patterns

The figures of merit obtained using the teacher and student models on the different external datasets are presented in Table 4.3. In this table, we also report the results obtained in [62], who resort to supervised training, and those achieved by employing the the baseline approaches. Furthermore, we include the confusion matrices associated to the obtained results in Fig. 4.4.

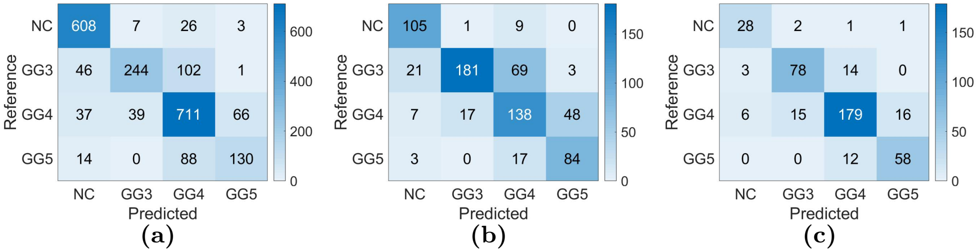


Figure 4.4: Confusion Matrix of the patch-level Gleason grades prediction done by Student CNN on the different test cohorts. (a): SICAP; (b): ARVANITI; (c): GERTYCH.

First, we will focus on the discussion of the results obtained by the different weakly supervised settings and the behavior of the teacher-student pair for this task. We can observe that teacher model achieved an inter-dataset average κ of 0.69 and 0.71 using max and AttMIL as aggregation functions, respectively. This represents a significant improvement compared to the Global-Assignment model (average κ = 0.47), which is limited by the noise introduced in the patches labeled from cancerous biopsies using the global label. Although similar inter-dataset κ is obtained for max and AttMIL aggregation functions in the teacher model, the former shows most promising results for Gleason grades

Method	**Grading**							**Binary C/NC**	
	ACC	**F1**					κ	**Sen.**	**Prec.**
		NC	GG3	GG4	GG5	Avg.			
Other settings									
Arvaniti et al. [83] (2018)*	–	–	–	–	–	–	0.55/0.49	–	–
Nir et al. [84] (2018)**	–	–	–	–	–	–	0.60	–	–
SICAP									
Supervised [62] (2020)	0.67	0.86	0.59	0.54	0.61	0.65	0.77	-	-
Global Assignment	0.505	0.075	0.440	0.744	0.095	0.338	0.465	0.732	0.039
Teacher - Max	0.722	0.788	0.642	0.642	0.217	0.604	0.636	0.663	**0.972**
Teacher - AttMIL	0.655	0.657	0.544	0.768	0.483	0.613	0.725	0.911	0.513
Student - Max	**0.797**	**0.901**	**0.714**	**0.798**	**0.601**	**0.754**	**0.830**	0.862	0.944
Student - AttMIL	0.663	0.653	0.563	0.760	0.544	0.630	0.728	**0.938**	0.501
ARVANITI									
Supervised [62] (2020)	0.586	0.566	0.685	0.469	0.560	0.570	0.641	-	-
Global Assignment	0.554	0.017	0.674	0.612	0.205	0.377	0.501	0.644	0.008
Teacher - Max	0.705	0.726	0.730	**0.682**	0.666	0.701	0.756	0.589	**0.947**
Teacher - AttMIL	0.655	0.271	0.725	0.647	0.725	0.592	0.716	0.760	0.165
Student - Max	**0.722**	**0.836**	**0.765**	0.623	**0.702**	**0.731**	**0.793**	**0.772**	0.913
Student - AttMIL	0.635	0.126	0.712	0.626	0.733	0.549	0.689	0.727	0.069
GERTYCH									
Supervised [62] (2020)	0.513	0.290	0.616	0.499	0.495	0.475	0.511	-	-
Global Assignment	0.562	0.014	0.693	0.761	0.267	0.434	0.531	0.462	0.007
Teacher - Max	0.789	0.697	0.795	0.835	0.666	0.748	0.694	0.555	**0.937**
Teacher - AttMIL	0.680	0.162	0.639	0.727	0.771	0.575	0.693	0.600	0.093
Student - Max	**0.830**	**0.811**	**0.821**	**0.848**	0.800	**0.820**	**0.826**	**0.756**	0.875
Student - AttMIL	0.707	0.111	0.636	0.760	**0.850**	0.589	0.731	0.500	0.062

* Results reported on different patch size and resolutions.

** Results reported on a different (private) dataset.

Table 4.3: Results for the patch-level Gleason grades classification performed by the different approaches on the different test cohorts. The metrics presented are the accuracy (ACC), the F1-Score (F1) and the Cohen's quadratic kappa (κ). Furthermore, precision and sensitivity are indicated for the non-cancerous class. Bold numbers highlight the best performing method. NC: non cancerous, GG: Gleason grade.

differentiation. In particular, AttMIL achieves an inter-dataset average F1 of 0.6360, 0.7140 and 0.6601 for GG3, GG4 and GG5, respectively. This shows the benefit of using attention mechanisms when training the teacher model, which enforces the model to focus on different patches to aggregate the instance-level predictions. Nevertheless, results shift when training the student model using the teacher model's predictions as pseudo-labels. We can observe that the student model using max aggregation obtained and inter-dataset average κ of 0.82 and a F1 of 0.77, an improvement of 8% and 12%, respectively, compared to its corresponding teacher model. On the other hand, the student model does not show any improvement when using AttMIL as aggregation function.

In the interpretation of these results, the process of label refinement between the teacher and the student models (see Section 4.3.2) plays a fundamental role. In this process, false negative patches from positive bags are discarded during training, while false positive instances cannot be detected. These instances classified wrongly as cancerous by the teacher model are the main source of noise in the student model training. Furthermore, we observe that max-pooling aggregation results on an inter-dataset precision in the detection of cancerous instances of 0.95, whereas AttMIL aggregation obtains only 0.26. This difference can be explained by the fact that the slide-level max pooling operation in the teacher model architecture produces backpropagation of the weights for only cancerous patterns that the model classifies with high confidence. Although this phenomenon increases the number of false negative for cancerous classes, these samples are discarded during the aforementioned label-refinement process. Thus, our proposed framework using max-pooling as MIL aggregation function in the teacher model does not introduce noise during the pseudo-labeling process, which results on a better performance of the student model.

Regarding previous state-of-the-art methods for patch-level Gleason grading based on supervised training on pixel-level annotations, our proposed teacher-student model using max-aggregation compared also favorable. In the supervised method in [62], which employs a CNN trained on annotated patches from SICAP database, a consistent drop in the κ metric was observed across the three datasets: SICAP test subset ($\kappa = 0.77$) to ARVANITI ($\kappa = 0.64$) and GERTYCH ($\kappa = 0.51$). In contrast, our weakly-supervised model obtained similar results in the three external datasets, with $\kappa = 0.83$ on SICAP, $\kappa = 0.79$ on ARVANITI and $\kappa = 0.82$ on GERTYCH. These values demonstrate that the proposed weakly supervised pipeline outperforms the methodology presented in [62], showing a higher generalization ability and requiring a weaker supervision during training. This comparison extends to other prior works using supervised learning, which reach k values of 0.55/0.49 in [83] and 0.61 in [84] under different setups. The reason for superior performance of the proposed weakly-supervised strategy could be due to the bias of the annotator produced in the supervised learning scenario, which is not present when using global-labels in our pipeline. Furthermore, the difficulty of obtaining large annotated datasets with heterogeneous patterns can also reduce the performance of fully supervised learning approaches. These benefits over previously proposed methods in the literature outweigh the disadvantages of the proposed strategy. In particular, we might identify as potential limitations the large computational requirements of processing all the patches of a biopsy in each iteration during the training of the teacher model, and the need to use large datasets for correct generalization. These drawbacks, however, are an inherent characteristic of weakly supervised strategies.

Finally, we would like to highlight the limitations observed to evaluate patch-level Gleason grading models in different, heterogeneous datasets. Although similar κ values were obtained across the three external datasets, differences can be observed when focusing on concrete figures of merit. For instance, the best results are obtained on GERTYCH dataset (average F1 of 0.82), whereas the worse results are reported on ARVANITI dataset (average F1 of 0.73). Although the overall results are similar on SICAP dataset (average F1 of 0.75), the student model performs poorly on the GG5 class (F1 of 0.60) and gives the best results on the NC class (F1 of 0.90). These differences could be related to different reasons. For instance, in each dataset the balance of the classes is not equal (see Table 4.2). Precisely, SICAP dataset presents a larger proportion of NC patches, while the proportion of GG5 cases is lower. In addition, SICAP dataset contains patches with mixed Gleason grades, whose label is assigned by majority voting. In these cases, the CNN could be mixing the features of the different classes, thus hampering the obtained performance. Examples of these patches are presented in Figure 4.5. Another limitation for Gleason grading assessment is the well-known inter-pathologist variability. Thus, specific patterns could be annotated with different Gleason grades by pathologists. This variability was quantified at patch level by Arvaniti et al. [83], obtaining a κ of 0.65. This fact enhances the importance of testing the proposed methods across different datasets to ensure the generalization capability of the CAD systems for Gleason grading.

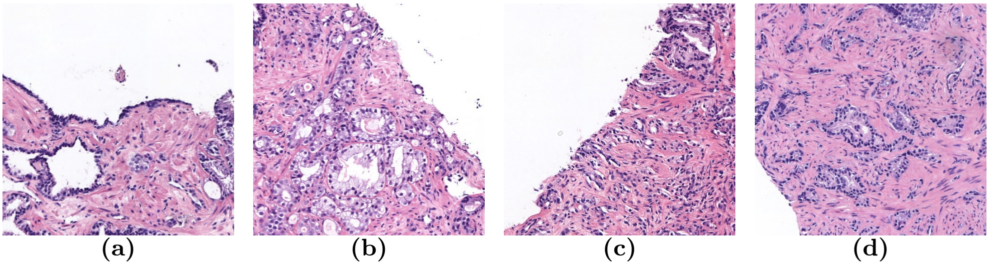


Figure 4.5: Histology regions with mixed Gleason grades from SICAP dataset. (a): region containing benign and GG3 glands, (b): region containing GG3 and GG4 glandular structures, (c) and (d): region containing GG4 and GG5 patterns. GG: Gleason grade.

4.5.2 Biopsy-level scoring

Table 4.4 reports the results obtained by the proposed approaches based on the student features, as well as those from the baseline methods (GG%). Also, results reported in previous works are indicated. Similarly, the confusion matrices of the Grade Group predictions using Student model as feature extractor are illustrated in Fig. 4.6.

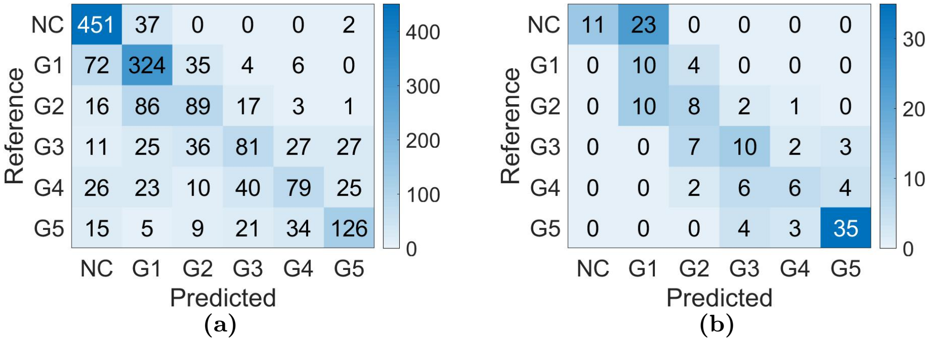


Figure 4.6: Confusion Matrix of the biopsy-level Grade Group prediction done by Student CNN features and k-Nearest Neighbor classifier, on the two test cohorts. (a): PANDA; (b): SICAP

Our proposed model based on aggregating patch-level features and kNN classifier outperformed previously state-of-the-art methods based on the use of the percentage of each Gleason grade in the tissue, reaching an

Method	**Gleason Score**	**Grade Group**
Other settings		
Arvaniti et al. [83] (2018)	0.75/0.71	–
Bulten et al. [117] (2020)	–	0.85*/0.72**
Strom et al. [119] (2020)	–	0.91*/0.82**
Otalora et al. [136] (2020)	–	0.44***
PANDA		
GG% + kNN	0.7936	0.8152
GG% + MLP	**0.8054**	0.8229
Features + Average + kNN	0.7773	0.7927
Features + Average + MLP	0.7954	**0.8245**
SICAP		
Supervised [62]	0.8177	–
GG% + kNN	0.5942	0.5221
GG% + MLP	0.4861	0.5082
Features + Average + kNN	**0.8299**	**0.8854**
Features + Average + MLP	0.3847	0.4421

* Results on test subset.

** Results reported on external datasets.

*** The used dataset does not include benign biopsies.

Table 4.4: Results of the biopsy-level Gleason scoring in the test subsets. The metric presented is the Cohen's quadratic kappa (κ).

average κ of 0.84 for both datasets. Although the results were similar for all models when testing on biopsies from the same center as in the training cohort (PANDA), the performance of neural-networks-based methods dropped on the external dataset (SICAP). The classification stage on neural networks showed overfitting to the training set characteristics in both Gleason grade percentage calculation (based on patch-level classification) and in the global score prediction using the Student model features. The non-parametric use of raw features and kNN generalized better on external datasets. The obtained results are promising, being most of the errors between adjacent classes. Moreover, the model differentiated well critical cases such Grade Group 2 and 3, whose main difference is the balance between Gleason grade 3 and 4 in the tissue (see Fig. 4.6). It is noteworthy to mention that the different results reported from other works are evaluated under different datasets and training conditions. Thus, direct comparison to those works is unfair.

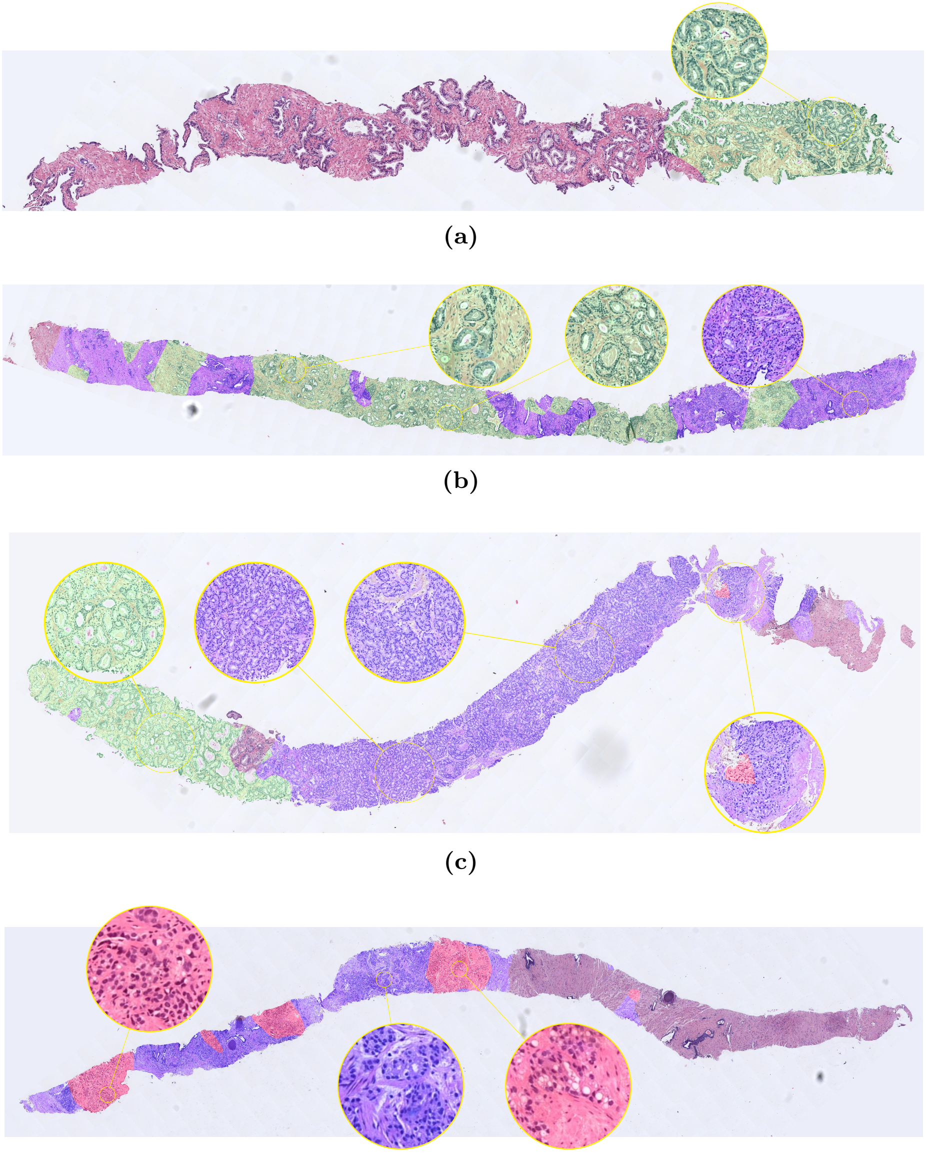
The obtained results are in line with previous literature for global Gleason scoring. The proposed method is comparable against works that use strong supervision via pixel-level annotations, i.e., Arvaniti et al. [83] (κ = 0.75 for Gleason scoring) and Bulten at al. [117] (κ = 0.85 for Grade Group scoring), as well as works that use only global biopsy-level labels, i.e., Strom et al. [119] (κ = 0.91 for Grade Group scoring) and Otálora et al. [136] (κ = 0.44 for Grade Group scoring). In accordance with the observations in our work, methods based on the Gleason grades proportion in the tissue suffer a performance drop on external datasets (κ = 0.72 in Bulten et al. and κ = 0.82 in Strom et al.) Finally, we would like to highlight the difficulty of establishing comparisons among different datasets, since most of the presented results are at the level of inter-pathologist variability for Gleason scoring. Different works have quantified the inter-observer variability on kappa values of 0.71 by Arvaniti et al. [83] or ranging 0.726-0.869 by Bulten et al. [117].

4.5.3 Qualitative evaluation

Representative examples of the obtained results using the CAD system on the external SICAP dataset are presented in Fig. 4.7. The pixel-level heatmaps are obtained by bilinearly interpolating the patch-level predicted probabilities of the closest patches in terms of euclidean distance. Then, the class with highest probability is assigned to each pixel. This figure is organized as follows: each row constitutes a case, and the Gleason grades in the tissue are highlighted in different colors. Also, the Grade Group predicted and the ground truth are indicated.

4.6 Conclusions

In this work, we have proposed a novel self-learning CNN strategy to perform both Gleason grading of local cancerous patterns and global scoring of prostate WSIs using only the global Gleason score during training. Our proposed framework is composed by a novel teacher model based on max-pooling of patch-level inferences of Gleason grades able to perform local classification of Gleason grade using biopsy-level labels. Based on the output of the teacher model and a label refinement post-processing, we propose the training of a patch-level student model on a pseudo-supervised dataset. In the experimental stage, we validate the



(a)

(b)

(c)

(d)

Figure 4.7: Examples of the proposed CAD system based on Self-Learning CNNs performance on the external SICAP dataset. In green: Gleason grade 3; blue: Gleason grade 4 and red: Gleason grade 5. The reference and predicted Grade Group (Reference - Predicted) are: (a): (2 - 1); (b): (2 - 2); (c): (3 - 3); (d): (5 - 5).

patch-level classification on three different external datasets. The student model reaches an inter-datasets average Cohen's quadratic kappa (κ) of 0.82 and an f1-score of 0.77. Our results outperformed previous works based on supervised learning with pixel-level annotations. Moreover, the results between the different test cohorts were similar, while previous supervised methods experimented a drop in performance when testing on external test images. Our proposed weakly-supervised method generalizes better than supervised methods for local Gleason grading, due to the absence of annotator bias and the capability of being trained on large heterogeneous datasets.

Then, the features learned by the patch-level trained models were used to predict the global Grade Group via an average aggregation and a linear classification layer. The method was tested on two different datasets, reaching an average κ of 0.84. This method was compared with the main approach in the literature for Grade Group prediction using the percentage of the different Gleason grades in the tissue. Our feature-based model showed to better generalize pathologist scoring biopsies than previous approaches.

The promising results presented in this work represent a significant advance in the literature of prostate histology. Using weakly-supervised learning it is possible to grade local patterns in gigapixel WSIs outperforming supervised methods which require laborious annotations by expert pathologists. Further research will focus on studying and improving the image-normalization process of prostate histology samples to use CAD systems in external datasets, which is a vital step for a successful generalization. Also, the proposed weakly-supervised models will be refined to decrease the number of biopsies required during training.

Chapter 5

Proportion constrained weakly supervised histopathology image classification

The content of this chapter corresponds to the author version of the following published paper: Silva-Rodríguez, J., Schmidt, A., Sales, M.A, Molina, M., & Naranjo, V. Proportion constrained weakly supervised histopathology image classification. Computers in Biology and Medicine, (2022).

Contents

Proportion constrained weakly supervised histopathology image classification

Julio Silva-Rodríguez[1], Arne Schmidt[2], Maria A. Sales[3], Rafael Molina[2] and Valery Naranjo[4]

[1]Institute of Transport and Territory, Universitat Politècnica de València, Valencia, Spain; [2]Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain; [3]Anatomical Pathology Service, University Clinical Hospital of Valencia, Valencia, Spain; [4] Institute for Research and Innovation in Bioengineering, Universitat Politècnica de València, Valencia, Spain

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 log-barrier 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.

5.1 Introduction

In the supervised learning paradigm, deep learning methods have shown promising performance in a wide range of medical imaging applications. Nevertheless, these methods usually require large amount of data for training, which must be labeled by expert clinicians. Obtaining these labeled datasets is a time-consuming process and is susceptible to inter-annotator variability, which complicates the use of these models in practice. This is the case for histology image analysis, whose large size of tissue images magnified on whole slide images (WSIs), patterns heterogeneity, and the high level of expertise required to annotate the data make this learning paradigm unfeasible. Considering these limitations, the most popular choice in this field has become the use of weakly supervised learning strategies under the multiple instance learning (MIL) paradigm. In particular, typically the training dataset is composed of bags (WSIs) that are known to have cancer or not. Each bag consists of instances (tissue tiles), of which the label is not accessible during training. Under this setting, different works have demonstrated outstanding results for both WSI-level cancer detection [47] and instance-level cancer localization [64]. Nevertheless, these methods require very large datasets (i.e. thousands of biopsies) to compensate for the absence of greater supervision. One common limitation is that these methods tend to focus on only a limited number of instances of each bag during training. Very recent literature has resort to instance-dropout [137] during training to alleviate this issue. Despite the improvement it produces, this solution does not involve classifying more positive instances systematically, but depends on the samples randomly discarded in the dropout, without prior knowledge. To improve the performance of MIL models with the help of prior knowledge, constraint deep learning has been proposed using previously estimated tumor size [49] to guide the weakly supervised optimization. Although this method shows a promising performance, in this case the tumor size estimation is a tedious task, which can be as costly as performing instance-level annotations. All these limitations are accentuated in the multi-label scenario, where it is desired to differentiate between different types of tissues, which may coincide in the same bag. In contrast to the binary scenario classification, multi-label MIL literature still remains scarce in histology image analysis [138].

Based on these observations, we propose a novel formulation for MIL in the multi-label scenario, applied to histology prostate cancer grading in WSIs. The key contributions of our work can be summarized as follows:

- A novel constrained formulation for instance-level MIL, which integrates an auxiliary term that forces to increase the number of instances classified on positive classes.
- In addition, our formulation leverages prior knowledge in terms of relative tissue proportions (i.e. primary cancerous grade in the WSI) by imposing inequality constraints on bag(WSI)-level class proportions.
- We benchmark the proposed model against a relevant body of literature on SICAP-MIL, a new publicly available dataset containing 350 prostate WSIs with global labels, as well as instance-level labels to test weakly-supervised methods on tumour localization.
- Comprehensive experiments demonstrate the superior performance of our model. By simply incorporating relative proportion information during training (easily accessible from medical records in many cancer types) we found improvements of nearly $\sim 3\%$ in mean AUC for bag-level classification and $\sim 13\%$ for instance-level cancer grading accuracy compared to prior MIL methods.

5.2 Related work

5.2.1 Multiple Instance Learning

In computer vision, multiple instance learning (MIL) is a learning paradigm that works with independent images (instances) that form groups (bags), and only bag-level information is known. In the multi-label scenario, each instance belongs to one class, but different classes could coincide at bag level [139]. Modern MIL methods using convolutional neural networks (CNNs) for feature extraction usually process each instance independently, and then combine the instance-level information into one bag-level output. Methods that combine instance-level features are known as embedding-based, which require a subsequent classification layer. In contrast, instance-based architectures combine directly instance-level predictions into the bag classification. Beyond the basic mean and maximum aggregation functions, recent methods have proposed the use of weighted-averaged embeddings, using instance-specific attention weights learned via a multi-layered perceptron projection [46] or recurrent neural networks [47]. It is noteworthy to mention that, although embedding-based approaches have yielded slightly better bag-level results in previous literature, they do not provide instance-level probability outputs. In this work, we are interested in both: instance and bag-level classification. Since we aim to include prior knowledge referred to class-wise proportions, our proposed method follows the instance-based learning paradigm.

5.2.2 Constrained classification

Constrained classification aims to guide the training of a CNNs towards a solution that satisfies a given condition, which takes advantage of additional knowledge to the main labels. This learning paradigm has gained popularity on weakly supervised scenarios (e.g. weakly supervised segmentation or MIL), since it allows to incorporate local information to the global annotations. In a usual constraint weakly supervised setting, an additional loss term enforces the sum of the instance-level predictions to match a given proportion using an L_2 penalty [140]. Similarly, it has been applied in unsupervised anomaly segmentation, to force attention maps to focus on all patterns of training images [60], or in semi-supervised learning, to match the predicted size distributions to the ones observed in the supervised subset using a KL-divergence term [141]. While the aforementioned equality-constrained formulations proposed in weakly supervised settings are very promising, they demand exact knowledge of the prior. For instance, in the case of histology tumour grading, this would require to know the cancerous tissue proportion extent. Therefore, recent works have preferred the use of inequality constraints to relax the prior assumptions, allowing more flexibility. This approach allows, for example, to set some tolerance margins on target size using L_2 penalties [142, 143], or Lagrangian optimization [48]. Following the example above, these works would require approximate knowledge of tumor size, and a tolerance margin would be applied to smooth the constraint. Unlike these works on weakly supervised classification, our formulation does not require prior information on the absolute size of the target. In contrast, we seek to constrain the training to account for relative relationships between proportions within the same global image. In the case of histological whole slide image classification in a multi-label setting, this formulation incorporates information about which tumor grade is in the majority (primary) and which is in the minority (secondary), so that the proportion of the primary grade must be greater than that of the secondary grade. Thus, we use inequality constraints to (i) encourage classification of instances to positive classes at the bag level, and (ii) incorporate relative relationships between class proportions within bags.

5.3 Methods

An overview of our proposed method is depicted in Figure 5.1. In the following, we describe the problem formulation, and each of the proposed components.

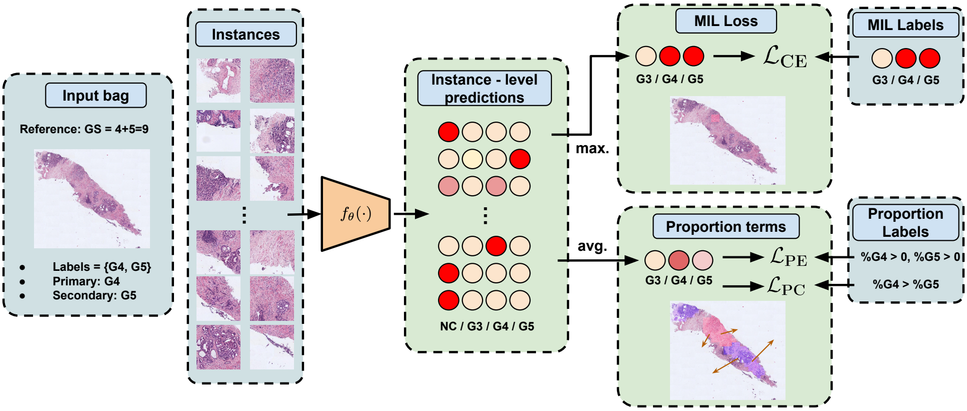


Figure 5.1: Method overview. In this work, we face weakly supervised histology image classification under the Multiple Instance Learning (MIL) paradigm. Each biopsy is a bag, while its patches are the instances conforming it. In the case of prostate analysis, expert labels are conformed by the Gleason score, that are the sum of the two most predominant tumour grades (i.e. G3, G4 or G5). In order to extract both instance and bag-level labels, an standard instance-level MIL with max aggregation is trained via cross-entropy loss, $\mathcal{L}_{ce}$ (see Eq. 5.3). Then, prior information is incorporated via inequality constraints that (i) force the classifier to predict instances that are present in the biopsy ($\mathcal{L}_{PE}$, see Eq. 5.5), and (ii) ensure that the proportion of the primary grade is superior than the secondary grade ($\mathcal{L}_{PC}$, see Eq. 5.7). Colored tissue indicates: blue: Gleason grade 4; red: Gleason grade 5. Circles in instance-level predictions indicate soft-max scores, $y_{n,k}$. The more intense the color, the higher the score.

Problem Formulation In the paradigm of Multiple Instance Learning (MIL), instances are grouped in bags of instances $X = \{x_n\}_{n=1}^{N}$, that exhibit neither dependency nor ordering among them, and its number N is arbitrary for each bag. In the multi-label scenario, there are multiple labels per bag, $Y = (Y_1, ..., Y_k, ..., Y_K)$, where $k \in \{1, ..., K\}$ denotes each one of the K categories. Also, individual labels, $y_{n,k} \in \{0, 1\}$, exist for each instance in the bags, but they remain unknown during training. In the standard MIL formulation, a bag label is considered positive if at least one instance in the bag is positive for that category. We can rewrite this assumption in the following forms:

$$Y_k = \begin{cases} 1, & \text{iff } \sum_n y_{n,k} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{5.1}$$

$$\equiv \qquad Y_k = \max_n \{y_{n,k}\} \tag{5.2}$$

Instance-based MIL In this work, we aim to training a model capable of extracting both: instance and bag-level labels, which falls into the instance-based MIL paradigm[1]. Let us denote a neural network model, $f_{\boldsymbol{\theta}}(\cdot) : \mathcal{X} \rightarrow \mathcal{H}^{K+1}$, parameterized by $\boldsymbol{\theta}$, which processes instances $x \in \mathcal{X}$ to predict softmax instance-level class scores, $\{h_k\}_{k=0}^{K} \in \mathcal{H}$, such that $\mathcal{H} \in [0, 1]$. Note that $k = 0$ represents a category for instances negative at all classes. Also, we use a parameter free aggregation function, $f_a(\cdot)$, in charge of pooling the instance-level scores into one global score $H = (H_1, ..., H_k, ..., H_K)$ such that $H = f_a(\{f_{\boldsymbol{\theta}}(x_n)\}_{n=1}^{N})$. Then, the optimization of $\boldsymbol{\theta}$ is driven by the minimization of cross entropy loss between reference and predicted bag-level score.

$$\mathcal{L}_{ce} = -\frac{1}{K} \sum_{k=1}^{K} Y_k log(H_k) + (1 - Y_k) log(1 - H_k) \tag{5.3}$$

5.3.1 Inequality constraints for MIL

Previous literature on instance-level MIL have proposed aggregation functions $f_a(\cdot)$ based on mean or maximum operator. The second solution is used based on the direct interpretation of maximum operation on MIL formulation (Eq. 5.2). Nevertheless, training a neural network via this aggregation produces well-known problems such as gradient vanishing of non-maximum instances. This limitation produces the network to focus only on discriminative instances during training, which leads to poor generalization performance on unseen samples. To alleviate this issue, we focus on the MIL formulation in Eq. 5.1, which interpretates a positive bag via an inequality that forces the sum of instances scores to be greater than zero. In this line, we incorporate to the base

[1] Based on the denomination proposed in [46]

instance-based MIL training a term that increases the proportion of positive instances classification for a given class k, $p_k = \frac{1}{N}\sum_n^N h_{n,k}$, by minimizing $-\lambda log(p_k)$. Nevertheless, this log-term is non-differentiable when $p_k \to 0$. To solve this limitation we resort to a smooth, duality-gap bound approximation. Concretely, we use the formulation proposed in [48] on constrained optimization that models inequality constraints using the approximation of log-barrier that is formally defined as:

$$\widetilde{\psi}_t(z) = \begin{cases} -\frac{1}{t}\log(z) & \text{if } z \geq \frac{1}{t^2} \\ -tz - \frac{1}{t}\log(\frac{1}{t^2}) + \frac{1}{t} & \text{otherwise,} \end{cases} \tag{5.4}$$

where t *controls* the barrier during training, and z is the objective term.

This log barrier extension is applied on the proportion term p_k of the bags that are positive for the class k at bag level (i.e. $Y_k = 1$). It is noteworthy to mention that this proportion is the objective term z in Eq. 7.4. Hereafter, we refer to this term as positives expansion (PE) constraint.

$$\mathcal{L}_{PE} = \sum_{k:Y_k=1} \widetilde{\psi}_{t_{PE}}(p_k) \tag{5.5}$$

Thus, we propose a MIL loss that combines the maximum formulation in Eq. 5.2 via the aggregation function $f_a(\cdot) = \max_n\{y_{n,k}\}$, and the PE term as follows:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_{PE}\mathcal{L}_{PE} \tag{5.6}$$

where $\lambda_{PE} \in \mathbb{R}^+$ weights the importance of each term during training. Note that the positives expansion term, $\mathcal{L}_{PE}$, is only applied for those positive categories at bag-level.

5.3.2 Incorporating proportion information

In some applications, prior knowledge of the bags is known. In this work, we focus on an information usually recorded on medical domains: data regarding the proportion of categories in the image (i.e. primary or secondary tumor grades in the tissue). This information can be formulated as an inequality constraint between categories proportions such that: $p_{k'} > p_{k''}$, where k' denotes the larger proportion category, and k'' its respective counterpart. Note that this relation can be established between any pair of positive categories in the bag for which we have this information available. Thus, we contemplate an arbitrary number of conditions I for each bag, which could give complete or partial information (i.e. the formulation could be applied for only few known inequalities). For each condition i, both major (k') and minor (k'') categories should be indicated. Again, we make use of extended log-barrier (see Eq. 7.4) to solve this inequality constraint, which has demonstrated good performance when multiple constraints are used [48]. In this case, the objective term z in Eq. 7.4 is the different between major and minor proportions in a given bag: $(p_{k'_i} - p_{k''_i})$. Hereafter, we refer to this additional term as proportion constraint (PC).

$$\mathcal{L}_{PC} = \sum_{i}^{I_b} \widetilde{\psi}_{t_{PC}}(p_{b,k'_i} - p_{b,k''_i}) \tag{5.7}$$

where b indicates the bag index over the complete dataset, $\lambda_{PC} \in \mathbb{R}^+$ weights the relative importance of the proportion term during training, t_{PC} controls the barrier slope over time. It is noteworthy to mention that the proportion term is not taken into account for bags with only one positive category, or which the proportion information is unknown.

Taking into account the different terms previously detailed, $\boldsymbol{\theta}$ is trained to solve the multi-label MIL formulation using the following optimization criteria via standard Gradient Descent:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_{PE}\mathcal{L}_{PE} + \lambda_{PC}\mathcal{L}_{PC} \tag{5.8}$$

5.4 Experimental setting

5.4.1 Datasets

In this work, we present a new dataset for prostate histological image analysis: SICAP-MIL[2]. This dataset is an extension of the previously published SICAP versions [62, 81], which is expanded with 168 new WSIs. The dataset introduced is composed of 350 WSIs from 271 patients. The samples were digitised using the Ventana iScan Coreo scanner at $40x$ magnification. The slides were analysed by a group of expert urogenital pathologists at Hospital Clínico of Valencia, and a combined Gleason score (GS) was assigned per biopsy. The Gleason score is the sum of the two main (primary and secondary) Gleason grades (GG) in the biopsy regarding its extent and severity. The clinical report specifies both the score and the primary and secondary grades that constitute the score. SICAP-MIL is specially design to serve as a benchmark for MIL methods. Each WSI is considered as a bag, from which instances are obtained by tiling the images using non-overlapped moving-windows of 512^2 pixels at $10\times$ of resolution level. Note that tiles with less than 20% of tissue were excluded. The dataset is divided into three class-wise balanced groups for training, validation and testing. A summary of the dataset in terms of the labelled Gleason scores and proposed partitions is presented in Table 5.1.

Partition	**NC**	**GS6**	**GS7**	**GS8**	**GS9**	**GS10**	**Total**
Train	77	10	61	7	25	8	188
Validation	19	2	26	5	10	2	64
Test	17	9	28	13	27	4	98
Total	111	21	115	25	62	14	350

Table 5.1: SICAL-MIL dataset. Whole slide images partition and Gleason scores (GS) distribution.

From the WSI-level Gleason scores, bag-level labels referred to the presence of each Gleason grade in the WSI are inferred. Also, the relative-proportion information of the primary and secondary grades is obtained

[2] SICAP-MIL is available at `https://cvblab.synology.me/PublicDatabases/SICAP_MIL.zip` or on the GitHub repository of the project: `https://github.com/jusiro/mil_histology`.

from this score. We show in Figure 5.2 the information regarding the primary and secondary Gleason grades for each WSI. It is observed that most cases present at least two tumor types, and thus two proportion expansion (PE) constraints and one proportion constraint (PC) in the proposed formulation. Also, the difficulty of training a classifier capable of distinguishing between different Gleason grades in a weakly supervised manner is appreciated, since the biopsy rarely presents a single tumor type.

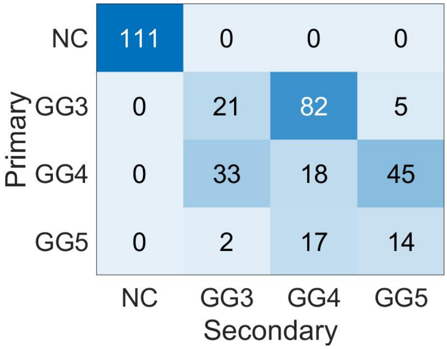


Figure 5.2: SICAP-MIL dataset description. The confusion matrix shows the distribution of global labels in terms of primary and secondary Gleason grades per Whole Slide Image. GG: Gleason grade. NC: non-cancerous.

In addition, SICAP-MIL includes instance-level annotations, which allow to test the capability of MIL methods to leverage instance classifications in a weakly-supervised manner. To do so, annotated WSIs are kept into the test subset. Note that instance-level labels are obtained from pixel-level annotations done by expert pathologist. Non-cancerous patches are obtained only from benign WSIs, while cancerous patch-level labels are obtained by majority voting of segmentation masks. The distribution of instance-level annotated subset from the test cohort is presented in Table 5.2.

Partition	NC	GG3	GG4	GG5
Test	448	289	632	132

Table 5.2: SICAP-MIL patch-level Gleason grade annotations used for testing. Distribution of the patches among non-cancerous (NC) and the different Gleason grades (GG).

5.4.2 Implementation details

The proposed methods were trained using the train subset from SICAP-MIL. The backbone $f_{\boldsymbol{\theta}}(\cdot)$ used was a VGG16 [90] pre-trained on Imagenet [94], which takes as input instances resized to 224×224 images. First, the PE setting was trained by empirically fixing $\lambda_{PE} = 0.1$ and $t_{PE} = 15$. Training was carried out during 100 epochs using a batch size of 1 bag and the SGD optimizer with a learning rate $\eta = 1 \cdot 10^{-2}$. After 50 epochs, η is decreased in a factor to 10×. During training, bag-level mAUC is monitored in the validation set, and early stopping is applied if this figure of merit does not improve during 20 epochs. Then, the PC formulation is trained keeping constant the PE hyperparameters, and empirically setting $\lambda_{PC} = 1$ and $t_{PC} = 5$. The training is carried out using the same training conditions as the PE setting. Nevertheless, instead of using mAUC from validation subset as early stopping criterion, we use the average proportion constraint satisfaction, $z = p_{b,k'_i} - p_{b,k''_i}$ in Eq. 5.7 from the training set to determine the best model. The hyperparameters and early stopping criterion used are further justified by means of ablation experiments. The code and trained models are publicly available on `https://github.com/jusiro/mil_histology`.

Instance-level Student In this work, we complement the proposed models for instance-level prediction with a second model, Student, trained with instance-level hard pseudo-labels as described in [64]. This second stage has demonstrated to increase model performance without any modification of the architecture as described in [64]. Note that we use as Teacher any trained instance-level classifier $f_{\boldsymbol{\theta}}(\cdot)$ under the MIL paradigm with the proposed methodology. A Student model with the same complexity as the Teacher is trained following the Noisy Student paradigm on semi-supervised learning [18]. Concretely, a dropout rate of 0.20 is applied over the instance embedding, and data augmentation is applied to all instances using random rotations, translations, Gaussian blur and color jittery. Student is trained during 60 epochs with mini-batches of 32 images using SGD optimizer and a learning rate of $\eta = 1 \cdot 10^{-2}$.

5.4.3 Baselines

With the aim of comparing our approach to state-of-the-art methods, we implemented and tested prior methodologies on MIL for both instnace-level and bag-level classification on SICAP-MIL dataset. **Instance-based MIL**. First, we compare our method with other instance-based MIL aggregation. Concretely, we use basic mean and max operations over the instance-level predictions to obtain the bag-level prediction. **Embedding-based MIL**. Secondly, we included embedding-based methods, which aim to obtain a bag-level embedding, on which a classifier is trained to predict bag-level labels. Aggregation methods of instance-level features include mean, max, attention mechanism, and recurrent neural networks (RNN). AttentionMIL [46] aims to obtain a weighted feature representation, which highlights positive instances in the bag. The weights are obtained using a multi-layered perceptron as detailed in [46]. We implemented the gated attention mechanism with an intermediate layer with $D = 128$ neurons. Campanella et al. [47] proposed a RNN based aggregation over the top-k positive instances of each bag to produce bag-level classifications. We increased $k = 10$ to support the multi-label scenario, and a RNN with a hidden state of 128 neurons was trained. All methods are train under the same training setup (i.e. backbone, learning rate, scheduler, batch size, etc.) as our baseline. Only the learning rate of the methods based on attention mechanisms was changed to $\eta = 1 \cdot 10^{-3}$. Note that embedding-based method don't make instance-level predictions, and is therefore only used as a comparison of the results at the bag level. Although attention-based methods include instance-level importance weights, these are not true predictions at the instance level, as they are sensitive to the number of instances in the bag.

5.4.4 Evaluation metrics

We evaluate the different models in this work using standard metrics on MIL for both instance and bag-level performance on the test subset. Concretely, for instance-level validation we obtain accuracy (Acc), and f1-score per class and micro-averaged. Also, as the Gleason grades constitutes a set of ordered classes, we obtain Cohen's quadratic kappa (κ) as figure of merit. Regarding the bag-level predictions, we evaluate them using the area under ROC curve (AUC). In the multi-label scenario, AUC is obtained class-wise, and it is averaged (mAUC). In order to facilitate the comparison of our methods with previous literature at the bag level, we also obtained the AUC for binary cancer vs. non-cancer detection by combining each class prediction and target via max-aggregation. For each experiment, the metrics shown are the mean of three consecutive repetitions (with its respective standard deviation) of the model training, to account for the variability of the stochastic factors in the process.

5.5 Results

5.5.1 Comparison to the literature

The quantitative results obtained by the proposed model and baselines on the test cohort are presented at instance level in Table 5.3, and at bag level in Table 5.5 and Figure 5.3. Also, we include results reported in a relevant body of literature for both tasks, using different datasets and experimental settings for instance level in Table 5.4, and at bag level in Table 5.6.

Method	**Acc**	**F1-score**					κ
		NC	**GG3**	**GG4**	**GG5**	**Avg.**	
mean	0.458	0.312	0.383	0.548	0.411	0.413	0.431
max	0.484	0.604	0.295	0.411	0.199	0.377	0.262
max (Student) [64]	0.573	0.716	0.398	0.529	0.320	0.490	0.454
max - w. PE	0.535	0.644	0.259	0.533	0.217	0.413	0.296
max - w. PE (Student)	0.610	0.748	0.302	0.616	0.341	0.502	0.481
max - w. PE w. PC	0.639	0.706	0.686	0.611	0.309	0.578	0.450
max - w. PE w. PC (Student)	**0.705**	**0.818**	**0.692**	**0.691**	**0.417**	**0.655**	**0.655**

Table 5.3: Quantitative comparison to prior literature at instance level on SICAP-MIL dataset. Results derived from the proposed methods in gray. Best results in bold. NC: non-cancerous; GG: Gleason grade; κ: Cohen's quadratic kappa.

Instance-level results. The proposed constrained formulation using a positive expansion constraint term (PE) to enhance positive instances prediction outperforms in $\sim$ 5% the accuracy for instance-level classification of max-aggregation baseline. Adding the Student stage, the model reaches an accuracy of 0.610, which outperforms on SICAP-MIL

Method	Paradigm	Training Dataset		Acc	F1-score					κ
		TMAs	WSIs		NC	GG3	GG4	GG5	Avg.	
Arvaniti et al. [83] (2018)	supervised	508	-	-	-	-	-	-	-	0.67/0.55
Nir et al. [85] (2019)	supervised	333	-	-	-	-	-	-	-	0.61
Silva-Rodríguez et al. [62] (2020)	supervised	-	160	0.67	0.86	0.59	0.54	0.61	0.65	0.77
Otálora et al. [144] (2020)	semi-supervised	508	171	-	-	-	-	-	-	0.59/0.55
Silva-Rodríguez et al. [64] (2021)	MIL	-	10,000	0.797	0.901	0.714	0.798	0.601	0.754	0.830
max - w. PE w. PC (Student)	MIL	-	188	0.705	0.818	0.692	0.691	0.417	0.655	0.655

Table 5.4: Quantitative comparison to prior literature at instance level. Results reported on different datasets, patch size and resolutions. Results derived from the proposed methods in gray. TMAs: tissue micro arrays; WSIs: whole slide images; NC: non-cancerous; GG: Gleason grade; κ: Cohen's quadratic kappa.

Method	Cancer Detection	Multilabel
Embedding + mean	0.952(0.013)	0.844(0.009)
Embedding + max	0.951(0.019)	0.834(0.002)
Embedding + RNN [47]	0.967(0.014)	0.855(0.011)
Embedding + AttentionMIL [46]	0.961(0.006)	0.848(0.007)
Instance + mean	0.701(0.090)	0.769(0.071)
Instance + max	0.955(0.012)	0.867(0.005)
Instance + max w. PE	0.962(0.009)	0.873(0.019)
Instance + max w. PE w. PC	**0.979**(**0.005**)	**0.899**(**0.007**)

Table 5.5: Quantitative comparison to prior literature at bag level on SICAP-MIL dataset. The metric presented is the Area Under ROC curve (AUC). Results derived from the proposed methods in gray. Best results in bold.

Method	Training WSIs	Cancer Detection	Multilabel
Campanella et al. [47] (2019)	24,859	0.994	–
Ström et al. [119] (2020)	6,682	0.997	–
Bulten et al. [117] (2020)	5,759	0.990	–
Li et al. [137] (2021)	9,638	0.982	–
max - w. PE w. PC (Student)	188	0.979(0.005)	0.899(0.007)

Table 5.6: Quantitative comparison to prior literature at bag level. Results reported on different datasets, patch size and resolutions. The metric presented is the Area Under ROC curve (AUC). Results derived from the proposed methods in gray. WSI: whole slide image.

the Teacher-Student strategy using only max aggregation in [64]. The observed improvement could be caused by the larger number of instances classified using the inequality constraint, which avoids over-fitting the model to focus only on very discriminative instances. Note that, although still the results reported in [64] in prior literature are better, the training dataset required to accomplish these results is too large: around

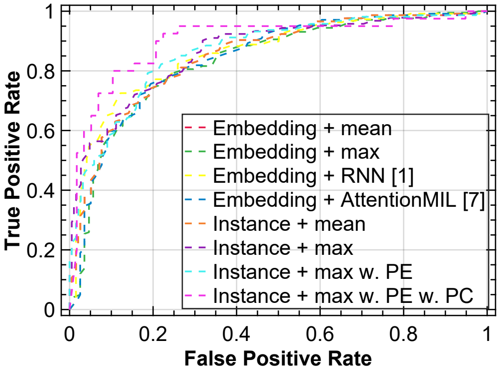


Figure 5.3: Overall receiver operating characteristic (ROC) curves for the multilabel bag-level prediction of proposed methods and baselines on SICAP-MIL dataset.

10, 000 WSIs. Once we introduce the proportion information in terms of primary and secondary classes in the bag via the proportion inequality constraint (PC), results reach an accuracy of 0.705 and average F1-score of 0.655. It is noteworthy to mention that these results are similar to the ones obtained in prior literature under full supervision on similar sized datasets [62, 83, 85, 144]. Under our proposed formulation, the model is capable of grading cancerous patches at the same performance of using pixel-level annotated datasets, by providing only WSI-level information about the most abundant grade.

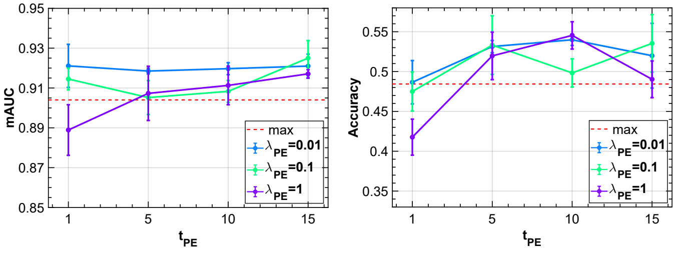
Bag-level results. Regarding the MIL bag-level results obtained, our PE formulation improved around $\sim$ 0.7% the baseline instance-based maximum aggregation. This modest improvement may be due to the fact that, because of the maximum-based inference, it is only necessary to locate one positive sample to get the bag-level prediction right. These observations are in line with previous literature, which highlights that the best classifier at the bag level need not be the best classifier at the instance level [29]. Once we incorporate the proportion information during training, the proposed model increases the multilabel mAUC in $\sim$ 3.3% from the baseline, and reaches mAUC of 0.899 in the multi-label scenario and 0.979 in the binary prediction (see Table 5.5). Note that this result almost reaches the ones reported in previous literature (see Table 5.6), which use thousands of WSIs during training. However, it is worth noting the limitations of this indirect comparison. The methods used in previous works may have different levels of supervision, and the datasets used are larger. Next, we perform a direct comparison of the weakly supervised methods in the database used in this work, SICAP-MIL (see Table 5.5). Specifically, we pay attention to embedding-based methods performance at bag level. The obtained results using mean and max aggregation are similar to the baseline instance-based max approach. However, in the multi-label scenario, these methods perform worse. Moreover, since they cannot provide instance-level labels, they cannot take advantage of the information referred to the proportion during training. It is notable that deep-learning based aggregation modules such as AttentionMIL or RNN do not perform properly in this training setting. This could be due to the complexity of having multiple classes in some bags, the over-fitting tendency of neural networks, and the incapacity of AttentionMIL to get class-specific attention weights. Finally, We would like to point out that a significant body of previous work validates multi-class methods at the bag level on the basis of Gleason scores. However, this score is beyond the scope of MIL. Its derivation involves a decision making according to the severity of the grades in the tissue by the clinical expert, which does not fit a proper formulation of MIL (see Eq. 5.1), based on the presence of each class in the bags of instances.

5.5.2 Ablation studies

In the following, we provide comprehensive ablation experiments to validate several elements of our model, and motivate the choice of the values employed in our formulation, as well as our experimental setting.

First, we optimized the proposed formulation only with the inequality constraint term in Eq. 5.6. Using the training setting previously described, validated different values of $\lambda_{PE} = \{0.01, 0.1, 1\}$ and slopes of the log-barrier inequality $t_{PE} = \{1, 5, 10, 15\}$. Using the mAUC on validation subset as an early stopping criteria, we obtained bag-level mAUC from the validation subset and instance-level accuracy from the test cohort. Results are presented in Figure 5.4. These show that the inclusion of the PE term improves both the performance at both bag-level and instance-level under most of the settings. Thus, we selected

$t_{PE} = 15$ and $\lambda_{PE} = 0.1$, which led the best results at bag level in the validation cohort.



(a) (b)

Figure 5.4: Ablation studies on positive expansion (PE) MIL formulation. Hyperparameters study for λ_{PE} and t_{PE} are performed for bag-level mAUC on validation set (a), and instance-level accuracy (b).

Then, using the best configuration reached for the PE term, we optimized the proportion constraint configuration (PC) in Eq. 5.8. During empirical experimentation, we appreciated that the instance-level model performance on the test subset did not always correlate with the bag-level performance on the validation or test cohort when applying early stopping based on mAUC metric. As the proposed PC loss term provides information about the correct prediction of proportions, we evaluated this term as an early stopping criterion. Thus, we also kept track of the epoch average of $z = \frac{1}{B}\sum_b p_{b,k'_i} - p_{b,k''_i}$. Among the full range of hyperparameter values, the ones that showed best stability during training were $\lambda_{PC} = \{0.1, 1\}$ and $t_{PC} = \{1, 5, 10\}$. We show the results obtained at bag-level and the instance-level accuracy on test cohort, as well as the proportion constraint satisfaction on the train subset for both early stopping criterion in Figure 5.5.

The figures of merit indicate that the criterion based on constraint satisfaction (dashed lines) consistently outperforms the validation mAUC criteria (solid line) at both instance and bag level for all settings. This could be explained by the possibles bias introduced using the validation subset due to class imbalance. Likewise, maximizing the difference in proportion between the majority and minority classes can help to better

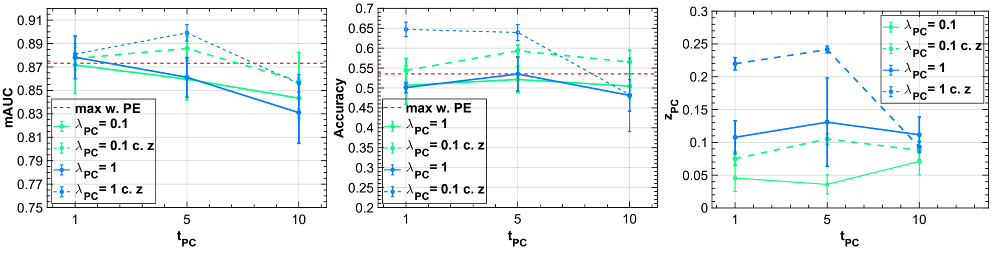


Figure 5.5: Ablation studies on proportion constraint (PC) MIL formulation. Hyperparameters study for λ_{PC} and t_{PC} are performed for bag-level mAUC on test set (a) and instance-level accuracy on test set (b). Also, two early stopping criterion are validated: mAUC on valdiation set (solid lines) and proportion constraint satisfaction z_{PC} (dashed lines), which values are illustrated in (c).

distinguish between them. The results obtained are in line with these observations, since lower values of t_{PC} seem to obtain better results. Due to the formulation of the barrier extension (Eq. 5.4), low values of t contribute not only to fulfill the constraint, but also to maximize it by using a slope proportional to $1/t$. Therefore, we selected the setting that gives the largest proportion of difference between the primary and secondary grade on the train cohort: $t_{PC} = 5$ and $\lambda_{PC} = 1$.

5.5.3 Qualitative evaluation

Finally, we want to get a more intuitive view of how the different terms of the proposed methodology are influencing the extraction of discriminative features. For that purpose, we depict the feature representation of the embedding space produced by the encoder networks on the instance-level labelled test cohort using the t-sne [145] in Figure 5.6. Concretely, we obtained the two-dimensional t-sne embedding using a perplexity value of 40, and 300 iterations. The t-sne representation is obtained on the instance-max setting (5.6a), instance-max with PE term (5.6b) and instance-max with PE and PC terms (5.6c) after Student model training.

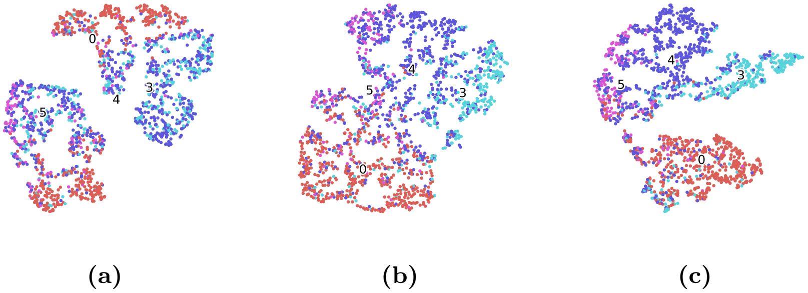


Figure 5.6: Visualization of the embedding space produced by baselines and the proposed method models on the labelled instances from SICAP-MIL test cohort. (a) instance-max; (b) instance-max w. PE; (c) instance-max w. PE w. PC. Red: non-cancerous; light blue: Gleason grade 3; dark blue: Gleason grade 4; purple: Gleason grade 5.

Features obtained using the basic max aggregation are quite overlapped on the cancerous classes. Although the PE term slightly improves this condition, only once the PC term is included it is possible to distinguish class-wise clusters between Gleason grades 3 and 4. These grades tend to coincide in WSIs, with Gleason score 7 (whole slide images that include both tumour growth patterns of grade 3 and 4) being the most common in the database used (see Table 5.1 and Figure 5.2). This fact produces noise during training, as many bags are positive for both classes simultaneously, making it difficult to distinguish between the two types of instances. However, when we introduce the relative proportion information of both classes during training, this facilitates the network to promote a distinction between them.

Also, we introduce in Figure 5.7 visualizations of the obtained instance-level classifications, compared to pathologists annotations and baselines. Instance-level predictions are performed on the test subset biopsies using an overlap of 75% between instances, to gain spatial resolution. Then, the instance-level scores are assigned to each pixel of the patch, and they are averaged among the overlapped patches. From the selected representative examples, it is observed how once the different proportion constraints are introduced, the model is able to differentiate best between the different Gleason grades (first and second rows), and locates more cancerous regions (third row).

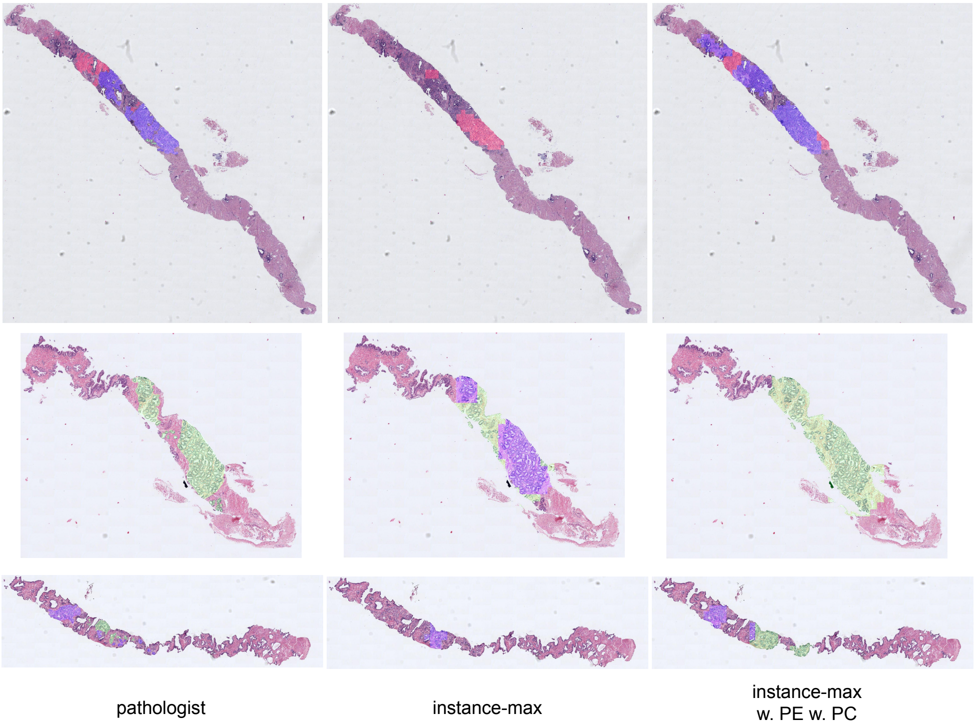


Figure 5.7: Visual examples of the proposed model performance on instance-level prostate cancer grading. In particular, the pathologists annotations are depicted with the instance-based MIL baseline using max aggregation, and the results when we introduce the proportion priors. In green: Gleason grade 3; blue: Gleason grade 4; red: Gleason grade 5.

5.6 Conclusions

In this work, we have presented a novel constrained multi-label instance-based MIL formulation that encourages the network to focus on many positive instances, and allows to impose restrictions about relative proportions of class size within the bag. In particular, we combine a standard instance-based max aggregation with additional inequality constrains terms via a flexible log-barrier extension. We validate the proposed formulation on a new publicly available dataset of prostate histology cancer WSIs images, SICAP-MIL. In the experimental stage, our method shows that forcing the network to classify more positive instances, the results improve in $\sim 5\%$ at instance level classification accuracy. By simply incorporating relative proportion information about the primary grade in the WSI, which is usually easily accessible from medical records, our method reports improvements of $\sim 9\%$ accuracy at instance level, and $\sim 3.3\%$ mAUC at bag level. In addition, the target relative proportion difference between primary and secondary classes in the bag has proven to be a good criterion when optimizing the model, obtaining more generalizable results than using the mAUC at the bag level. The obtained results are comparable to prior works using similarly-sized datasets under the supervised paradigm, which require tedious instance-level annotations.

Chapter 6

Supervised contrastive learning-guided prototypes on axle-box accelerations for railway crossing inspections

The content of this chapter corresponds to the author version of the following published paper: Silva-Rodríguez, J., Salvador, P., Naranjo, V., & Insa, R. Supervised contrastive learning-guided prototypes on axle-box accelerations for railway crossing inspections. Expert Systems with Applications, (2022).

Contents

Supervised contrastive learning-guided prototypes on axle-box accelerations for railway crossing inspections

Julio Silva-Rodríguez[1], Pablo Salvador[1], Valery Naranjo[2] and Ricardo Insa[1]

[1]Institute of Transport and Territory, Universitat Politècnica de València, Valencia, Spain; [2] Institute for Research and Innovation in Bioengineering, Universitat Politècnica de València, Valencia, Spain

Abstract

Increasing demands on railway structures have led to a need for new cost-effective maintenance strategies in recent years. Current dynamic railway track monitoring systems are usually based on the analysis of axle-box accelerations to automatically detect track singularities and defects. These methods rely on hand-crafted feature extraction and classifiers for different tasks. However, the low performance shown in previous literature makes it necessary to complement these analyses with in-situ inspections. Very recent works have proposed the use of deep learning systems that allow extracting more generalizable features from time-frequency spectrograms. However, the lack of specific public domain datasets and the finite number of track singularities in a railway structure have limited the development of deep learning based systems. In this paper, we propose a method capable of outstanding in low-data scenarios. In particular, we explore the use of supervised contrastive learning to cluster class embeddings nearly in the encoder latent space, which is used during inference for prototypical distance-based class assignment. We provide comprehensive experiments demonstrating the performance of our method in comparison to previous literature for detecting worn-out crossings.

6.1 Introduction

Railway structures are one of the main components of any country's transportation system. Railway maintenance plays a key role in achieving a high-performance, safe and cost-effective system. [146]. The increase in demand for passenger and cargo rail transport services has led to an increase in the maintenance needs of the rail network in recent years. Specifically, European countries invest between 15 and 25 billion euros annually in the maintenance and renewal of these structures [147]. With the advent of the Industry 4.0 paradigm and the development of enabling technologies such as sensing devices and artificial intelligence systems, predictive maintenance has been projected as a promising tool for cost-effective maintenance strategies.

In this work, among the different challenges on railway maintenance, we focus on track surveying. Different technologies have been proposed to support the maintenance process: vision camera-based methods, acoustic recording, laser sensors, etc. [148]. Among these procedures, the use of axle-box accelerometers have proved to be versatile enough to sense different track irregularities of different wavelengths and occurrence [149–151]. Some of its advantages are that this technology is not limited to any field of view, and it is able to perform a dynamic surveying of the direct interaction between the track and the railway. The presence of characteristic track element patterns and their deterioration in axle-box acceleration on time-frequency domain has been extensively studied in previous literature [149]. In addition, some models based on hand-crafted feature extraction based on traditional image processing methods and machine learning models have been proposed and used on maintenance practice [152]. However, the low performance of these methods makes it necessary to supplement these predictions with on-site visual inspections by operators.

The emergence of deep learning has led to an increase of performance of different computer-vision based industrial applications. In particular, very recent works have shown the benefits of using convolutional neural networks (CNNs) for axle-box track surveying characterization [153–155]. Under the supervised learning paradigm, deep learning models have achieved remarkable performance in a wide range of applications. Nevertheless, a main limitation of these models is the large amount of labeled data required for training. These limitations are accentuated in track surveying applications. The absence of domain-specific datasets makes it difficult use pre-trained fine-tuned models and the annotation process is costly, while the number of track elements is limited [156]. This encourages the development of novel strategies, capable of withstanding low data scenarios, to achieve robust and reliable automatic systems that may be used in decision making systems for dynamic track surveying.

Based on these observations, in this paper we propose a novel end-to-end system able to detect worn crossings using axle-box accelerations and deep-learning based features via convolutional neural networks (see Figure 6.1). The key contributions of our work can be summarized as follows:

- We propose to deal with the scarcity of labelled training data inherent to track surveying applications by means of non-parametric prototypical inference over the feature encoding.
- Specifically, unlike previous work, class embeddings are distributed in the latent space indirectly, using a subspace guided by supervised contrastive losses.
- We compare the proposed system with previous methods in the literature. In-depth experiments demonstrate the superior performance of our approach, with accuracy gains of $\sim 8\%$.
- In addition, we report extensive ablation experiments to provide further insights into feature preprocessing, CNN architectures, and learning strategies in a deep learning-based analysis of axle-box accelerations.

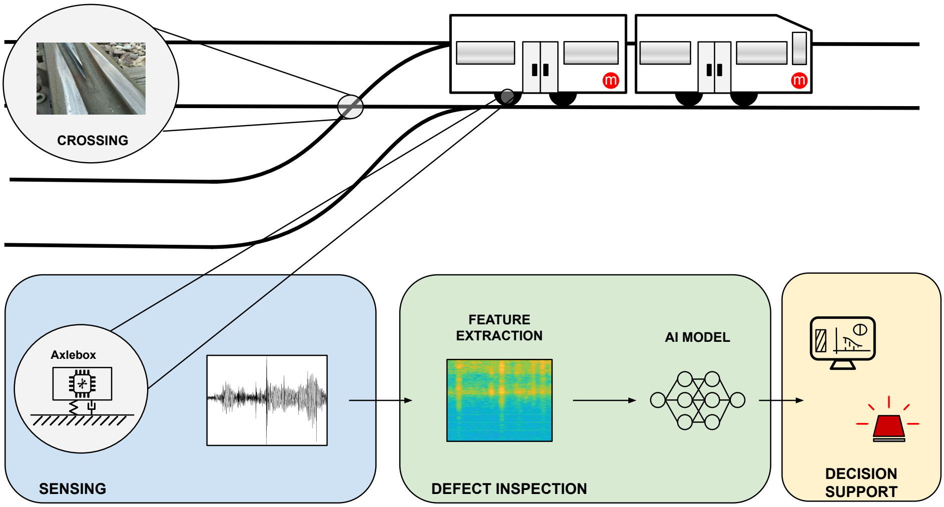


Figure 6.1: System overview. In this work, we propose a deep-learning based system able to locate worn crossing on railway surveying maintenance. The sensing technology is based on axle-box vertical accelarations (Section 6.3.1). First, signals are transformed to time-frequency distributions (Section 6.3.2). Then, normalized features are used as input to an artificial intelligence model (Section 6.3.3) to detect worn crossings. The proposed model can be trained on scenarios with scarce training examples. This pipeline can be scaled to other analysis on dynamic railway surveying.

6.2 Related Work

6.2.1 Railway track surveying

Automatic track surveying is based on patter analysis over sensed signals and images. Among sensing devices, different technologies such as thermal resistors [157], acoustic sensors [158, 159], video recording [160–165]; [166, 167]) or accelerators [149, 153–155, 168–187] have been proposed. In particular, the use of acceleration sensors on axle-box has become more popular for detecting track irregularities of different wavelengths and occurrence. Concretely, different applications include wheel flat [171], crossings monitoring [174, 175], rail corrugation [176, 185]; [188]), roughness derivation [178, 179], rail joints [155], settlement and dipped joint ([189]) and other railway elements [149, 190]. In the aim of predictive maintenance, first works focused on visual description of the patterns that elements and defects produce on time-frequency domain [149, 168, 169, 177–179, 182–184, 190]. Among time-frequency distributions, both standard short-time Fourier transform and Wavelets have been used alike. Further on, some works described a set of features based on classic image processing such as peak intensity, frequency-band relative intensity, or other statistics. Then, first classifiers were used on these features, such as SVMs [176] to predict rail corrugation, random forest for railway lifetime prediction [174, 175] or simple costume decision trees for fault detection [185], or recent neural networks classifiers [189, 191]. Very recent works [153–155, 187] have proposed the use of deep learning models via CNNs to characterize acceleration spectrograms on predictive tasks. In line to recent advance on computer vision, these works have perform superior than previous approaches based on handcrafted feature extraction [153, 155, 187]. Although these works have shown promising results, models are usually trained on small datasets, with scarce labelled data [156]. On vision camera-based methods, the vast amount of publicly available databases of natural images facilitates the use of previous knowledge for fine-tuning rich, pre-trained models [162]. Thus, camera-based surveying methods in the literature have been able to successfully train CNNs architecture such as UNets for track segmentation and fault classification [166] or YOLO networks for surface defect localization [167]. Nevertheless, time-frequency distribution of acceleration spectrograms are a too specific domain to apply such knowledge. To deal with this issue, different strategies have been proposed. For instance, some works use synthetic data to train CNNs directly on acceleration signals [189, 191]. Still, the reliability of synthetic data is not clear in comparison with in situ data. Other works have resort to self-training strategies such us autoencoders ([154]), which use unlabelled data to learn rich features. Regarding the CNNs training, the main strategy ([153, 155, 189]) is still the use of standard cross-entropy based supervised training of deep networks, which tend to generalize poorly when trained from scratch on small datasets.

6.2.2 Learning from limited data

In the context of deep learning, the branch that covers low-data training is few-shot learning. In this scenario the goal is to train a model capable of making predictions that can be generalized to new classes, of which few examples (K-shots) are given during inference. This model, instead of simply characterizing given classes on a standard supervised scenario, should be able to project a feature space from images, where samples from new, unknown concepts, behave similar. Although this setup has gained popularity on recent years, it is sometimes difficult to apply it in real applications, which need to prove its performance when all classes are used during both training and inference. Nevertheless, methods proposed on the few-shot learning paradigm tend also to generalize best on standard supervised scenarios train on very small data, as it is our case. Among different approaches in few-shot learning classification, metric-based methods aim to learn a good embedding space, where novel class samples can be nicely categorized. This categorization has been done learning a deep distance metric on matching [37] or relational networks [50], but also using memory-based nearest neighbour classifier (so-called prototypical networks) based on class-level prototypes via l2 (Euclidean) [38] or cosine distance [51, 192]. These methods are trained on an episodic way, where training examples are divided between queries and support to simulate the few labeled examples encountered during inference. Nevertheless, recent works have demonstrated that such training strategy is data-inefficient, and produces detriments in model performance [52]. Methods that learn to cluster samples in a non-episodic way resemble contrast-based learning methods, which have recently demonstrated leading results on classification tasks in self-training [33], and in standard supervised learning [53]. In the last case, clusters of points belonging to the same class are pulled together in a hyper-sphere subspace, while simultaneously pushing apart clusters of samples from different classes, in a mini-batch way. In this work, we investigate the use of contrastive learning on low-data scenarios for learning embeddings subsequently used via a prototypical-based inference.

6.3 Methods

6.3.1 Data acquisition

In this work, we study the dynamic train-track interaction as a system of masses, springs and dampers. In this model, any significant alteration in any of the elements will affect the rest of the system. Thus, it is possible to survey alterations on railway track status by recording the interaction on later elements of the system. The dynamic surveying of the railway status is performed by means of vertical accelerometers placed on the axle-boxes of the wheelsets for the left and right rails. From this interaction, we intend to train a classifier capable of recognizing whether a crossing is worn or not. Hereafter, we will refer to $x[n]$ as the signal acquired for any of the channels in a given window, which contains a crossing.

6.3.2 Feature extraction

The recorded signals $x[n]$ on time domain are transformed into the time-frequency spectrograms using the short-time Fourier transform, $X[m, \omega]$ such that:

$$X[m, \omega] = \sum_{n=-\infty}^{N-1} x[n]w[n-m]e^{-j\omega n} \tag{6.1}$$

where $w[n]$ is a hamming window, with length W samples. Each window, $w[n]$, get chunks of the original signal, overlapped by O to reduce artifacts. Note that, in the following, we refer to $X[m, \omega]$ as X for simplicity.

Then, spectrograms are scaled to improve model convergence and fasten training. Concretely, we propose to use a dynamic-margin normalization of the input spectrogram to ensure that $X \in [0, 1]$, and use all the intensity range. This operation is parameterized by the desired dynamic margin in decibels, γ, such that:

$$X' = \frac{20\ log_{10}(\frac{X}{W/2} + \epsilon) + \gamma}{\gamma} \tag{6.2}$$

where $\epsilon = 10^{(\frac{-\gamma}{20})}$. In the following, we refer to X' as X for notational simplicity

Feature extraction is applied to axle-box signals from both railways, and their features are concatenated into a two-channel tensor for both model training and inference.

6.3.3 Supervised contrastive feature learning

An overview of our algorithm for crossing wear detection is presented in Figure 6.2. Below, we describe each component proposed for model training and inference.

Let us denote a set of I crossing features $\{X_i\}_{i=1}^{I}$, and their respective labels by $\{y_{i,k}\}_{i=1}^{I}$. Each individual label, $y_{i,k}$, is composed by a one-hot-encoding ground-truth that indicates if that crossing is worn, such that $y_{i,k} \in \{0,1\}$, with $k = \{0,1\}$. We also define an encoder, $f_{\boldsymbol{\theta}}(\cdot) : \mathcal{X} \rightarrow \mathcal{Z}$, parameterized by $\boldsymbol{\theta}$, that is trained to characterize each crossing into an embedding of lower dimensionality D_E, such that $\mathbf{z} \in \mathcal{Z} \subset \mathbb{R}^{D_E}$. Then, we aim to train $f_{\boldsymbol{\theta}}(\cdot)$ such that the embedding representation of normal and anomalous crossings are discriminated. In this line, we propose to use a supervised contrastive strategy. Thus, we define a projection head, $f_{\boldsymbol{\phi}}(\cdot) : \mathcal{Z} \rightarrow \mathcal{R}$, parameterized by $\boldsymbol{\phi}$, which is composed by a two-layered perceptron with relu activations that maps the embedding space to a lower dimensionality, such that $\mathbf{z} \in \mathcal{R} \subset \mathbb{R}^{D_E/F_c}$, with F_c a system hyper-parameter. Then, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are trained via gradient descent to minimize the supervised contrastive loss [53] defined as:

$$\mathcal{L}_c = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} log \frac{\exp(\mathbf{r}_i \cdot \mathbf{r}_p/\tau)}{\sum\limits_{a \in A(i)} \exp(\mathbf{r}_i \cdot \mathbf{r}_a/\tau)} \tag{6.3}$$

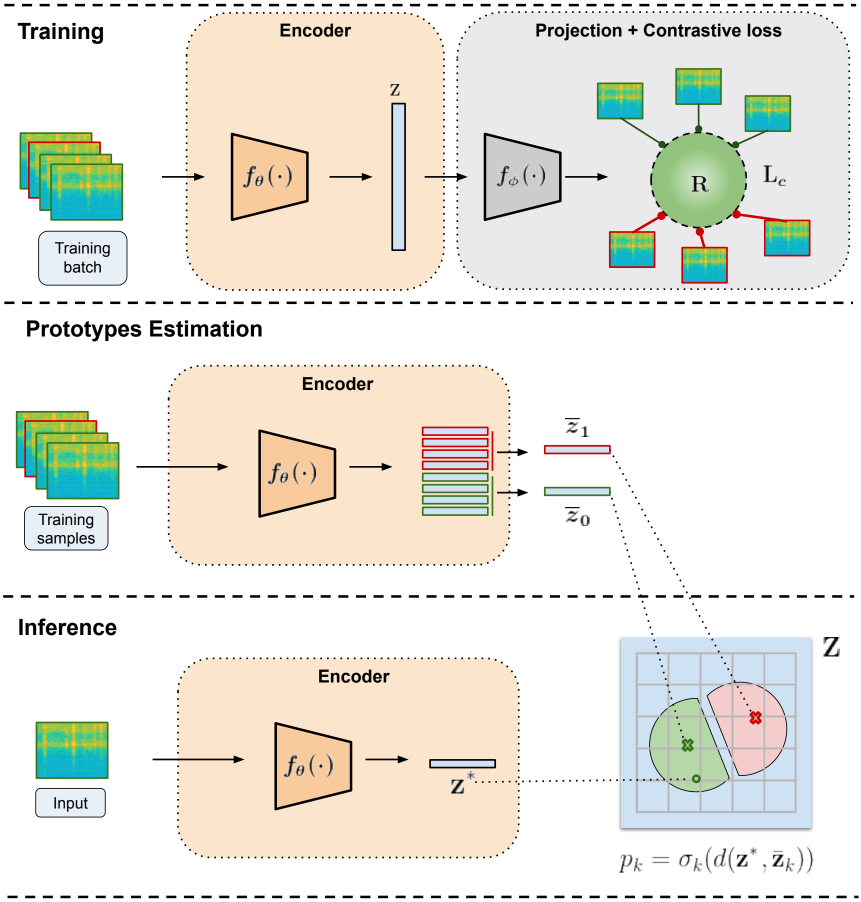


Figure 6.2: Method overview. An encoder is trained to minimize supervised contrastive loss in Eq. 6.3 after projecting the produced embedding $\mathbf{z}$ into a subspace $\mathbf{r}$ that falls into an unit hyper-sphere. During inference, new queries are classified on the latent space projected by the encoder. Concretely, a non-parametric prototypical classifier is implemented using class-wise prototypes $\bar{\mathbf{z}}_k$ from the training set given by Eq. 6.4. In particular, the class of nearest prototype in terms of l2-distance is assigned to the new query sample.

where $\cdot$ denotes the inner product, $\tau \in \mathbb{R}^+$ is a temperature parameter, $A(i) \equiv I \setminus \{i\}$ indicates all instances other than i, and $P(i) \equiv p \in A(i) : y_p = y_i$ refers to the set of instances positives, with $|P(i)|$ its cardinality.

It is noteworthy to mention that $\mathbf{r}$ are l2-normalized features, to apply the criterion on an unity hyper-sphere. Using supervised contrastive loss, points belonging to the same class (positives) are pulled together in the projected space, while simultaneously pushing apart clusters of samples from different classes (negatives).

6.3.4 Prototypical inference

For inference, contrastive-based methods usually train a linear classifier on top of the frozen representations $\mathbf{z}$ using a cross-entropy loss. In this work, we study the use of non-parametric inference strategy, to avoid overfitting on scenarios with limited data available. Concretely, we use prototypical-based inference [38], a memory-based approach that assigns predicted labels according to the distance in the latent space between new queries and precomputed representations of each class, called prototypes. This method creates softer decision boundaries compared to learned-based architectures. As we support later on our experiments, it generalizes better in the setting under study. Prototypes are calculated using all samples from training set such that:

$$\bar{\mathbf{z}}_k = \frac{1}{I} \sum_i \mathbf{z}_i \tag{6.4}$$

Given a new query sample, X^*, the wear prediction $\hat{y}_k$ is given by its relative distance to each prototype as follows:

$$\hat{y}_k = \sigma_k(d(f_{\boldsymbol{\theta}}(X^*), \bar{\mathbf{z}}_k)) \tag{6.5}$$

where σ_k indicates a softmax activation over classes, and $d(\cdot)$ indicates the Euclidean distance.

6.4 Experimental setting

6.4.1 Dataset

The experiments described in this work were carried out using a private dataset of dynamic railway surveying on line 3 of Metrovalencia. 25 km of railway surveying were recorded using the data acquisition setup described in Section 6.3.1, with accelerometers of model KS76C100 manufactured by MMF and sampling frequencies of 3.2 KHz. The train used in the tests was an Electrical Multiple Unit (EMU 4300 series), which has four cars of two bogies each one, being motorised the wheelsets of the last car. The run tests had a maximum speed of 80 km/h, and included ballasted track with single-block concrete sleepers, and Stedef slab track. From the entire path, 33 crossing points were selected and manually on-site evaluated by experienced operators in terms of wear. Of this dataset, 17 crossings points showed damages that required follow-up and maintenance actions. Observed deterioration included spalls, burrs and squats. Examples of the deteriorated crossings are presented in Figure 6.3. The acceleration signals recorded were windowed using 4 seconds around each crossing point.

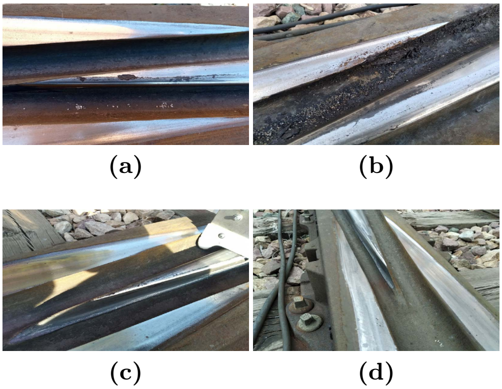

(a) **(b)** **(c)** **(d)**

Figure 6.3: Examples of deteriorated railway crossings included in the used dataset. Anomalies indlude squats (a-b-d), spalls (c), and burrs (d).

6.4.2 Implementation details

The 4 seconds crossing signals acquired as detailed in Section 6.3.1 are transformed to time-frequency spectrograms as detailed in Section 6.3.2. Concretely, based upon the studies in [149], hamming windows of $W = 0.25$ seconds with an overlap of $O = 95\%$ were used to compute the short-time Fourier transforms. Then, spectrograms were normalized using the dynamic-margin standardization with $\gamma = 20$, and resized to 256×320 pixels to reduce computational requirements. Using a 4-fold cross validation strategy, the encoder for crossing characterization was trained as described in Section 6.3.3. Concretely, ResNet-18 [91] was used as base architecture for the encoder. The architecture used included an initial convolutional layer to adapt the number of channels, and was composed of 2 residual blocks. The spatial features were reduced to a one-dimensional embedding $z \in \mathbb{R}^{64}$ via a global-average pooling. Regarding the projection head, a multi-layered perceptron that reduced the embedding size in an order of $F_c = 4$ with relu activation was used. The different modules were trained during 200 iterations, using ADAM optimizer with a learning rate of $1e{-}4$ and mini-batches of 8 samples. Finally, test samples form each fold are infered as described in Section 6.3.4, using all samples from training subset to compute class-wise prototypes. The code and trained models are publicly available on (`https://github.com/cvblab/contrastive_prototypes_railway`).

6.4.3 Baselines

In order to compare our approach to state-of-the-art methods, we implemented proposals of prior works on accelerometer-based automatic railway maintenance, and validated them on the dataset used, under the same conditions. Due to the scarce literature on this field, we only differentiated three proposed approaches: hand-crafted feature-based methods, standard supervised learning using CNNs and cross entropy loss, and self-training ones via autoencoder features. ***Hand-crafted features methods***, aim to describe a series of features obtained by classic signal processing methods on time and frequency domains using human knowledge about the problem. Concretely, from the windowed crossing signal, we used as features the intensity peak amplitude, relative intensity at different bandwiths, entropy and other statistics such as skewness, similarly to [176]. Then, a support-vector machine (SVM) classifier with Gaussian kernell was trained to predict the wear crossing. ***Self-training methods***, aim to leverage knowledge on large amounts of unlabelled data from dynamic surveyings. Concretely, an autoencoder is trained to compress the spectrogram information into an embedding space, which is trained to minimize the reconstruction error using a trained decoder. Then, the resultant embedding space is used for clustering purposes. In our work, we implemented an autoencoder trained on the full dataset (including unlabelled data). Concretely, the same architecture with residual blocks used for our proposed method was used as encoder, and a symmetrical decoder was used to reconstruct the input spectrogram. The autoencoder architecture was pre-trained during 100 iterations using ADAM optimizer with a learning rate of $1e-4$ and mini-batches of 32 samples. Then, the non-parametric prototypical inference described in Section 6.3.4 was used for classification using the features extracted from the encoder. ***CNNs using cross-entropy loss***: Also, we include as an independent baseline the same CNN architecture trained using simply the binary cross entropy loss instead of the proposed learning method, as it has been used by [153, 155, 189].

6.4.4 Evaluation metrics

We use standard metrics on classification tasks to evaluate the proposed system performance on crossing wear detection. In particular, accuracy, precision and recall are calculated using the expert and system labels. From precision and recall F1-score (FS) is calculated to summarize both figures of merit. For each experiment, the metrics shown are the mean of ten consecutive repetitions of the model training, to account for the variability of the stochastic factors involved in the process.

6.5 Results

6.5.1 Crossing wear detection

The quantitative results obtained by the proposed model and baselines on the cross-validation partitions are presented in Table 6.1. We can observe that the proposed methodology outperforms previous approaches by a large margin, with a substantial increase of $\sim 8\%$ in both accuracy and F1-score. Although the hand-crafted features baseline reached promising results (0.6124 accuracy), deep-learning methods outperformed this approach, which aligns to recent literature on railway surveying [155]. Finally, the features learned by the autoencoder approach, even though it is trained on large quantities of data, obtained results inferior to those of the proposed method. This may be because the cross wear classification task requires specific features. In contrast, the autoencoder learns general features to reconstruct the original image that do not seem suitable for the supervised task.

Method	**Metric** ($\mu \pm \sigma$)			
	Accuracy	**F1-Score**	**Precision**	**Recall**
CNNs + BCE ([155, 189])	0.5875 ± 0.0945)	0.6099 ± 0.1227	0.6111 ± 0.0863	0.6529 ± 0.2254
Hand-crafted Features + SVMs [152, 185]	0.612 ± 0.061	0.652 ± 0.014	0.604 ± 0.022	0.651 ± 0.054
Autoencoder Features [154]	0.648 ± 0.025	0.652 ± 0.014	0.652 ± 0.014	0.720 ± 0.064
Proposed	$\mathbf{0.715 \pm 0.715}$	$\mathbf{0.735 \pm 0.050}$	$\mathbf{0.726 \pm 0.0496}$	$\mathbf{0.747 \pm 0.064}$

Table 6.1: Quantitative results on railway crossing wear detection for the proposed method and implemented baselines. Best results in bold.

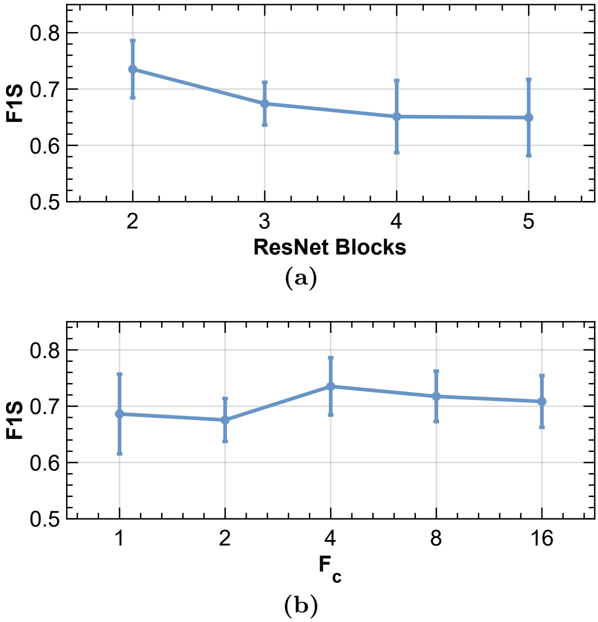
6.5.2 Ablation studies

In the following, we provide comprehensive ablation experiments to validate several elements of our model, and motivate the choice of the values employed in our formulation, as well as our experimental setting.

Studies on model complexity We first studied the configuration of the encoder used, ResNet-18, for the feature extraction stage. Concretely, we validated the proposed model using different number of residual blocks. Results are presented in Figure 6.4a, from which we can observe how the less residual blocks are used, the best the classification performance is. These results could be explained in two different ways: first, deep networks are over parameterized under scarce data conditions, and second, visual characterization on acceleration spectrograms are made up of by simple patterns, which are modeled on early layers of CNNs, together with intensity information.

Contrastive learning setup Next, we study the multi-layered perceptron block used on the contrastive head. Concretely, ablation experiments are performed on the dimensionality of the unity hyper-sphere used to contrast samples, as a fraction of the dimension of the features extracted by the encoder. Concretely, the compression factor F_c is evaluated at $F_c = \{1, 2, 4, 8, 16\}$. Results are illustrated in Figure 6.4b. These show that reducing the dimension on the hyper-sphere used for contrastive losses produces slight benefits, with improvements around 3% on F1-score.

Feature normalization As previously mentioned, one of the main steps on deep learning systems is feature normalization. Concretely, the time-frequency spectrogram intensity should be constrained to small amplitudes, such that $x \in [0, 1]$. For this purpose, our method uses a dynamic-margin normalization described in Section 6.3.2. We now validate the proposed normalization, comparing both quantitatively and qualitatively with other well-known methods. In particular, we use minimum-maximum normalization, and z-score standardization on log-magnitude spectrograms. Results are presented in Table 6.2, while normalized spectrograms are presented in Figure 6.5. Results demonstrate that benefits of dynamic-margin normalization, which outperforms other approaches by up to $\sim$ 8% in terms of F1-score. Qualitative evaluations show that the most large-intensity excited frequencies are contrasted from background on the spectrogram, the best the results are.



(a)

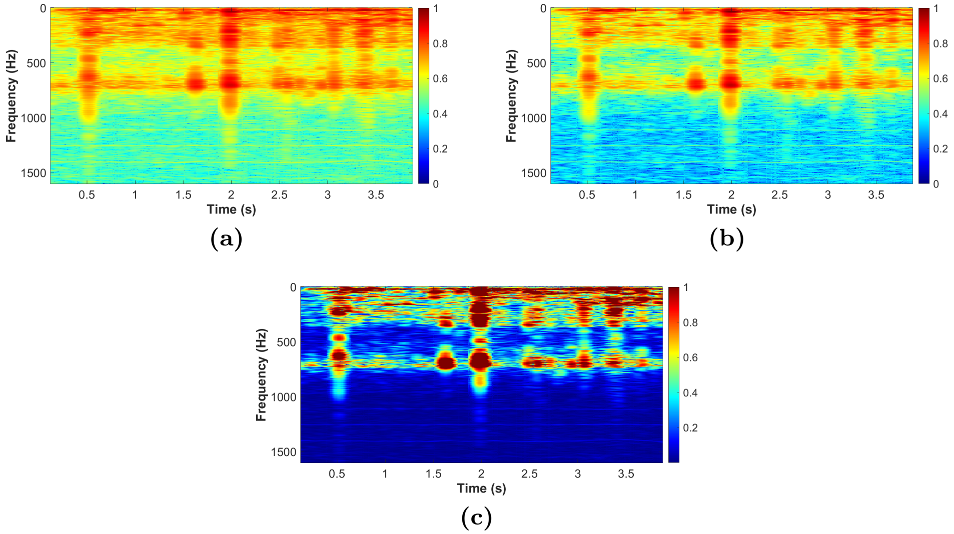
(b)

Figure 6.4: Ablation studies on network architecture. Accuracy and F1-score are presented for each possible configuration. Best performance highlighted in bold. (a) Encoder complexity; (b) Contrastive head compression factor.

Normalization	**Metric** $(\mu \pm \sigma)$	
	Accuracy	**F1- score**
z-score	0.6124(0.0619)	0.6045(0.0223)
min-max	0.6484(0.0259)	0.6529(0.0147)
dynamic-margin	**0.7156(0.0715)**	**0.7352(0.0508)**

Table 6.2: Ablation study on feature normalization methods. Best results in bold.

Learning strategies In the following, we benchmark the proposed contrastive-based feature learning and prototypical inference with other common methods. Concretely, we train the proposed model using a linear classification layer and binary cross-entropy (BCE) loss to compare both contrastive and BCE-based training. For fair comparisons and to avoid the over-parametrization of densely classification, we also implement the prototypical inference on the BCE-trained model (BCE+Prototypes) as described in [193]. Finally, we also include a purely prototypical



(a)

(b)

(c)

Figure 6.5: Qualitative assessment of different normalization strategies. (a) min-max; (b) z-score; (c) dynamic margin.

learning strategy (Prototypical), using episodic training and minimizing l2-distance between support and query samples as proposed in the original publication [38]. Concretely, the number of query and support samples used during training was 4. The encoder architecture and hyper-parameters were the same to the ones optimized for our proposed method (see Section 6.4.2). Results for different methods are presented in Figure 6.6 in terms of accuracy and F1-score. The proposed supervised contrastive learning model and prototypical inference outperforms by a large margin the BCE method, and shows greater stability in the results among experiment repetitions. Although results consistently improve using prototypical memory-based inference, our method reaches the best performance, which shows the benefits of contrastive learning strategies.

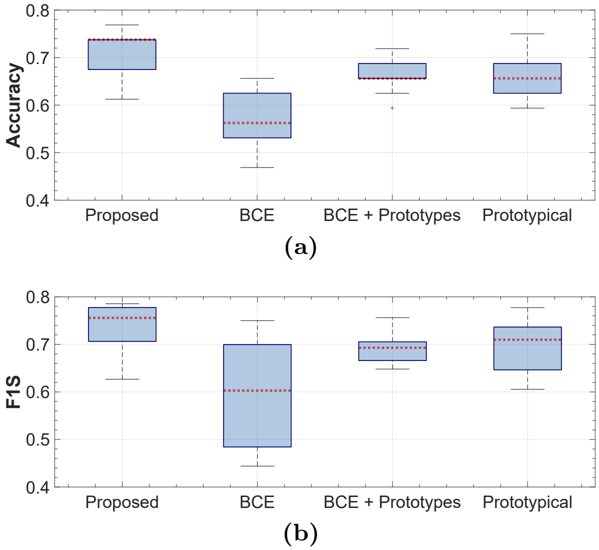


Figure 6.6: Ablation studies on learning strategies. The illustrated metrics are accuracy (a) and F1-score (b).

On the role of each element of the system Different components have been presented to optimize the proposed method: dynamic margin normalization, prototypical inference, and constrastive feature learning have been the best performing settings. Nevertheless, it is still unclear the individual contribution of each element. For this reason, in the following, we discuss the incremental improvement of each module of the system. First, we focus on normalization methods, where dynamic margin normalization performed the best on the proposed setting (see Table 6.2). In addition, as shown in Figure 6.7, this type of normalization is also indispensable to obtain promising results when we simply use a CNN with linear classifier, trained using cross-entropy (BCE). Thus, we consider this standardization to be an indispensable step for the operation of the system. Next, if we introduce an inference based on prototypes (BCE+Prototypes), improvements of $\sim$ 8% are obtained (see Figure 6.6). Finally, when we get rid of entropy-based objective functions, using the proposed contrastive learning setting, improvements of $\sim$ 5% are obtained for both accuracy and F1-score figures of merit (see Figure

6.6). Thus, we see that what most damages the model is the use of dense classifiers during inference, in the scenario studied with sparse data. Next, direct training of the model to generate prototypes based on contrastive learning also produces a substantial improvement.

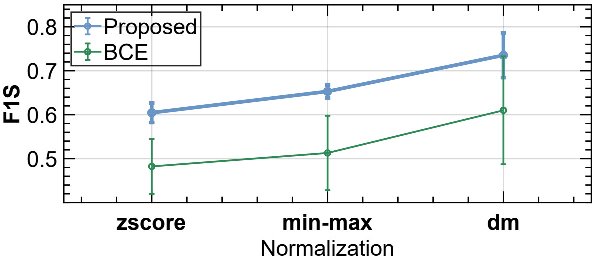


Figure 6.7: Ablation study on feature normalization methods. In particular, performance using zscore, min-max, and the proposed dynamic margin (dm) normalization is compared for the proposed method and a CNN using linear classifier (BCE).

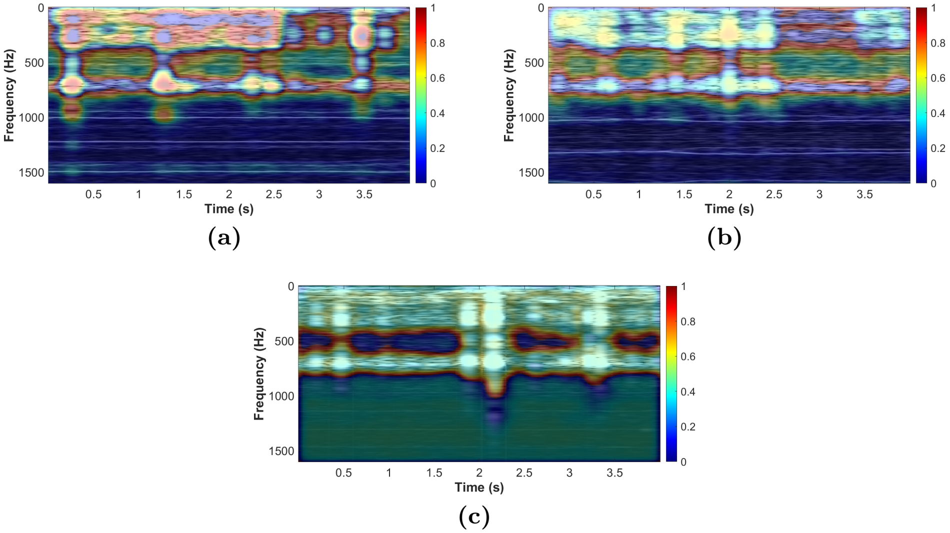
6.5.3 On system explainability

Explainability on AI-based systems have become a relevant topic on the field that aims to prevent bias on learning systems and demonstrate the robustness of the model [194]. In the following, we explore the explainability of the proposed model in order to provide confidence in its use during railway maintenance practice. Thus, we shed light into the features learned by the trained CNN to detect wear crossings using gradiend-guided class activation maps (CAMs) [40]. For a given input image $\mathbf{x}$ its corresponding attention map is computed as: $a = \Sigma(\sum_k^K \alpha_k f_{\boldsymbol{\theta}}^s(\mathbf{x})_k)$ where K is the total number of filters of that encoder layer, Σ a sigmoid operation, and α_k are the generated gradients such that:

$$\alpha_k = \frac{1}{|\mathbf{a}|} \sum_{t \in \Omega_T} \frac{\partial \hat{y}_1}{\partial \mathbf{a}_{k,t}} \tag{6.6}$$

where Ω_T is the spatial features domain.

Generated CAMs of representative cases are visualized overlaid to the input spectrogram features on Figure 6.8. These heat-maps highlight the important regions in the image for predicting a crossing as anomalous. Concretely, we can appreciate that CAMs focus on the band-with between 650 to 850 relaxation frequencies. These findings are consistent with previous literature [149], that identified wider patterns and higher relative amplitude on this band related to crossings points on spectrograms.



(a)

(b)

(c)

Figure 6.8: Qualitative evaluation of the proposed model on wear crossing detection. For explainability, class-activation maps are obtained on true positive (a-b) and true negative (c) predictions, and overlaid over the input spectrogram.

6.6 Conclusions

A deep learning system capable of detecting worn crossings in dynamic railway inspections via axle-box accelerations sensing has been presented. Specifically, the system processes time-frequency spectrograms using convolutional neural networks through a novel combination of prototypical inference guided by supervised contrastive learning. The use of narrow CNNs showed the best results, as they extract mostly basic patterns, similar to those found in time-frequency spectrograms. Furthermore, normalization of these distributions using a dynamic margin scaling approach outperforms standard normalization in computer vision tasks. This method improves the contrast between the excited frequencies and the background, leading to better characterization. In addition, the supervised contrastive learning strategy has shown a promising performance for learning on small data sets. It outperforms standard cross-entropy based supervised learning by a wide margin, and improves other metric learning strategies from the few-shot learning domain, which resort to episodes-based training. The presented method achieves F1-score values of 0.7352 in a cross-validation, and outperforms previous literature by $\sim 8\%$ for defect crossing classification. The presented system and its methods could be used to detect a wide range of singularities and defects in railway surveying.

Chapter 7

Constrained unsupervised anomaly segmentation

The content of this chapter corresponds to the author version of the following published paper: Silva-Rodríguez, J., Naranjo, V., & Dolz, J. Constrained unsupervised anomaly segmentation. Medical Image Analysis, (2022). This article is an extension of the following paper presented at the international conference: Silva-Rodríguez, J., Naranjo, V., & Dolz, J. Looking at the whole picture: constrained unsupervised anomaly segmentation in The 32nd British Machine Vision Conference (BMVC) (2021).

Contents

Constrained unsupervised anomaly segmentation

Julio Silva-Rodríguez[1], Valery Naranjo[2] and Jose Dolz[3]

[1]Institute of Transport and Territory, Universitat Politècnica de València, Valencia, Spain; [2] Institute for Research and Innovation in Bioengineering, Universitat Politècnica de València, Valencia, Spain; [3]École de Technologie Supérieure, Montreal, Canada

Abstract

Current unsupervised anomaly localization approaches rely on generative models to learn the distribution of normal images, which is later used to identify potential anomalous regions derived from errors on the reconstructed images. To address the limitations of residual-based anomaly localization, very recent literature has focused on attention maps, by integrating supervision on them in the form of homogenization constraints. In this work, we propose a novel formulation that addresses the problem in a more principled manner, leveraging well-known knowledge in constrained optimization. In particular, the equality constraint on the attention maps in prior work is replaced by an inequality constraint, which allows more flexibility. In addition, to address the limitations of penalty-based functions we employ an extension of the popular log-barrier methods to handle the constraint. Last, we propose an alternative regularization term that maximizes the Shannon entropy of the attention maps, reducing the amount of hyperparameters of the proposed model. Comprehensive experiments on two publicly available datasets on brain lesion segmentation demonstrate that the proposed approach substantially outperforms relevant literature, establishing new state-of-the-art results for unsupervised lesion segmentation, and without the need to access anomalous images.

7.1 Introduction

Deep learning models are driving progress in a wide range of visual recognition tasks, particularly when they are trained with large amounts of annotated samples. This learning paradigm, however, carries two important limitations. First, obtaining such curated labeled datasets is a cumbersome process prone to annotator subjectivity, limiting the access to sufficient training data in practice. This problem is further magnified in the context of medical image segmentation, where labeling involves assigning a category to each image pixel or voxel. In addition, even if annotated images are available, there exist some applications, such as brain lesion detection, where large intra-class variations are not captured during training, failing to cover the broad range of abnormalities that might be present in a scan. This results in trained models which are potentially tailored to discover lesions similar to those seen during training. Thus, considering the scarcity and the diversity of target objects in these scenarios, lesion segmentation is typically modeled as an anomaly localization task, which is trained in an unsupervised manner. In this setting, the training dataset contains only *normal* images and *abnormal* images are not ideally accessible during training.

A popular strategy to tackle unsupervised anomaly segmentation is to model the distribution of normal images in the training set. To this end, generative models, such as generative adversarial networks (GANs) ([59, 195–199]) and variational auto-encoders (VAEs) ([58, 200–203]) have been widely employed. In particular, these models are trained to reconstruct their input images, which are drawn from a normal, i.e., *healthy*, distribution. At inference, input images are compared to their reconstructed normal counterparts, which are recovered from the learned distribution. Then, the anomalous regions are identified from the reconstruction error.

As an alternative to these methods, a few recent works have integrated class-activation maps (CAMs) during training [60, 61]. In particular, [60] leverage the generated attention maps as an additional supervision cue, enforcing the network to provide attentive regions covering the whole context in normal images. This term was formulated as an equality constraint with the form of a L_1 penalty over each individual pixel. Nevertheless, we found that explicitly forcing the network to produce maximum attention values across each pixel does not achieve satisfactory results in the context of brain lesion segmentation. In addition, recent literature in constrained optimization for deep neural networks suggests that simple penalties –such as the function used in [60]– might not be the optimal solution to constraint the output of a CNN ([48]).

Based on these observations, we propose a novel formulation for unsupervised semantic segmentation of brain lesions in medical images. The key contributions of our work can be summarized as follows:

- A novel constrained formulation for unsupervised lesion segmentation, which integrates an auxiliary constrained loss to force the network to generate attention maps that cover the whole context in normal images.
- In particular, we leverage *global* inequality constraints on the generated attention maps to force them to be activated around a certain target value. This contrasts with the previous work in [60], where *local* pixel-wise equality constraints on Grad-CAMs [40] are employed. In addition, to address the limitations of penalty-based functions, we resort to an extended version of the standard log-barrier.
- Furthermore, we consider an alternative regularization term that maximizes the Shannon entropy of the attention maps, reducing the amount of hyperparameters with respect to the extended log-barrier model, while yielding at par performances.
- We benchmark the proposed model against a relevant body of literature on two public lesion segmentation benchmarks: BraTS and Physionet-ICH datasets. Comprehensive experiments demonstrate the superior performance of our model, establishing a new state-of-the-art for this task.

This journal version provides a substantial extension of the conference work presented in [204]. First, we extended the literature survey, particularly for unsupervised medical image segmentation. Then, in terms of methodology, the current version introduces several important modifications. In particular, we further investigate the role of the gradients on the attention maps derived from Grad-CAM in the task of unsupervised anomaly detection. Based on our empirical observations, we modify the formulation in [204] to constraint directly the activation maps without involving any gradient information. Furthermore, we propose an alternative learning objective for our constrained problem based on the Shannon entropy. More concretely, we replace our log-barrier formulation by a maximizing entropy term on the softmax activation of brain tissue pixels, which reduces the complexity in terms of hyperparameters with respect to the former model. Last, we add comprehensive experiments to empirically validate our method, including an additional dataset and extensive ablation studies on several design choices.

7.2 Related Work

7.2.1 Unsupervised anomaly segmentation

Unsupervised anomaly segmentation aims at identifying abnormal pixels on test images, containing, for example, lesions on medical images ([58, 198]), defects in industrial images ([55, 60, 61]) or abnormal events in videos ([56, 197]). A main body of the literature has explored unsupervised deep (generative) representation learning to learn the distribution from normal data. The underlying assumption is that a model trained on normal data will not be able to reconstruct anomalous regions, and the reconstructed difference can therefore be used as an anomaly score. Under this learning paradigm, generative adversarial networks (GAN) ([205]) and variational auto-encoders (VAE) ([57]) are typically employed. Nevertheless, even though GAN and VAE model the latent variable, the manner in which they approximate the distribution of a set of samples differs. GAN-based approaches ([59, 195–199]) approximate the distribution by optimizing a generator to map random samples from a prior distribution in the latent space into data points that a trained discriminator cannot distinguish. On the other hand, data distribution is approximated in VAE by using variational inference, where an encoder approximates the posterior distribution in the latent space and a decoder models the likelihood ([201, 206]). Recent literature on unsupervised anomaly segmentation also includes non VAE and GAN based approaches. For instance, [23] exploits the teacher-student learning paradigm, highlighting anomalies on those outputs where the student networks and teacher model predictions differ. Additionally, feature-based methods [23, 207], which identify anomalies in the feature space can be also employed.

7.2.2 Unsupervised anomaly segmentation in medical imaging

In the context of medical images, most current literature resorts to VAEs, proposing several improvements to overcome specific limitations of simple VAEs [58, 200, 202, 208]. For example, to handle the lack of consistency in the learned latent representation on prior works, [58] included a constraint that helps mapping an image containing abnormal anatomy close to its corresponding healthy image in the latent space. [208] presented a context-encoding VAE that combines reconstruction-with density-based anomaly scoring to capture the high-level structure present in the data. More recently, a probabilistic model that uses a network-based prior as the normative distribution on the latent-variable model was proposed in [202]. In particular, this model penalized large deviations between the reconstructed and original input images, reducing false positives in pixel-wise predictions. Generative models have been also employed to tackle the unsupervised lesion segmentation task [198, 209]. While SteGANomaly [198] integrated a CycleGAN-based style-transfer framework to map samples in the latent space much closer to the training distribution, [209] mask out random regions of the input data before they are fed to the GAN model. Note that a detailed survey on unsupervised anomaly localization in medical imaging can be found in [210]. However, despite the recent popularity of these methods, the results from the Medical Out-of-Distribution Analysis Challenge 2020 [211] highlight their suboptimal performance on anomaly segmentation, which might impede their usability in clinical practice, as stressed by [212].

More recently, [60] integrate attention maps derived from Grad-CAM ([40]) during the training as supervisory signals. In particular, in addition to standard learning objectives, authors introduce an auxiliary loss that tries to maximize the attention maps on normal images by including an equality constraint with the form of a L_1 penalty over each individual pixel.

7.2.3 Constrained segmentation

Imposing global constraints on the output predictions of deep CNNs has gained attention recently, particularly in weakly supervised segmentation. These constraints can be embedded into the network outputs in the form of direct loss functions, which guide the network training when fully labeled images are not accessible. For example, a popular scenario is to enforce the softmax predictions to satisfy a prior knowledge on the size of the target region. [49] employed a L_2 penalty to impose equality constraints on the size of the target regions in the context of histopathology image segmentation. In [213], authors leverage the target properties by enforcing the label distribution of predicted images to match an inferred label distribution of a given image, which is achieved with a KL-divergence term. Similarly, [141] proposed a novel loss objective in the context of partially labeled images, which integrated an auxiliary term, based on a KL-divergence, to enforce that the average output size distributions of different organs approximates their empirical distributions, obtained from fully-labeled images.

While the equality-constrained formulations proposed in these works are very interesting, they assume exact knowledge of the target size prior. In contrast, inequality constraints can relax this assumption, allowing much more flexibility. In [214], authors imposed inequality constraints on a latent distribution –which represents a “fake” ground truth– instead of the network output, to avoid the computational complexity of directly using Lagrangian-dual optimization. Then, the network parameters are optimized to minimize the KL divergence between the network softmax probabilities and the latent distribution. Nevertheless, their formulation is limited to linear constraints. More recently, inequality constraints have been tackled by augmenting the learning objective with a penalty-based function, e.g., L_2 penalty, which can be imposed within a continuous optimization framework ([48, 142, 143]), or in the discrete domain ([215]). Despite these methods have demonstrated remarkable performance in weakly supervised segmentation, they require that prior knowledge, *exact* or *approximate*, is given. This contrasts with the proposed approach, which is trained on data without anomalies, and hence the size of the target is zero.

7.3 Methods

An overview of our method is presented in Fig. 7.1. In what follows, we describe each component of our methodology.

Preliminaries Let us denote the set of unlabeled training images as $\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^{N}$, where $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^{\Omega_i}$ represents the i^{th} image and Ω_i denotes the spatial image domain. This dataset contains only normal images, e.g., healthy images in the medical context, and has therefore no segmentation mask associated with each image. We now define an encoder, $f_{\boldsymbol{\theta}}(\cdot) : \mathcal{X} \rightarrow \mathcal{Z}$, parameterized by $\boldsymbol{\theta}$, which is optimized to project normal data points in $\mathcal{D}$ into a manifold represented by a lower dimensionality d, $\mathbf{z} \in \mathcal{Z} \subset \mathbb{R}^d$. Furthermore, a decoder $f_{\boldsymbol{\phi}}(\cdot) : \mathcal{Z} \rightarrow \mathcal{X}$ parameterized by $\boldsymbol{\phi}$ aims at reconstructing an input image $\mathbf{x} \in \mathcal{X}$ from $\mathbf{z} \in \mathcal{Z}$, which results in $\hat{\mathbf{x}} = f_{\boldsymbol{\phi}}(f_{\boldsymbol{\theta}}(\mathbf{x}))$.

7.3.1 Vanilla VAE

A Variational Autoencoder (VAE) is an encoder-decoder style generative model, which is currently the dominant strategy for unsupervised anomaly location. Training a VAE consists in minimizing a two-term loss function, which is equivalent to maximize the evidence lower-bound (ELBO) ([57]):

$$\mathcal{L}_{VAE} = \mathcal{L}_R(\mathbf{x}, \hat{\mathbf{x}}) + \beta \mathcal{L}_{KL}(q_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \tag{7.1}$$

where $\mathcal{L}_R$ is the reconstruction error term between the input and its reconstructed counterpart. The right-hand term is the Kullback-Leibler

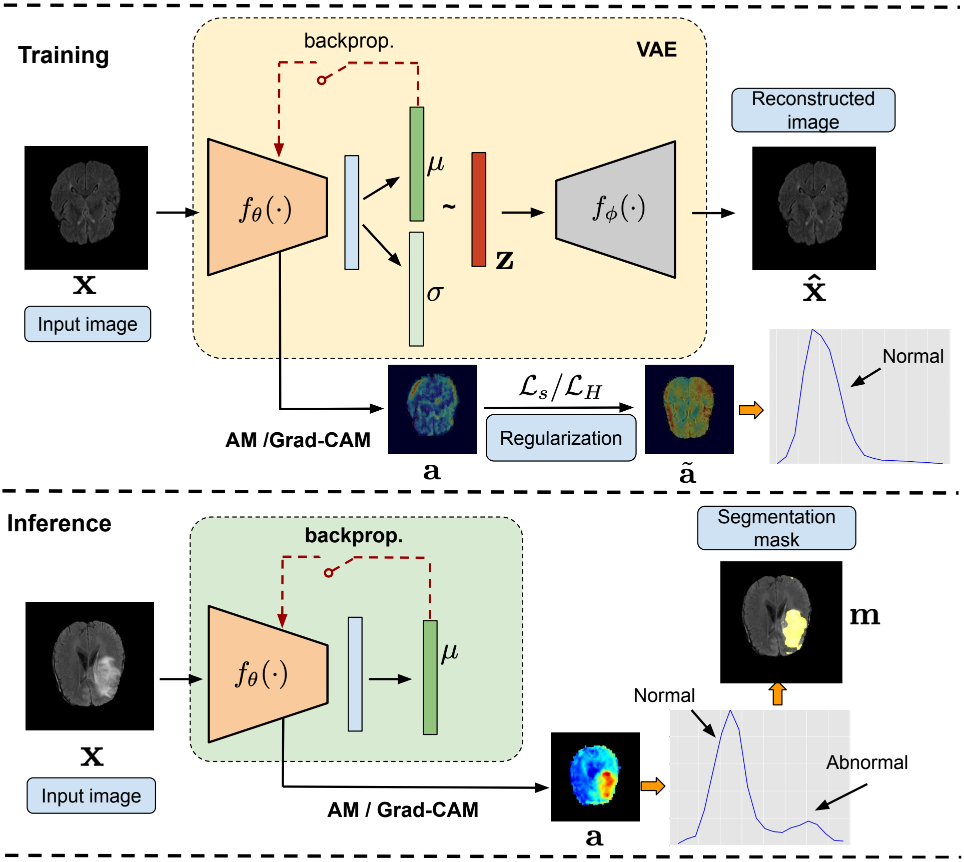


Figure 7.1: Method overview. Following the standard literature, the VAE is optimized to maximize the evidence lower bound (ELBO), which satisfies Eq. 7.1. In addition, we include an attention constraint (in the form of a size-constrained loss $\mathcal{L}_s$ or entropy proxy $\mathcal{L}_H$) on the attention maps $\mathbf{a}$, to force the network to search in the whole image. At inference, the attention map is thresholded to obtain the final segmentation mask $\mathbf{m}$.

(KL) divergence (weighted by β) between the approximate posterior $q_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ and the prior $p(\mathbf{z})$, which acts as a regularizer, penalizing approximations for $q_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ that differ from the prior.

7.3.2 Size regularizer via VAE attention

Very recent literature ([60, 61]) has explored the use of attention maps for anomaly localization. In particular, attention maps $\mathbf{a} \in \mathbb{R}^{\Omega_i}$ are generated from the latent mean vector $\mathbf{z}_\mu$, by using Grad-CAM ([40]) via backpropagation to an encoder block output $f^s_{\boldsymbol{\theta}}(\mathbf{x})$, at a given network depth s. Thus, for a given input image $\mathbf{x}^n$ its corresponding attention map is computed as follows:

$$\mathbf{a}^n = \sigma(\sum_k^K \alpha_k f^s_{\boldsymbol{\theta}}(\mathbf{x}^n)_k) \tag{7.2}$$

where K is the total number of filters of that encoder layer, σ a sigmoid operation, and α_k are the generated gradients such that: $\alpha_k = \frac{1}{|\mathbf{a}^n|}\sum_{t\in\Omega_T}\frac{\partial \mathbf{z}_\mu}{\partial \mathbf{a}^n_{k,t}}$, where Ω_T is the spatial features domain.

In [60], authors leveraged the Grad-CAMs based attention maps (Eq.7.2) by enforcing them to cover the whole normal image. To achieve this, their loss function was augmented with an additional term, referred to as expansion loss, which takes the form of: $\mathcal{L}_s = \frac{1}{|\mathbf{a}|}\sum_{l\in\Omega_i}(1-\mathbf{a}^n_l)$. We can easily observe that this term resembles to multiple equality constraints, one at each pixel, forcing the class activation maps to be maximum at the whole image in a pixel-wise manner (i.e., it penalizes each single pixel individually). Contrary to this work, we integrate supervision on attention maps by enforcing inequality constraints on its global target size. Note that the use of the inequality constraints is motivated by the choice of the barrier function in the constrained problem, which is further detailed in Section 7.3.3. Hence, we aim at minimizing the following constrained optimization problem:

$$\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \quad \mathcal{L}_{VAE}(\boldsymbol{\theta},\boldsymbol{\phi}) \qquad \text{s.t.} \quad f_c(\mathbf{a}^n) \leq 0, \quad n = 1,...,N \tag{7.3}$$

where $f_c(\mathbf{a}^j) = (1 - \frac{1}{|\Omega_i|}\sum_{l\in\Omega_i}\mathbf{a}^n_l)$ is the constraint over the attention map from the *j-th* image, which enforces the generated attention map to cover the whole image. It is well-known in optimization that a penalty does not act as a barrier near the boundary of the feasible set [216]. In other words, a constraint that is satisfied results in a null penalty and gradient. Therefore, at a given gradient update, there is nothing that prevents a satisfied constraint from being violated, causing oscillations between competing constraints and ultimately resulting in a potential unstable training. This is further exacerbated in the case of many multiple constraints (i.e., [60]), motivating the use of a single global constraint to achieve a maximum coverage of class-activation maps over the whole image in our scenario. From Eq. 7.3 we can derive an approximate unconstrained optimization problem by employing a penalty-based method, which takes the hard constraint and moves it into the loss function as a penalty term ($\mathcal{P}(\cdot)$): $\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \mathcal{L}_{VAE}(\boldsymbol{\theta},\boldsymbol{\phi})+\lambda\mathcal{P}(f_c(\mathbf{a}))$. Thus, each time that the constraint $f_c(\mathbf{a}^n) \leq 0$ is violated, the penalty term $\mathcal{P}(f_c(\mathbf{a}^n))$ increases.

7.3.3 Extended log-barrier as an alternative to penalty-based functions

Despite having demonstrated a good performance in several applications ([49, 214, 217, 218]) penalty-based methods have several drawbacks. First, these unconstrained minimization problems have increasingly unfavorable structure due to ill-conditioning ([219, 220]), which typically results in an exceedingly slow convergence. Second, finding the optimal penalty weight is not trivial. In addition, we advocate for the use of the log-barrier extension versus penalties due to the strictly positive gradient of the latter becomes higher when a satisfied constraint approaches violation during optimization, pushing it back towards the feasible set (See Figure 1 in [48]). As explained in the previous section, this contrasts with penalties, as they deliver null gradients if a given constraint is satisfied. To address these limitations, we replace the penalty-based functions by the approximation of log-barrier[1] presented in [48]. We would like to stress that barrier methods require the interior of the feasible sets to be non-empty and they are used, therefore, in constrained optimization problems with inequality constraints, such as the one

[1] Note that this function is convex, continuous and twice-differentiable.

defined in Eq. 7.3 (note that there is no interior for equality constraints). Thus, we can formally define the approximation of log-barrier as:

$$\widetilde{\psi}_t(z) = \begin{cases} -\frac{1}{t}\log(-z) & \text{if } z \leq -\frac{1}{t^2} \\ tz - \frac{1}{t}\log(\frac{1}{t^2}) + \frac{1}{t} & \text{otherwise,} \end{cases} \tag{7.4}$$

where t *controls* the barrier during training, and z is the constraint $f_c(\mathbf{a}^n)$. Thus, by taking into account the approximation in 7.4, we can solve the following unconstrained problem by using standard Gradient Descent:

$$\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \quad \underbrace{\mathcal{L}_{VAE}(\boldsymbol{\theta},\boldsymbol{\phi})}_{\text{Standard VAE loss}} + \underbrace{\lambda_s \sum_{n=1}^{N} \widetilde{\psi}_t(1 - \frac{1}{|\Omega_i|}\sum_{l\in\Omega_i}\mathbf{a}_l^n)}_{\mathcal{L}_s\text{: Size regularizer}} \tag{7.5}$$

In this scenario, for a given t, the optimizer will try to find a solution with a good compromise between minimizing the loss of the VAE and satisfying the constraint $f_c(\mathbf{a}^n)$. In the following, we refer to this formulation of gradient-CAM constraint as GradCAMCons setting.

7.3.4 On the role of gradients in VAEs

Even though there exist a few initial attempts to integrate attention maps on the task of unsupervised anomaly detection, how gradient-based attention behave on anomalous patterns remains unclear. For instance, [61] argue that anomalies produce larger gradients in the learned latent representation, which results in higher activated attention maps. On the other hand, [60] states that the VAE only focus on normal patterns (with which it has been trained), thus anomalous regions produce smaller absolute value gradients. These inconsistencies in the literature have motivated us to analyze the underlying role of the gradients in the context of brain images analysis. Thus, we performed several experiments to analyze the behaviour of grad-CAMs in anomaly localization compared to non-weighted activation maps (AMs), which are computed as:

$$\mathbf{a}^n = \frac{1}{K}\sum_k^K f^s_{\boldsymbol{\theta}}(\mathbf{x}^n)_k \tag{7.6}$$

In particular, we could not find any benefit on gradients weighting other than serving as a scaling factor for attention maps to fall on non-saturated range of values of typically used activation functions, such as the sigmoid operation in Eq. 7.2 (see Figure 7.2, where we show that the values obtained by both types of attention are highly correlated). Furthermore, we found that the reconstructed images derived from the gradient-based attention contained more errors compared to those reconstructed with attention on the activation maps (Eq 7.6). We refer the reader to Section 1 of Supplemental Material for the detailed results concerning the role of the gradients.

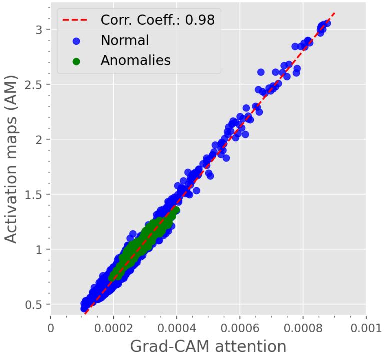


Figure 7.2: Relation between the activation values and gradient-weighted attention maps in an unconstrained VAE. These results demonstrate that the values obtained by Grad-CAM based attention are highly correlated (correlation coefficient = 0.98) to those obtained by the attention maps, suggesting that the gradient basically contributes as a scaling factor on the attention maps.

7.3.5 Entropy maximization as a proxy for the constraint

Based on our previous findings, we advocate that the use of non-weighted activation maps (AMs) should be preferred over their gradient-based counterpart. Nevertheless, this solution has a main limitation that hinders the use of size constraints. As the activation maps are not normalized, the arbitrary activation value to impose the constraint loses the sense of *size* or *proportion.* The activation values produced by neural networks can vary in each application, as well as with the architecture used, which makes it difficult to establish generalizable restrictions on their value. For this reason, we propose to use attention maps derived from normalizing the activation maps over all the pixels of the image, via a softmax activation, similarly to [46], such that: $p^n = \tau_{\Omega_B}(\mathbf{a}^n)$[2]. Since these attention maps are normalized across pixels and not over classes, the use of global constraints is meaningless, as the sum over all the pixels post-softmax will be equal to 1.0. Nevertheless, we still aim at regularizing the attention distribution p^n to focus on all patterns in the image *homogeneously.* To this end, we propose to minimize the KL distance $D_{KL}(p||q) = H(p,q) - H(p)$ between the attention distribution p, and a constant distribution q, where $H(p,q)$ represents the cross-entropy between both distributions, and $H(p) = H(p,p)$ is the Shannon entropy of the intensity distribution such that $H(p) = -\frac{1}{I}\sum_i p_i \cdot log(p_i)$. In the scenario where we want p to match a constant distribution, it is straightforward to see that minimizing the KL distance is equivalent to maximizing the entropy $H(p)$:

$$D_{KL}(p||q) = H(p,q) - H(p) =^c -H(p) \tag{7.7}$$

where $=^c$ indicates equality up to an additive constant.

Thus, the proposed constrained optimization problem integrating an entropy maximization term, referred to as $\mathcal{L}_H$, offers a softer attention constraint compared to the solution in Eq. 7.5. Furthermore, this formulation allows the VAE to keep the most suitable activation values, while requiring less hyper-parameters to be optimized. Analogously to

[2] Note that τ is the softmax activation on the brain tissue instances, Ω_B.

Eq. 7.5, we solve the constrained optimization problem with $\mathcal{L}_H$ by using standard Gradient Descent:

$$\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \quad \underbrace{\mathcal{L}_{VAE}(\boldsymbol{\theta},\boldsymbol{\phi})}_{\text{Standard VAE loss}} \underbrace{-\lambda_H \frac{1}{N}\sum_{n=1}^{N} H(\tau_{\Omega_B}(\mathbf{a}^n))}_{\mathcal{L}_H\text{: Entropy regularizer}} \tag{7.8}$$

Hereafter, we will refer to this formulation as AMCons.

7.3.6 Inference

During inference, we use the generated attention as an anomaly saliency map. For the Grad-CAMs based settings we replaced the sigmoid operation by a minimum-maximum normalization in order to avoid saturation caused by large activations. During the experimental stage, we found that anomalies produce larger activation on attention maps than the constrained normal samples, in line to prior literature ([61]). Then, the map is thresholded to create an anomaly mask of the image.

7.4 Experimental setting

7.4.1 Datasets

The experiments described in this work are carried out in the context of brain lesions localization. Concretely, we use two relevant neuroimaging challenges: tumour segmentation in MRI volumes and intracranial hemorrhage (ICH) segmentation in CT scans.

Brain tumor segmentation For this task, we used the popular BraTS 2019 dataset ([221–223]), which contains 335 multi-institutional multi-modal MR scans with their corresponding Glioma segmentation masks. Following [54], from every patient, 10 consecutive axial slices of FLAIR modality of resolution 224 × 224 pixels were extracted around the center to get a pseudo MRI volume. Then, the dataset is split into training, validation and testing groups, with 271, 32 and 32 patients, respectively.

Following the standard literature, during training only the slices without lesions are used as normal samples. For validation and testing, scans with less than 0.01% of tumour are discarded, following the standard practices in the literature.

Intracranial hemorrhage segmentation We use the Physionet-ICH dataset ([224–226]) to localize intracranial hemorrhage lesions. The dataset is composed of 82 non-contrast CT scans of subjects with traumatic brain injury. From those, 36 cases are diagnosed with intracranial hemorrhage of different types: Intraventricular, Intraparenchymal, Subarachnoid, Epidural and Subdural. ICH Lesions were slice-wise delineated by two expert radiologists. In our work, we join the different ICH types into one single label for binary lesion segmentation. CT scans are skull-stripped, intensity-normalized, and co-registered into a reference scan. Similar to the BraTS dataset, 10 consecutive axial slices of resolution 224 × 224 pixels around the center were extracted to get CT pseudo volumes. The dataset is divided into training, validation and testing splits. The first one contains only non-ICH cases (n=46), while cases with labeled lesions were used for validation (n=6) and testing (n=30). Although the main core of ablation experiments in this work are described on the BraTS dataset, we use the Physionet-ICH dataset to demonstrate the generalization capabilities of our proposed method on different brain lesions and imaging modalities.

7.4.2 Evaluation metrics

We resort to standard metrics for unsupervised brain lesion segmentation, as in [210]. Concretely, we compute the dataset-level area under precision-recall curve (AUPRC) at pixel level, as well the area under receptive-operative curve (AUROC). From the former, we obtain the operative point (OP) as threshold to generate the final segmentation masks. Then, we compute the best dataset-level Sørensen-Dice score ($\lceil$DICE$\rceil$) and intersection-over-union ($\lceil$IoU$\rceil$) over these segmentation masks. Finally, we compute the average Sørensen-Dice score (DICE) over single scans. For each experiment, the metrics reported are the average of three consecutive repetitions of the training, to account for the variability of the stochastic factors involved in the process.

7.4.3 Implementation details

The VAE architecture used in this work is based on the recently proposed framework in [60]. Concretely, the convolution layers of ResNet-18 ([91]) are used as the encoder, followed by a dense latent space $\mathbf{z} \in \mathbb{R}^{32}$. For image generation, a residual decoder is used, which is symmetrical to the encoder. It is noteworthy to mention that, even though several methods have resorted to a spatial latent space ([54, 60]), we observed that a dense latent space provided better results, which aligns to the recent benchmark in [210]. To train the GradCAMCons formulation in eq. 7.5 we first trained the VAE during 50 epochs without any expansion to stabilize the convergence using $\beta = 1$. Then, the proposed regularizer was integrated (equation 7.5) with $t = 10$ and $\lambda_s = 10^3$ applied to the Grad-CAMs obtained from the first convolutional block of the encoder during 250 epochs. We use a batch size of 8 images, and a learning rate of $1e{-}5$ with ADAM as optimizer. The reconstruction loss, $\mathcal{L}_R$, in eq. (7.1) is the binary cross-entropy. Similarly, the AMCons formulation in eq. 7.8 was trained by using $\beta = 10$ and $\lambda_H = 0.1$, using a learning rate of $1e{-}4$. Ablation experiments to motivate the choice of values used are presented in Section 7.5.2 and Section 3 of supplemental materials. The code and trained models are publicly available on (`https://github.com/jusiro/constrained_anomaly_segmentation/`).

7.4.4 Baselines

In order to compare our approach to state-of-the-art methods, we implemented prior works and validated them on the dataset used, under the same conditions. First, we use residual-based methods to match the recently benchmark on unsupervised lesion localization in [210]. Then, we implement up-to-date methods based on contrast adjustment on the input image via histogram equalization. We also include recently proposed methods that integrate CAMs to locate anomalies. For both strategies, the AE/VAE architecture was the same as the one used in the proposed method. ***Residual methods***, given an anomalous sample, aim to use the AE/VAE to reconstruct its normal counterpart. Then, they obtain an anomaly localization map using the residual between both images such that $\mathbf{m} = |\mathbf{x} - \hat{\mathbf{x}}|$, where $|\cdot|$ indicates the absolute value. On the AE/VAE scenario, we include methods which propose modifications over vanilla versions, including context data augmentation in Context AE [208], Bayesian AEs ([200]), Restoration VAEs ([202]), an adversarial-based VAEs, AnoVAEGAN ([54]) and a recent GAN-based approach, F-anoGAN ([59]). For methods including adversarial learning, DC-GAN [26] is used as discriminator. During inference, residual maps are masked using a slight-eroded brain mask, to avoid noisy reconstructions along the brain borderline. ***Equalization-based methods***: very recent methods have highlighted the limits of residual-based approaches to properly discern brain lesions [212, 227]. In contrast, they propose to apply an equalization of the histogram of the input image, and to set a threshold on the preprocessed image, considering that brain lesions often show hyperintense patterns in different modalities. Concretely, we include the method proposed in [227], which we refer to as HistEq. ***CAMs-based***: we use Grad-CAM VAE ([61]), which obtains regular Grad-CAMs on the encoder from the latent space $\mathbf{z}_\mu$ of a trained vanilla VAE. Concretely, we include a disentanglement variant of CAMs proposed in this work, which computes the combination of individually-calculated CAMs from each dimension in $\mathbf{z}_\mu$, referred to as Grad-CAM$_D$ VAE. We also use the recent method in [60] (CAVGA), which applies a L1 penalty on the generated CAM to maximize the attention. In contrast to our model and [61], the anomaly mask in [60] is generated by focusing on the regions not activated on the saliency map such that $\mathbf{a} = 1 - CAM$, hypothesizing that the network has learnt to focus only on normal regions. Then, $\mathbf{a}$ is thresholded with 0.5 to obtain the final anomaly mask $\mathbf{m} \in \mathbb{R}^{\Omega_i}$. For both methods, the network layer to obtain the Grad-CAMs is the same as in our method.

7.5 Results

7.5.1 Comparison to the literature.

The quantitative results obtained by the proposed model and baselines on the test cohort are presented in Table 7.1. Results from residual-based baselines range between [0.056-0.511](AUPRC) and [0.188-0.525] (DICE), which are in line with previous literature [210]. We can observe that the proposed formulations outperform these approaches by a large margin. Concretely, the AMCons method provides a substantial increase of ∼34% and ∼26% in terms of AUPRC and DICE, respectively, compared to the best model, i.e., F-anoGAN. Furthermore, the model integrating the $\mathcal{L}_H$ term significantly outperforms our previous method in [204]. This supports our hypothesis that using non-weighted attention maps with a maximization entropy term as constraint is indeed a better solution for the unsupervised lesion segmentation task. Finally, in comparison with the very recently proposed method of histogram equalization, HistEq, our proposed formulation brings improvements of nearly ∼10% in the main figures of merit.

Method	**AUROC**	**AUPRC**	$\lceil$**DICE**$\rceil$	$\lceil$**IoU**$\rceil$	**DICE** ($\mu \pm \sigma$)
CAVGA ([60])	0.726(0.001)	0.056(0.005)	0.188(0.001)	0.104(0.002)	0.182(0.004)±0.096(0.002)
Bayesian VAE ([200])	0.922(0.002)	0.193(0.005)	0.342(0.005)	0.206(0.005)	0.329(0.005)±0.115(0.005)
AnoVAEGAN ([54])	0.925(0.020)	0.232(0.052)	0.359(0.074)	0.221(0.053)	0.349(0.071)±0.115(0.015)
Bayesian AE ([200])	0.940(0.002)	0.279(0.009)	0.389(0.012)	0.242(0.009)	0.375(0.010)±0.130(0.011)
AE	0.937(0.002)	0.261(0.011)	0.397(0.011)	0.248(0.008)	0.386(0.010)±0.125(0.004)
Grad-CAM$_D$ VAE ([61])	0.941(0.003)	0.312(0.010)	0.400(0.009)	0.250(0.012)	0.361(0.014)±0.164(0.005)
Restoration VAE ([202])	0.934(0.028)	0.352(0.111)	0.403(0.099)	0.252(0.069)	0.345(0.075)±0.186(0.044)
Context VAE ([208])	0.939(0.004)	0.271(0.017)	0.406(0.020)	0.255(0.016)	0.394(0.017)±0.126(0.007)
Context AE ([208])	0.940(0.003)	0.278(0.012)	0.411(0.014)	0.259(0.011)	0.399(0.013)±0.126(0.005)
VAE ([54, 203])	0.940(0.002)	0.273(0.010)	0.411(0.012)	0.259(0.009)	0.399(0.010)±0.127(0.004)
F-anoGAN ([59])	0.946(0.026)	0.511(0.190)	0.525(0.147)	0.369(0.131)	0.494(0.138)±0.151(0.038)
GradCAMCons w. $\mathcal{L}_S$ (L2 penalty)	0.969(0.015)	0.567(0.138)	0.620(0.085)	0.455(0.086)	0.586(0.079)±0.184(0.028)
HistEq ([227])	0.972(0.000)	0.725(0.000)	0.705(0.000)	0.545(0.000)	0.653(0.000)±0.233(0.000)
GradCAMCons w. $\mathcal{L}_S$ (Log Barrier)	0.982(0.001)	0.746(0.034)	0.698(0.034)	0.537(0.041)	0.677(0.021)±0.215(0.019)
AMCons w. $\mathcal{L}_H$	**0.988**(0.000)	**0.850**(0.011)	**0.786**(0.009)	**0.648**(0.013)	**0.741**(0.009)±**0.153**(0.001)

Table 7.1: Comparison to prior literature on BraTS dataset. Results derived from the proposed methods in gray. Best results in bold. The values in parentheses indicate the standard deviation over the three training repetitions.

7.5.2 Ablation experiments

The following ablation studies aim at demonstrating, in an empirical way, the motivation of employing the proposed models. First, we provide quantitative evidences about the better performance of using global constraints (model in Eq. 7.5) over pixel-level constraints (i.e., [60]). Second, we show that resorting to the extended log-barrier function is a better alternative than standard L2 penalty functions. Then, we perform an in-depth analysis of the optimal hyperparameters values for the entropy-guided model (Eq. 7.8), as well as other important design choices.

Image vs. pixel-level constraint The following experiment demonstrates the benefits of imposing the constraint on the whole image rather than in a pixel-wise manner, such as in [60]. In particular, we compare the two strategies when the constraint is enforced via a L2-penalty function, whose results are presented in Table 7.2. In particular, we can easily see that imposing the constraint at image-level consistently outperforms pixel-level constraints. These results support our hypothesis that global constraints, such as the proposed formulation in Eq. 7.5, should be preferred over multiple pixel-wise constraints, similar to [60].

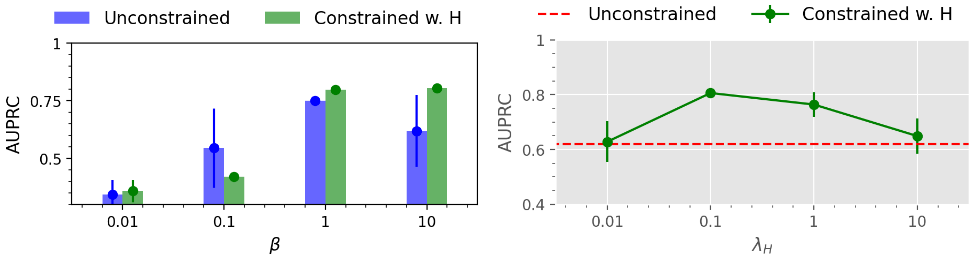
	L2 (pixel-level)	L2 (image-level)	Log-Barrier (image-level)
AUPRC	0.489(0.098)	0.550(0.160)	0.728(0.034)

Table 7.2: Quantitative comparison, in terms of AUPRC, between enforcing the constraint at pixel-level (i.e., [60]) or at image-level (i.e., proposed approach), and for the impact of the type of regularization.

Extended log-barrier vs. penalty-based functions To motivate the choice of employing the extended log-barrier over standard penalty-based functions in the constrained optimization problem in Eq. (7.3), we compare them in Table 7.2. It can be observed that imposing the constraint with the extended log-barrier consistently outperforms the L_2-penalty, with substantial performance gains.

On the impact of entropy-guided constraints We now perform an in-depth analysis of the effect of integrating the entropy-guided constraint in Eq. 7.8 for anomaly localization, as well as an extensive validation of the values of the balancing terms β and λ_H. First, we study the impact of $\mathcal{L}_H$ across different β values (i.e. $\beta = \{0.01, 0.1, 1, 10\}$), by fixing its balancing term λ_H to 0.1, a value that empirically showed good stability. These results, which are reported in Figure 7.3a, show that the VAE with and without entropy constraint presents different optimal values for β. Nevertheless, the best results are obtained when the contribution of the regularization term is large (i.e. $\beta \geq 1$), and the entropy-based regularization over the activation maps included (i.e., green bars). Furthermore, this configuration is shown to be more stable once a large β weight is set, particularly for the constrained formulation. Then, based on the best configuration ($\beta = 10$), we study how different

λ_H weights $\{0.01, 0.1, 1, 10\}$ impact the model performance. These results (Figure 7.3b) show that incorporating the entropy regularization always contributes to performance gains, with an optimum weight value of $\lambda_H = 0.1$.



(a) (b)

Figure 7.3: Ablation study on the AMCons setting. Concretely, the role of the KL regularization (β) in the VAE and the entropy constraint on attention maps (λ_H) from our formulation is studied. (a) Entropy constraint effect and dependency on β. (b) Ablation study on λ_H.

In the next experiment, we show how adding the L_H term in our formulation impacts the activation maps (AM). Concretely, we first show in Figure 7.4 the AM distribution for a normal sample for both the constrained and unconstrained configurations. It can be observed that, in our constrained formulation, the distribution of activation values is more homogeneous (in orange), unlike the more spread values found in its unconstrained counterpart (in green). Furthermore, we show its impact on unseen, anomalous samples, where the benefits of our model are better highlighted. In particular, we represent the AM distribution for normal and anomalous pixels on the unconstrained formulation (i.e. $\lambda_H = 0$) in Figure 7.5 (*top*), and the effect of integrating the L_H term (Figure 7.5, *bottom*). Similarly to the normal samples, the distribution of normal pixels produced by the unconstrained setting spreads over a larger range, resulting in a higher overlapping with the distribution of anomalous pixels. Note that, in addition to the overlapping regions, there exist values of normal pixels which overpass anomalous values. In contrast, the more compact distribution provided by the proposed formulation favors a smaller overlap between normal and anomalous pixel intensity distributions. This results in an easier identification of normal *versus* anomalous pixels.

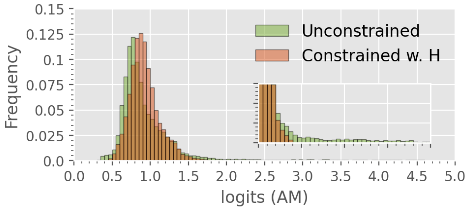


Figure 7.4: Influence of the entropy constrained term on the attention maps for AMCons on normal images.

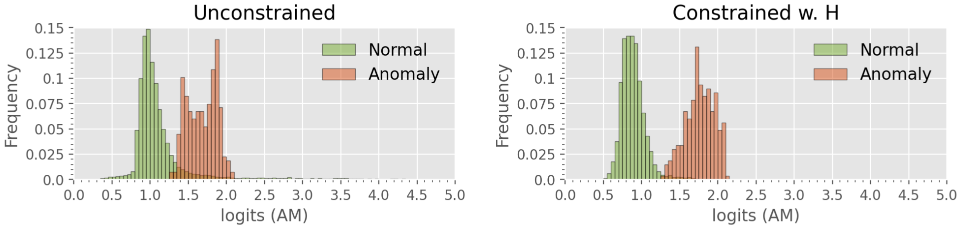


Figure 7.5: Influence of the entropy constrained term on the attention maps for AMCons on images with anomalies.

In the following, we explore how the entropy constraint favors the smallest overlap between normal and anomalous distribution on the objective criteria, compared to previous literature. To do so, we depict in Figure 7.6 the distribution of both populations for the proposed methods, AMCons and GradCAMCons, and the most promising baselines, F-anoGAN and Histeq. Furthermore, we obtain the overlap between both distributions by dividing the number of samples in the overlapped region of the histograms by the total number of samples. It can be seen how the proposed method based on entropy maximization obtains the smallest overlap (10.2%) and produces a narrower distribution of normal samples in comparison with the GradCAMCons method, based on size constraints.

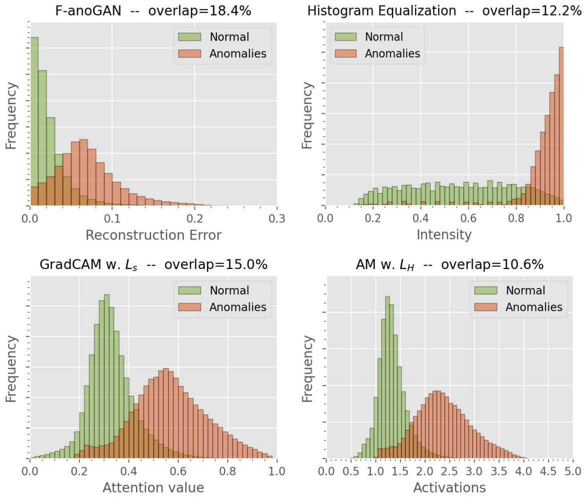


Figure 7.6: Histogram analysis on the overlap of normal and anomalous samples for the different proposed methods and baselines (on the whole BRATS dataset).

Using statistics from normal domain for anomaly localization threshold A common practice on unsupervised anomaly segmentation is to use anomalous images to define the threshold to obtain the final segmentation masks. In particular, these methods look at the AUPRC on the anomalous images, which is then used to compute the optimal threshold value. We refer to this technique in our experiments as OP (Operative Point). To alleviate the need of anomalous samples during the validation stage, several methods ([54]) have discussed the possibility of using a given percentile from the normal images (i.e., no anomalies) distribution to set the threshold. Motivated by this, an ablation study on the percentile value is presented in Table 7.3 for our proposed formulations and the best performing baselines. First, we can observe that under the OP strategy (i.e., accessing to anomalous images to identify the optimal threshold), both of our models bring substantial improvements over the state-of-the-art on residual-based approaches, ranging from 14% to 22%. If we resort to the percentiles instead, the performance improvements observed are very similar to the OP scenario, with our models outperforming F-anoGAN by a large margin. Nevertheless, we observed that the best results are obtained with different percentile values. While F-anoGAN and AMCons w. $\mathcal{L}_H$ yields the best performance using the 98% percentile, GradCAMCons w. $\mathcal{L}_S$ follows previous observations in [54], performing better using the 95% percentile.

	OP	th=0.5	p85	p90	p95	p98
F-anoGAN	0.525	–	0.310	0.390	0.505	0.488
HistEq	0.690	–	0.298	0.404	0.624	0.620
GradCAMCons w. $\mathcal{L}_S$	0.693	0.583	0.512	0.611	0.663	0.587
AMCons w. $\mathcal{L}_H$	0.743	–	0.189	0.201	0.265	0.720

Table 7.3: Ablation study on threshold values from normal images. pX indicates the average percentile used on the training set (normal images) to compute the segmentation threshold. OP indicates the operative point from area under precision-recall curve, using all validation dataset, which contains anomalous images. The metric presented is the dataset-level DICE.

This suggests that, even though not used directly, anomalous images are still required to find the optimal threshold value. However, the proposed method GradCAMCons shows special properties that suggest that they can achieve large performance gains without having access to anomalous images to define the threshold, unlike prior works. In particular, our GradCAM-based formulation restricts the attention values to $[0, 1]$, which allows to set a typical threshold to 0.5, with still large performance gains (+7%) compared to the baselines. Nevertheless, we can observe that if we resort to the percentile strategy, our method based on maximizing the entropy of the attention maps (i.e., AMCons) is very sensitive to the selected value.

Number of slices to generate the pseudo-volumes In our experiments, we followed the standard literature ([210]) to generate the pseudo-labels for validation and testing. Nevertheless, we concede that this scenario is unrealistic, as the appropriate number of slices used from the MRI scans in unsupervised anomaly detection should be unknown. We now explore the impact of including more slices in these pseudo-volumes, which increase the variability of normal samples. For instance, it is well-known that the target regions in slices farther from the center are incrementally smaller. In this line, we hypothesize that the dimension of the VAE latent space and the importance of the KL regularization may be a determining factors in absorbing this increased variability. Regarding the latent space, the appropriate $\mathbf{z}$ dimension is unclear in the literature. For instance, [210] uses $\mathbf{z} = 128$, while [54] uses $\mathbf{z} = 64$, and we obtained better results using $\mathbf{z} = 32$. To validate the proposed experimental setting and latent space dimension, we now present results using increasing number of slices around the axial midline $N = \{10, 20, 40\}$, and two different latent space dimensions $\mathbf{z} = \{32, 128\}$ for both a standard VAE and our proposed models, in Figure 7.7a. We can observe that despite the gap between the baselines and the attention based methods is reduced as the number of slides is increased, this difference is still significant, and the relative performance drop is similar for all methods. Finally, we can observe that an increasing on $\mathbf{z}$ dimension (solid *versus* dotted lines in Fig 7.7a) does not produce gains in performance in any case. Note that the model hyperparameters used are optimized for $z = 32$, and $N = 10$, which also could produce some underestimation of the proposed model performance when N increases. In the following, we study the performance of the proposed AMCons method using different β values ($\beta = \{1, 10\}$) in the KL term of eq. 7.1 across different number of slices, whose results are presented in Figure 7.7b. We can observe that, by decreasing the value of β as the number of employed slices increases, we can alleviate the performance degradation observed with a fixed β. Since the KL regularization directly affects the capacity of the VAE for learning different samples, the optimization of its balancing term when increasing the domain of samples used seems necessary. The similar behaviour between the proposed method and baselines suggest that this could be a limitation of self-training features based on VAEs, which struggle to encode heterogeneous sample information.

7.5.3 Generalization to other datasets

In order to empirically demonstrate the generalization properties of the proposed methodology, we evaluate its performance on a different dataset for brain lesion detection. Concretely, as previously described, we resort to Physionet-ICH dataset for non-contrast CT on ICH localization. Implementation details are analogous as the ones used on the BraTS dataset, although we decreased the learning rate to $1e-5$, and we set a larger latent dimension, i.e. $\mathbf{z} \in \mathbb{R}^{128}$, along all baselines and

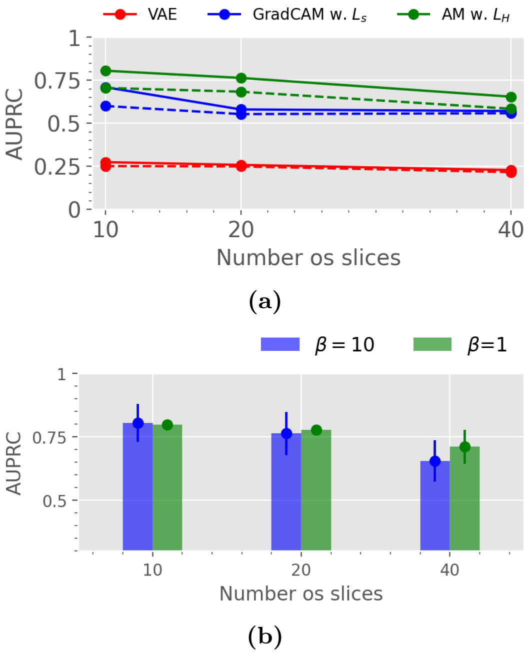


Figure 7.7: Ablation study on the effect of increasing the number of axial slices around the center used from MR brain volumes. (a) Study of latent space dimension for the proposed models and an standard VAE. Solid lines indicate $\mathbf{z} = 32$, and dashed lines denote $\mathbf{z} = 128$. (b) Study of the KL component importance (β term) using the proposed AMCons method.

methods to favour model convergence. Obtained results for anomaly localization are reported in Table 7.4. Even though there exist slight differences in the comparison between residual methods in the literature compared to the results obtained on BraTS dataset (i.e. the simple AE outperforms variations approaches), the proposed attention-based anomaly localization methods still achieve remarkable results. Again, the AMCons configuration yields the best performance, and it reaches improvements of nearly ∼25% and ∼18% in terms of AUPRC and DICE, respectively, compared to previous literature. The observed results suggest that the proposed methodology is able to generalize to other unsupervised brain lesion segmentation challenges, even using different imaging modalities. It should be noted, however, that the absolute results in terms of segmentation are lower than those obtained in BraTS. Among other reasons, this may be due to the greater heterogeneity observed in the ICH dataset, the lower degree of standardization and size of the database used, and the small size of ICH lesions, which penalizes metrics such as DICE. Nevertheless, the values obtained are in line with the scarce previous literature on ICH segmentation, as reflected in Table 7.4. Indeed, the obtained results are at par with previous works using a fully supervised learning approach [225], which shows the difficulty of the task.

Method	**AUROC**	**AUPRC**	**$\lceil$DICE$\rceil$**	**$\lceil$IoU$\rceil$**	**DICE ($\mu \pm \sigma$)**
Other works					
Karkkainen et al. (2021) (Unsupervised)*	–	–	–	–	0.197 ± 0.222
Hssayeni et al. (2020) (Supervised)	–	–	–	–	0.315 ± 0.211
Physionet-ICH dataset					
CAVGA ([60])	0.919(0.004)	0.061(0.003)	0.094(0.005)	0.062(0.004)	0.053(0.004)±0.161(0.002)
Grad-CAM$_D$ VAE ([61])	0.955(0.003)	0.157(0.009)	0.275(0.011)	0.159(0.005)	0.178(0.005)±0.175(0.003)
Bayesian AE ([200])	0.961(0.001)	0.188(0.006)	0.309(0.009)	0.183(0.007)	0.242(0.008)±0.181(0.003)
VAE ([54, 203])	0.962(0.000)	0.167(0.005)	0.319(0.002)	0.190(0.002)	0.245(0.004)±0.192(0.003)
AnoVAEGAN ([54])	0.961(0.000)	0.167(0.003)	0.313(0.006)	0.185(0.004)	0.239(0.006)±0.192(0.002)
Bayesian VAE ([200])	0.964(0.000)	0.178(0.010)	0.323(0.007)	0.193(0.005)	0.248(0.008)±0.191(0.004)
Context VAE ([208])	0.963(0.002)	0.170(0.013)	0.321(0.023)	0.191(0.016)	0.243(0.014)±0.191(0.009)
Restoration VAE ([202])	0.962(0.001)	0.183(0.005)	0.327(0.002)	0.187(0.001)	0.233(0.004)±0.189(0.003)
Context AE ([208])	0.962(0.001)	0.195(0.005)	0.359(0.010)	0.219(0.007)	0.276(0.004)±0.198(0.004)
F-anoGAN ([59])	0.961(0.000)	0.173(0.007)	0.343(0.007)	0.207(0.005)	0.268(0.007)±0.191(0.005)
AE	0.961(0.001)	0.176(0.006)	0.344(0.007)	0.208(0.006)	0.266(0.002)±0.202(0.005)
GradCAMCons w. $\mathcal{L}_S$ (L2 penalty)	0.967(0.009)	0.261(0.013)	0.361(0.067)	0.231(0.029)	0.276(0.046)±0.243(0.029)
HistEq ([227])	0.963(0.000)	0.313(0.000)	0.385(0.000)	0.239(0.000)	**0.348**(0.000)±**0.213**(0.000)
GradCAMCons w. $\mathcal{L}_S$ (Log Barrier)	0.970(0.008)	0.295(0.073)	0.401(0.044)	0.251(0.049)	0.286(0.076)±0.233(0.039)
AMCons w. $\mathcal{L}_H$	**0.971**(0.006)	**0.420**(0.068)	**0.522**(0.046)	**0.354**(0.043)	0.319(0.054)±0.266(0.011)

* Results reported on a different (private) dataset.

Table 7.4: Comparison to prior literature on Physionet-ICH dataset, and previous works on ICH segmentation. Results derived from the proposed methods are depicted in gray, and best results are indicated in bold.

7.5.4 Qualitative evaluation

Visual results of the proposed and existing methods for both datasets are depicted in Figure 7.8. We can observe that our approach identifies as anomalous more complete regions of the lesions, whereas existing methods are prone to produce a significant amount of false positives (*first, third and seventh* rows) and fail to discover many abnormal pixels (*third row*). These visual results are in line with the quantitative validation performed in previous sections. However, there is a known problem about segmenting only hyperintense regions in the state-of-the-art methods of unsupervised anomaly localization of brain lesions ([227]). Although the proposed method still suffers from this limitation (*fourth row, red arrow*), the positive results regarding true negative segmentation obtained in some normal, hyperintense tissue (*second row, green arrow*) suggest an improvement in relation to this problem.

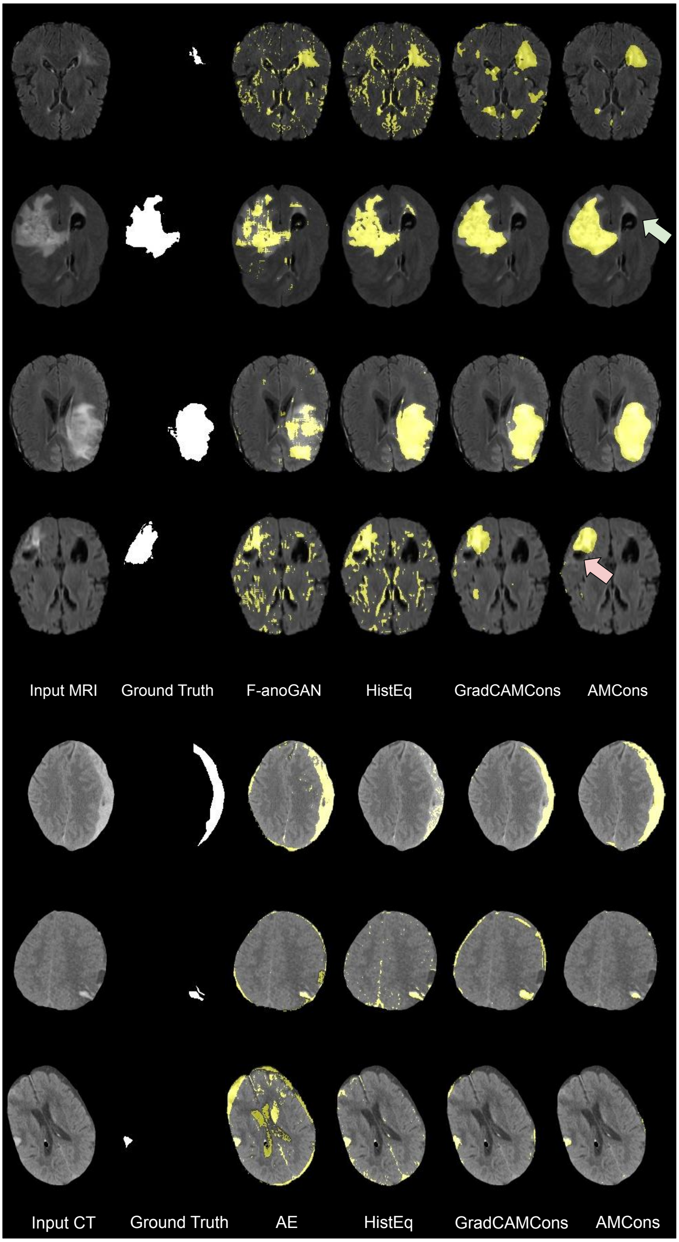


Figure 7.8: Qualitative evaluation of the proposed and existing high-performing methods for anomaly localization on BraTS MRI flair volumes (top) and on Physionet-ICH non-contrast CT images (bottom). A failure case is depicted with the red arrow (*fourth column*).

7.6 Conclusions

Despite the recent advances of unsupervised anomaly segmentation in medical problems, existing literature still provides limited performance, with most methods yielding suboptimal results in popular segmentation benchmarks. In this work, we have presented a novel approach that substantially differs from prior literature in several aspects.

First, we resort to generated attention maps to identify anomalous regions, which contrasts with most existing works that rely on the pixel-wise reconstruction error. Second, our formulation integrates a size-constrained loss that enforces the attention maps to cover the whole image in normal images. This differs from very recent works [60], as we tackle this problem by imposing inequality constraints on the whole target attention maps. Another important difference lies on the manner the constrained problem is addressed. While [60] leverages a L2 penalty function, we resort to an extension of standard log-barrier methods, which overcome the well-known limitations of penalty-based methods. Quantitative results demonstrate that this model significantly outperforms prior literature on unsupervised lesion segmentation.

A drawback of the log-barrier based formulation is that it requires to find the optimal value for several hyperparameters. Motivated by this, we have proposed an alternative model, which integrates a regularization term that maximizes the Shannon entropy on the generated attention maps. This new formulation only adds the entropy balancing term $\mathcal{L}_H$, which reduces the complexity compared to the constrained problem in eq. 7.5. Furthermore, as reported in the results, the maximum-entropy model yields better performance than the size regularizer formulation. Note, in addition, that the alternative entropy-based model better separates the intensity distributions between normal and abnormal tissue. This allows us to employ a higher percentile value to obtain the final anomalous regions, with a substantial performance improvement compared to previous methods. Thus, based on the reported empirical validation, the proposed models represent a novel state-of-the-art for unsupervised anomaly segmentation.

We believe that there exist potential research directions to further improve the performance of unsupervised segmentation methods. For example, brain images are typically acquired along multiple modalities. Learning how to combine multiple modalities in the scenario of anomalous regions detection might indeed enhance the learned representation by the VAE, ultimately resulting in better identification of abnormal pixels. In addition, unsupervised segmentation methods have been only evaluated from a discriminative perspective. Nevertheless, assessing their performances in terms of the quality of the uncertainty estimates, i.e., calibration, might give a better overview of the quality of a segmentation model.

Chapter 8

Final conclusions

This chapter relates the findings from each paper to the final aim of the PhD thesis. It summarises concluding remarks and suggests future research lines for each proposed learning framework.

Contents

8.1 Global remarks

In this thesis, we have designed, developed and validated novel *not-so-supervised* methods to solve real-world computer vision challenges using deep learning. Concretely, we have focused on weakly supervised, few-shot and unsupervised learning strategies. We have proposed self-supervised learning algorithms on weakly supervised learning for gigapixel prostate histology image classification, able to leverage highly accurate instance-level tumor grades. Then, this thesis has explored prototypical few-shot learning methods, based on the recently popularized contrastive learning optimization, to detect crossing defects on train railway surveying. Finally, we have presented a novel formulation on unsupervised anomaly localization, applied to brain lesion segmentation, based on attention regularization. In a transversely fashion, this thesis has explored how to incorporate prior knowledge for each application, in the form of constraint formulations. In particular, inequality constraints have been explored to include relative class proportions in weakly supervised classification, or attention homogenization on VAEs for anomaly localization. The methods proposed in this thesis have been extensively validated and, when possible, *not-so-supervised* solutions have been compared with their supervised counterparts.

8.2 Specific remarks

Weakly supervised classification of histology images

In Chapter 2 we first introduced the use of deep learning models based on CNNs for prostate histology biopsies grading, under the supervised learning paradigm. The difficulties and slowness encountered in the preparation of a database with local annotations, and the problems of generalization in external databases, have motivated the study of weakly supervised techniques in this thesis. Thus, in Chapter 3 we introduced an instance-based CNN able to leverage pixel-level labels from global labels on histology patches, under the weakly supervised segmentation paradigm, in a multi-label scenario. Results show the promising performance of this method, being comparable to a supervised counterpart using an UNet architecture. The difficulty for pathologists to perform accurate annotations at the pixel level leads to noisy annotations, coupled with the heterogeneity of the patterns, and the imbalance of the different categories may be behind this phenomenon. Then, in Chapter 4 we have gone further, processing entire WSIs in a weakly supervised manner under the multiple instance learning paradigm. In this line, we proposed a self-supervised learning framework based on a Teacher CNN that leverages instance labels, which are further post-processed to feed a Noisy Student trained with pseudolabels. Different Teacher architectures have been explored. The use of max-pooling on instance-based MIL showed the best performance, since the pseudolabels predicted were of high specificity. This framework was trained on the large PANDA dataset, reaching a Cohen's quadratic kappa of ~ 0.80. It is noteworthy to mention that the proposed self-supervised learning approach has been tested on three different datasets, showing robust performance, which are comparable to inter-pathologist variability. The observed results are largely better than the ones reached in Chapter 2 under the supervised learning paradigm. Thus, the capability of weakly supervised methods for training CNNs on large datasets allows a best generalization than its supervised counterparts on smaller datasets. Still, collection such large number of samples could be unfeasible on other applications with less prevalence. For that reason, we have presented in Chapter 5 a constraint optimization able to incorporate prior knowledge, in the form of relative class proportions, to the multiple instance learning formulation. By using this information during training, which is easily accessible on medical records, results are comparable with the supervised methods developed in Chapter 2 using similarly sized datasets. Interestingly, the results improve notably in the grades that coincide the most in the training bags. This shows the benefits of the method to better discern categories at the instance level, with less effort on the part of the annotators. As a qualitative sample, the database developed in Chapter 2 took years to compile, while the one used in Chapter 5 with global constraints was obtained in a matter of weeks.

Few-shot learning for railway crossing surveying

In Chapter 6 we have presented a few-shot learning formulation for railway crossing defect detection. The computer vision system takes as input axle-box accelerations, preprocessed using short-time Fourier transform spectrograms, to detect via CNNs worn crossings. The use of supervised contrastive learning to learn discriminative latent spaces, that are further used by memory-based k-nearest neighbours on inference, has shown a promising performance when dealing with small datasets. In particular, in comparison with standard episodic-based prototypical training on few-shot learning, the proposed formulation reaches improvements of $\sim 8\%$ F1-score. Likewise, it also substantially outperforms cross-entropy based standard supervised learning losses. The direct regularization of latent space, contrary to cross-entropy losses, which require to create non-calibrated decision boundaries, produces best embeddings for classification tasks on small datasets. In that refers to crossings surveting, we observed how narrower models produced best results, since spectrograms contain low-level features present in early layers of CNNs. Likewise, we explored different normalization techniques for input features. On the contrary to standard computer vision application on natural images, max or zscore normalization perform worse than a dynamic-margin standarization that allows to maintain the intensity information across samples. The methods introduced in Chapter 6 outperform largely previous, scarce literature in this field, and has a promising future ahead of it.

Unsupervised anomaly segmentation of brain lesions

Finally, we have studied the use of unsupervised deep learning, in the context of unsupervised anomaly segmentation of brain lesions, in Chapter 7. In line to very recent literature, using constrained attention maps on VAEs has shown better performance than residual-based approaches. This may be due to the difficulty that generative models have in producing accurate, high-resolution images. In addition, the incorporation of image-level constraints, instead previously proposed pixel-level, led to best results for lesion localization. In this line, two different formulations have been presented that aim to homogenize the attention maps: one based on size constraints, and the other based on Shannon entropy maximization. The second one reaches best results, needing fewer parameters for optimization. This suggests that applying softer constraints, which have led to better reconstructions in the VAE, also improves the representation of normal domain images. Thus, anomalous pixels differ more from those observed during training in the attention maps. It is worth mentioning that the constraint formulation produces compact activation distributions on normal pixels, which favors a smaller overlap between normal and anomalous pixel intensity distributions. The best results obtained with respect to the state of the art, in two different applications for brain injury detection, suggest that the method is feasible generalizable on different scenarios. Although the obtained results are still far form the supervised scenario performance, our work substantially bridges the gap, and bodes well for the future of this field.

8.3 Future work

In the present thesis, different not-so-supervised methods have been explore to palliate the deep learning necessity of large, labelled datasets to perform properly. Still, there are many research possibilities to go further in this line. To begin with, the methods studied in this thesis contained only pure methodologies. In real-world scenarios, despite focusing on weakly supervised or unsupervised methods, some labeled samples are usually available. This opens the range of possibilities for mixed methods, including semi-supervised learning mechanisms. Among other possibilities, using few labelled samples allow predicting pseudolabels on weakly- or unlabeled data, using class-specific data augmentation, distilling knowledge of supervised CNNs into less-supervised ones, or incorporating subcategory exploration methods with deep clustering techniques. Similarly, the use of few labeled data opens up other challenges to consider: uncertainty, category imbalance, and inter-annotator variability should be incorporated into the model. This can be done using Bayesian networks for example, or by incorporating prior knowledge about uncertainty in the annotations, in the form of constraint formulations, among other approaches.

Finally, we would like to point out some future perspectives on using computer vision on real-world applications. Most deep learning focus on training a model able to perform accurately in a given task. Then, this model is directly used in practice. Still, this model remains static along time, even if new, unseen data domains are used, or if it produces noisy predictions on its daily use. Further research lines should focus on active learning paradigms, able to deploy dynamic models that are updated given new data, and discontinuous supervision on its use. Challenges in this field include how to incorporate new categories to the model in the form of continual learning, how to deal with uncertainty of non-supervised data and corrections from the annotations of noisy outputs, how to avoid catastrophic forgetting, or how to integrate multicentric knowledge under the federated learning paradigm. Active learning topics are of great interest to today's scientific community and foresee an exciting future for the successful inclusion of computer vision in everyday life.

Merits

Journal papers

Silva-Rodríguez, J., Naranjo, V., & Dolz, J. Constrained unsupervised anomaly segmentation. Medical Image Analysis (2022).

Silva-Rodríguez, J., Salvador, P., Naranjo, V., & Insa, R. Supervised contrastive learning-guided prototypes on axle-box accelerations for railway crossing inspections. Expert Systems with Applications (2022).

Silva-Rodríguez, J., Schmidt, A., Sales, M.A, Molina, R., & Naranjo, V. Proportion constrained weakly supervised histopathology image classification. Computers in Biology and Medicine (2022).

Schmidt, A., **Silva-Rodríguez, J.**, Molina, R., & Naranjo, V. Coupling Semi-Supervised and Multiple Instance Learning for Histopathological Image Classification. IEEE Access (2022).

Silva-Rodríguez, J., Colomer, A., Dolz, J., & Naranjo, V. Self-learning for weakly supervised Gleason grading of local patterns. IEEE Journal of Biomedical and Health Informatics (2021).

Silva-Rodríguez, J., Colomer, A., & Naranjo, V. WeGleNet: A weakly-supervised convolutional neural network for the semantic segmentation of Gleason grades in prostate histology images. Computerized Medical Imaging and Graphics (2021).

Silva-Rodríguez, J., Colomer, A., Sales, M.A, Molina, R., & Naranjo, V. 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 (2020).

International conferences

Silva-Rodríguez, J., Naranjo, V., & Dolz, J. *Looking at the whole picture: constrained unsupervised anomaly segmentation* in *The* 32*nd* *British Machine Vision Conference (BMVC)* (2021).

Kalapahar, A., **Silva-Rodríguez, J.**, López-Mir, F., Colomer, A. & Naranjo, V. *Gleason grading of histology prostate images through semantic segmentation via residual u-net* in *2020 IEEE International Conference on Image Processing (ICIP)* (2020).

Silva-Rodríguez, J., Payá-Bosch, E., García, G., Colomer, A., & Naranjo, V. *Prostate gland segmentation in histology images via residual and multi-resolution u-net* in *International Conference on Intelligent Data Engineering and Automated Learning (IDEAL)* (2020).

Silva-Rodríguez, J., Colomer, A., Meseguer, M., & Naranjo, V. Predicting the success of blastocyst implantation from morphokinetic parameters estimated through CNNs and sum of absolute differences in *2019 27th European Signal Processing Conference (EUSIPCO)* (2019).

Research awards

Silva-Rodríguez, J., Payá-Bosch, E., García, G., Colomer, A., & Naranjo, V. (2020). Prostate gland segmentation in histology images via residual and multi-resolution u-net. *Best Paper on Image Processing Award.*

Bibliography

1. Sonka, M., Hlavac, V. & Boyle, R. *Image processing, analysis and machine vision* (1993).
2. Elgendy, M. *Deep Learning for Vision* (2020).
3. Ojala, T., Pietikäinen, M. & Harwood, D. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition* **29,** 51–59 (1996).
4. Dalal, N. & Triggs, B. *Histogram of oriented gradients for human detection* in *Computer Vision and Pattern Recognition* (2005).
5. Low, D. G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision,* 91–110 (2004).
6. Saitta, L. Support-Vector Networks. **297,** 273–297 (1995).
7. Breiman, L. Random forests. *Machine Learning* (2001).
8. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016).
9. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436–444 (2015).
10. Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics* **36,** 193–202 (1980).
11. LeCun, Y. *et al. Backpropagation applied to digit recognition* 1989.
12. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86,** 2278–2323 (1998).
13. Scherer, D., Müller, A. & Behnke, S. *Evaluation of pooling operations in convolutional architectures for object recognition* in *20th International Conference on Artificial Neural Networks (ICANN)* (2010), 92–101.

14. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* **15** (2014).
15. Chellapilla, K., Puri, S. & Simard, P. High Performance Convolutional Neural Networks for Document Processing. *Tenth International Workshop on Frontiers in Handwriting Recognition* (2006).
16. Krizhevsky, A., Sutskever, I. & Hinton, G. E. *ImageNet classification with deep convolutional neural networks* in *NeurIPS* (2012).
17. Cires, D. C., Meier, U., Masci, J. & Gambardella, L. M. *Flexible, High Performance Convolutional Neural Networks for Image Classification* in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence Flexible* (2013).
18. Xie, Q., Luong, M.-T., Hovy, E. & Le, Q. V. *Self-training with Noisy Student improves ImageNet classification* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020).
19. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. *You only look once: Unified, real-time object detection* in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (2016). ISBN: 9781467388504.
20. Schulter, S., Vernaza, P., Choi, W. & Chandraker, M. *Deep network flow for multi-object tracking* in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).
21. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Medical Image Analysis* **42,** 60–88 (2017).
22. Wan, Z. *et al. Bringing old photos back to life* in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (2020).
23. Bergmann, P., Fauser, P., Sattlegger, D. & Steger, C. *Uninformed Students : Student – Teacher Anomaly Detection with Discriminative Latent Embeddings* in *Proceedings of the IEEE/CVF Conference on Computer Visionand Pattern Recognition (CVPR)* (2020).
24. Dalzochio, J. *et al.* Machine learning and reasoning for predictive maintenance in Industry 4.0: Current status and challenges. *Computers in Industry* **123** (2020).

25. Lashin, V. & Rahtu, E. *Multi-modal dense video captioning* in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR)* (2020).
26. Radford, A., Metz, L. & Chintala, S. *Unsupervised representation learning with deep convolutional generative adversarial networks* in *International Conference on Learning Representations (ICLR)* (2016).
27. Sun, C., Shrivastava, A., Singh, S. & Gupta, A. *Revisiting unreasonable effectiveness of data in deep learning era* in *IEEE International Conference on Computer Vision (ICCV)* (2017).
28. Hestness, J. *et al.* Deep learning scaling is predictable, empirically. *ArXiv Preprint* (2017).
29. Cheplygina, V., de Bruijne, M. & Pluim, J. P. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis* **54,** 280–296 (2019).
30. Oquab, M., Bottou, L., Laptev, I. & Sivic, J. *Learning and transferring mid-level image representations using convolutional neural networks* in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (2014).
31. Shorten, C. & Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of Big Data* **6** (2019).
32. Henaff, O. J. *et al. Data-efficient image recognition with contrastive predictive coding* in *International Conference on Machine Learning (ICML)* (2020).
33. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. *A simple framework for contrastive learning of visual representations* in *37th International Conference on Machine Learning (ICML)* (2020).
34. Misra, I. & Maaten, L. V. D. *Self-supervised learning of pretext-invariant representations* in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (2020).
35. Donahue, J. & Simonyan, K. *Large scale adversarial representation Learning* in *Advances in Neural Information Processing Systems (NeurIPS)* (2019).
36. Oquab, M., Bottou, L., Laptev, I. & Sivic, J. *Is object localization for free? - Weakly-supervised learning with convolutional neural networks* in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015).

37. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K. & Wierstra, D. *Matching networks for one shot learning* in *Advances in Neural Information Processing Systems (NeurIPS)* (2016).
38. Snell, J., Swersky, K. & Zemel, R. *Prototypical networks for few-shot learning* in *Advances in Neural Information Processing Systems (NeurIPS)* (2017).
39. Chang, Y. T. *et al.* Weakly-supervised semantic segmentation via sub-category exploration. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* 8988–8997. ISSN: 10636919 (2020).
40. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* **128,** 336–359 (2020).
41. Patel, G. & Dolz, J. Weakly supervised segmentation with cross-modality equivariant constraints. *Medical Image Analysis* **77** (2021).
42. Wang, Y., Zhang, J., Kan, M., Shan, S. & Chen, X. *Self-Supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation* in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (2020).
43. Durand, T., Mordan, T., Thome, N. & Cord, M. *WILDCAT: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation* in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).
44. Chan, L., Hosseini, M. S., Rowsell, C., Plataniotis, K. N. & Damaskinos, S. *HistoSegNet : Semantic Segmentation of Histological Tissue Type in Whole Slide Images* in *IEEE/CVF International Conference on Computer Vision (ICCV)* (2019).
45. Pinheiro, P. O. & Collobert, R. *From image-level to pixel-level labeling with Convolutional Networks* in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (2015).
46. Ilse, M., Tomczak, J. M. & Welling, M. *Attention-based deep multiple instance learning* in *35th International Conference on Machine Learning (ICML)* (2018).
47. Campanella, G. *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine* **25,** 1301–1309 (2019).

48. Kervadec, H. *et al.* Constrained Deep Networks: Lagrangian Optimization via Log-Barrier Extensions. *ArXiv Preprint* (2019).
49. Jia, Z., Huang, X., Chang, E. I. & Xu, Y. Constrained Deep Weak Supervision for Histopathology Image Segmentation. *IEEE Transactions on Medical Imaging* **36,** 2376–2388 (2017).
50. Sung, F. *et al. Learning to Compare: Relation Network for Few-Shot Learning* in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).
51. Chen, W. Y., Wang, Y. C. F., Liu, Y. C., Kira, Z. & Huang, J. B. *A closer look at few-shot classification* in *7th International Conference on Learning Representations (ICLR)* (2019).
52. Laenen, S. & Bertinetto, L. *On episodes,prototypical networks, and few-shot learning* in *Advances in Neural Information Processing Systems (NeurIPS) meta-learning workshop* (2020).
53. Khosla, P. *et al. Supervised Contrastive Learning* in *Advances in Neural Information Processing Systems (NeurIPS)* (2020).
54. Baur, C., Wiestler, B., Albarqouni, S. & Navab, N. *Deep autoencoding models for unsupervised anomaly segmentation in brain MR images* in *Medical Image Computing and Computer Assisted Intervention (MICCAI)* (2019).
55. Bergmann, P., Löwe, S., Fauser, M., Sattlegger, D. & Steger, C. *Improving unsupervised defect segmentation by applying structural similarity to autoencoders* in *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)* (2019).
56. Abati, D., Porrello, A., Calderara, S. & Cucchiara, R. *Latent space autoregression for novelty detection* in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).
57. Kingma, D. P. & Welling, M. *Auto-encoding variational bayes* in *2nd International Conference on Learning Representations, (ICLR)* (2014).
58. Chen, X. & Konukoglu, E. *Unsupervised Detection of Lesions in Brain MRI using constrained adversarial auto-encoders* in *Medical Imaging with Deep Learning (MIDL)* (2018).
59. Schlegl, T., Seeböck, P., Waldstein, S. M., Langs, G. & Schmidt-Erfurth, U. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis* **54,** 30–44 (2019).

60. Venkataramanan, S., Peng, K. C., Singh, R. V. & Mahalanobis, A. *Attention Guided Anomaly Localization in Images* in *Proceedings of the European COnference on Computer Vision (ECCV)* (2020).
61. Liu, W. *et al. Towards Visually Explaining Variational Autoencoders* in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (2020).
62. Silva-rodríguez, J., Colomer, A., Sales, M. A., Molina, R. & Naranjo, V. 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* **195** (2020).
63. Silva-Rodríguez, J., Colomer, A. & Naranjo, V. WeGleNet: A weakly-supervised convolutional neural network for the semantic segmentation of Gleason grades in prostate histology images. *Computerized Medical Imaging and Graphics* **88** (2021).
64. Silva-Rodriguez, J., Colomer, A., Dolz, J. & Naranjo, V. Self-learning for weakly supervised Gleason grading of local patterns. *IEEE Journal of Biomedical and Health Informatics* **25,** 3094–3104 (2021).
65. World Cancer Research Foundation. *Prostate Cancer Statistics* 2019.
66. World Health Organization. *Global Cancer Observatory* 2019.
67. Gleason, D. F. *Histologic grading of prostate cancer: A perspective, human pathology* (1992).
68. Gordetsky, J. & Epstein, J. Grading of prostatic adenocarcinoma: Current state and prognostic implications. *Diagnostic Pathology* **11,** 2–9 (2016).
69. Epstein, J. I. *et al.* 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* **40,** 244–252 (2016).
70. Sharma, M. & Miyamoto, H. Percent Gleason pattern 4 in stratifying the prognosis of patients with intermediate-risk prostate cancer. *Translational Andrology and Urology* **7,** S484–S489 (2018).
71. Hassan, O. & Matoso, A. Clinical significance of subtypes of Gleason pattern 4 prostate cancer. *Translational Andrology and Urology* **7,** S477–S483 (2018).
72. Van der Kwast, T. & van Leenders, G. J. On cribriform prostate cancer. *Translational Andrology and Urology* **7,** 145–154 (2018).
73. Remotti, H. in *Methods in Molecular Biology* (2013).

74. Khouja, M. H., Baekelandt, M., Sarab, A., Nesland, J. M. & Holm, R. Limitations of tissue microarrays compared with whole tissue sections in survival analysis. *Oncology letters* **1,** 827–831 (2010).
75. Voduc, D., Kenney, C. & O. Nielsen, T. Tissue Microarrays in Clinical Oncology. *Bone* **18,** 88–97 (2009).
76. Doyle, S. *et al. Automated grading of prostate cancer using architectural and textural image features* in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro* (2007).
77. Gertych, A. *et al.* Machine learning approaches to analyze histological images of tissues from radical prostatectomies. *Computerized Medical Imaging and Graphics* **46,** 197–208 (2015).
78. Jiménez del Toro, O. *et al. Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade Gleason score* in *SPIE Medical Imaging* (2017).
79. Ren, J., Sadimin, E., Foran, D. J. & Qi, X. *Computer aided analysis of prostate histopathology images to support a refined Gleason grading system* in *SPIE Medical Imaging* (2017).
80. Ing, N. *et al. Semantic segmentation for prostate cancer grading by convolutional neural networks* in *SPIE Medical Imaging* (2018).
81. Esteban, A. E. *et al.* A new optical density granulometry-based descriptor for the classification of prostate histological images using shallow and deep Gaussian processes. *Computer Methods and Programs in Biomedicine* **178,** 303–317 (2019).
82. Lucas, M. *et al.* Deep learning for automatic Gleason pattern classification for grade group determination of prostate biopsies. *Virchows Archiv* **475,** 77–83 (2019).
83. Arvaniti, E. *et al.* Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *Scientific Reports* **8,** 1–11 (2018).
84. Nir, G. *et al.* Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. *Medical Image Analysis* **50,** 167–180 (2018).
85. Nir, G. *et al.* Comparison of Artificial Intelligence Techniques to Evaluate Performance of a Classifier for Automatic Grading of Prostate Cancer From Digitized Histopathologic Images. *JAMA network open* **2** (2019).
86. García, G., Colomer, A. & Naranjo, V. First-stage prostate cancer identification on histopathological images: Hand-driven versus automatic learning. *Entropy* **21** (2019).

87. Ma, Y. *et al.* Generating region proposals for histopathological whole slide image retrieval. *Computer Methods and Programs in Biomedicine* **159,** 1–10 (2018).
88. Li, W. *et al.* Path R-CNN for prostatecCancer diagnosis and Gleason grading of histological images. *IEEE Transactions on Medical Imaging* **38,** 945–954 (2019).
89. Lin, M., Chen, Q. & Yan, S. *Network In Network* in *International Conference of Learning Representations (ICLR)* (2014), 1–10.
90. Simonyan, K. & Zisserman, A. *Very Deep Convolutional Networks for Large-Scale Image Recognition* in *International Conference on Learning Representations (ICLR)* (2014).
91. He, K., Zhang, X., Ren, S. & Sun, J. *Deep residual learning for image recognition* in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).
92. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. *Rethinking the inception architecture for computer vision* in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).
93. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. *Densely connected convolutional networks* in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).
94. Deng, J. *et al. ImageNet: A Large-Scale Hierarchical Image Database* in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (2009).
95. Weintein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer Analysis Project. *NIH Public Access* **4518,** 219–223 (2013).
96. *OpenSeadragon*
97. Cohen, J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* **70,** 213–220 (1968).
98. McHugh, M. L. Interrater reliability: the kappa statistic. *Lessons in biostatistics* **22,** 276–282 (2012).
99. Swets, J. A. Measuring the Accuracy of Diagnostic Systems. *Science* **240,** 1285–1293 (1988).
100. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. *Learning Deep Features for Discriminative Localization* in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (2016).

101. Vahadane, A. *et al. Structure-preserved color normalization for histological images* in *Proceedings of the International Symposium on Biomedical Imaging (ISBI)* (2015).

102. Kweldam, C. F. *et al.* Gleason grade 4 prostate adenocarcinoma patterns: an interobserver agreement study among genitourinary pathologists. *Histopathology* **69,** 441–449 (2016).

103. Burchardt, M. *et al.* Interobserver reproducibility of Gleason grading: Evaluation using prostate cancer tissue microarrays. *Journal of Cancer Research and Clinical Oncology* **134,** 1071–1078 (2008).

104. Komura, D. & Ishikawa, S. Machine Learning Methods for Histopathological Image Analysis. *Computational and Structural Biotechnology Journal* **16,** 34–42 (2018).

105. Carbonneau, M. A., Cheplygina, V., Granger, E. & Gagnon, G. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition* **77,** 329–353 (2018).

106. Lee, J., Kim, E., Lee, S., Lee, J. & Yoon, S. *Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference* in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).

107. Ahn, J., Cho, S. & Kwak, S. *Weakly supervised learning of instance segmentation with inter-pixel relations* in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).

108. Papandreou, G., Chen, L. C., Murphy, K. P. & Yuille, A. L. *Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation* in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2015).

109. Bency, A. J., Kwon, H., Lee, H., Karthikeyan, S. & Manjunath, B. S. *Weakly supervised localization using deep feature maps* in *Proceedings of the European Conference on Computer Vision (ECCV)* (2016). ISBN: 9783319464473.

110. Krähenbühl, P. & Koltun, V. *Efficient inference in fully connected crfs with Gaussian edge potentials* in *Advances in Neural Information Processing Systems (NeurIPS)* (2011).

111. Chan, L., Hosseini, M. S. & Plataniotis, K. N. A Comprehensive Analysis of Weakly-Supervised Semantic Segmentation in Different Image Domains. *International Journal of Computer Vision* **1** (2020).

112. Courtiol, P., Tramel, E. W., Sanselme, M. & Wainrib, G. Classification and Disease Localization in histopathology Using Only Global Labels: a Weakly Supervised Approach. *ArXiv,* 1–13 (2017).

113. Arvaniti, E. & Claassen, M. Coupling weak and strong supervision for classification of prostate cancer histopathology images. *ArXiv* (2018).

114. Wang, X. *et al.* Weakly Supervised Deep Learning for Whole Slide Lung Cancer Image Analysis. *IEEE Transactions on Cybernetics,* 1–13 (2019).

115. Li, J., Speier, W., Ho, K. C. & Sarma, K. V. An EM-based semi-supervised deep learning approach for semantic segmentation of histopathological images from radical prostatectomies. *Computerized Medical Imaging and Graphics* **69,** 125–133 (2016).

116. Xu, Y., Zhu, J. Y., Chang, E. I., Lai, M. & Tu, Z. Weakly supervised histopathology cancer image segmentation and classification. *Medical Image Analysis* **18,** 591–604 (2014).

117. Bulten, W. *et al.* Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *The Lancet Oncology* **21,** 233–241 (2020).

118. Otálora, S., Atzori, M., Andrearczyk, V., Khan, A. & Müller, H. Staining invariant features for improving generalization of deep convolutional neural networks in computational pathology. *Frontiers in Bioengineering and Biotechnology* **7,** 1–13 (2019).

119. Ström, P. *et al.* Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *The Lancet Oncology* **21,** 222–232 (2020).

120. Singh, K. K. & Lee, Y. J. *Hide-and-Seek: Forcing a Network to be Meticulous for Weakly-Supervised Object and Action Localization* in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2017).

121. Ronneberger, O., Fischer, P. & Brox, T. *U-net: Convolutional networks for biomedical image segmentation* in *Medical Image Computing and Computer Assisted Intervention (MICCAI)* (2015).

122. Zhang, Z., Liu, Q. & Wang, Y. Road extraction by deep residual U-Net. *IEEE Geoscience and Remote Sensing Letters* **15,** 749–753 (2018).

123. Kalapahar, A., Silva-Rodríguez, J., Colomer, A., López-Mir, F. & Naranjo, V. *Gleason Grading of Histology Prostate Images through Semantic Segmentation via Residual U-Net* in *Proceedings of the IEEE International Conference on Image Processing (ICIP)* (2020).

124. Epstein, J. I. A new contemporary prostate cancer grading system. *Annales de Pathologie* **35,** 474–476 (2015).

125. Scudder, H. J. Probability of Error of Some Adaptive Pattern-Recognition Machines. *IEEE Transactions on Information Theory* **11,** 363–371 (1965).

126. Yalniz, I. Z., Jégou, H., Chen, K., Paluri, M. & Mahajan, D. *Billion-scale semi-supervised learning for image classification* in *ArXiv Preprint* (2019).

127. Veit, A. *et al. Learning from noisy large-scale datasets with minimal supervision* in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).

128. Bazzani, L., Bergamo, A., Anguelov, D. & Torresani, L. Self-taught object localization with deep networks. *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)* (2016).

129. Sangineto, E., Nabi, M., Culibrk, D. & Sebe, N. Self paced deep learning for weakly supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41,** 712–725 (2019).

130. Jie, Z., Wei, Y., Jin, X., Feng, J. & Liu, W. *Deep self-taught learning for weakly supervised object localization* in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).

131. Chen, T., Kornblith, S., Swersky, K., Norouzi, M. & Hinton, G. *Big self-supervised models are strong semi-supervised learners* in *Advances in Neural Information Processing Systems (NeurIPS)* (2020).

132. Patacchiola, M. & Storkey, A. *Self-Supervised Relational Reasoning for Representation Learning* in *Advances in Neural Information Processing Systems (NeurIPS)* (2020).

133. Nagpal, K. *et al.* Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *npj Digital Medicine* **2** (2019).

134. Bulten, W. *et al.* Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nature Medicine* (2022).

135. Zhou, Z. H., Sun, Y. Y. & Li, Y. F. *Multi-instance learning by treating instances as non-I.I.D. samples* in *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)* (2009).

136. Otálora, S. *et al.* A systematic comparison of deep learning strategies for weakly supervised Gleason grading, 20 (2020).

137. Li, J. *et al.* A multi-resolution model for histopathology image classification and localization with multiple instance learning. *Computers in Biology and Medicine* **131** (2021).

138. Srinidhi, C. L., Ciga, O. & Martel, A. L. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis* **67** (2021).

139. Zhou, Z. H. & Zhang, M. L. *Multi-instance multi-label learning with application to scene classification* in *Advances in Neural Information Processing Systems (NeurIPS)* (2007).

140. Jia, Z., Huang, X., Chang, E. I. & Xu, Y. Constrained Deep Weak Supervision for Histopathology Image Segmentation. *IEEE Transactions on Medical Imaging* **36,** 2376–2388 (2017).

141. Zhou, Y. *et al. Prior-aware neural network for partially-supervised multi-organ segmentation* in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2019).

142. Kervadec, H., Dolz, J., Granger, E. & Ben Ayed, I. *Curriculum Semi-supervised Segmentation* in *Medical Image Computing and Computer Assisted Intervention (MICCAI)* (2019).

143. Bateson, M., Dolz, J., Kervadec, H., Lombaert, H. & Ayed, I. B. Constrained Domain Adaptation for Image Segmentation. *IEEE Transactions on Medical Imaging* **40,** 1875–1887 (2021).

144. Otálora, S., Marini, N., Müller, H. & Atzori, M. *Semi-weakly Supervised Learning for Prostate Cancer Image Classification with Teacher-Student Deep Convolutional Networks* in *Interpretable and Annotation-Efficient Learning for Medical Image Computing. IMIMIC 2020, MIL3ID 2020, LABELS 2020.* (2020).

145. Van Der Maaten, L. & Geoffrey, H. Visualizing data using t-SNE. *Journal of Machine Learning Research* **9,** 2579–2605 (2008).

146. Tzanakakis, K. *The railway track and its log term behaviour* 279–292 (2013).

147. Lidén, T. Railway infrastructure maintenance - A survey of planning problems and conducted research. *Transportation Research Procedia* **10,** 574–583 (2015).

148. Kouroussis, G. *et al.* Review of trackside monitoring solutions: From strain gages to optical fibre sensors. *Sensors* **15** (2015).

149. Salvador, P., Naranjo, V., Insa, R. & Teixeira, P. Axlebox accelerations: Their acquisition and time-frequency characterisation for railway track monitoring purposes. *Measurement: Journal of the International Measurement Confederation* **82,** 301–312 (2016).

150. Chia, L., Bhardwaj, B., Lu, P., Bridgelall, R. & Member, S. Railroad track condition monitoring using inertial sensors and digital signal processing: A review. *IEEE Sensors Journal* **19,** 25–33 (2019).

151. Jing, L., Wang, K. & Zhai, W. Impact vibration behavior of railway vehicles: a state-of-the-art overview. *Acta Mechanica Sinica* **77** (2021).

152. Nadarajah, N., Shamdani, A., Hardie, G., Chiu, W. K. & Widyastuti, H. Prediction of railway vehicles' dynamic behavior with machine learning algorithms. *Electronic Journal of Structural Engineering* **18,** 38–46 (2018).

153. Chellaswamy, C., Santhi, P. & K, V. Deep learning based intelligent rail track health monitoring system. *International Journal of Innovative Technology and Exploring Engineering* **9,** 5111–5122 (2019).

154. Niebling, J., Baasch, B. & Kruspe, A. *Analysis of railway track irregularities with convolutional autoencoders and clustering algorithms* in *Communications in Computer and Information Science* (2020).

155. Yang, C., Sun, Y., Ladubec, C. & Liu, Y. Developing machine learning-based models for railway inspection. *Applied Sciences* **11,** 1–15 (2021).

156. Chenariyan Nakhaee, M., Hiemstra, D., Stoelinga, M. & van Noort, M. *The recent applications of machine learning in rail track maintenance: A survey* in *Proceedings of the International Conference on Reliability, Safety, and Security of Railway Systems (RSSRail)* **11495 LNCS** (2019), 91–105.

157. Bosso, N., Gugliotta, A. & Zampieri, N. Design and testing of an innovative monitoring system for railway vehicles. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit* **232,** 445–460 (2018).

158. Zhang, X., Wang, K., Wang, Y., Shen, Y. & Hu, H. *An improved method of rail health monitoring based on CNN and multiple acoustic emission events* in *Proceedings of the IEEE International Instrumentation and Measurement Technology Conference (I2MTC)* (2017).

159. Chen, S. X., Zhou, L., Ni, Y. Q. & Liu, X. Z. An acoustic-homologous transfer learning approach for acoustic emission–based rail condition evaluation. *Structural Health Monitoring* **20,** 2161–2181 (2021).

160. Giben, X., Patel, V. M. & Chellappa, R. *Material classification and semantic segmentation of railway track images with deep convolutional neural networks* in *Proceedings of the IEEE International Conference on Image Processing (ICIP)* (2015).

161. Faghih-Roohi, S. *et al. Deep convolutional neural networks for detection of rail surface defects* in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)* (2016).

162. Mittal, S. & Rao, D. Vision based railway track monitoring using deep learning. *ArXiv* (2017).

163. Gibert, X., Patel, V. M. & Chellappa, R. Deep multitask learning for railway track inspection. *IEEE Transactions on Intelligent Transportation Systems* **18,** 153–164 (2017).

164. Zhang, H., Jin, X., Wu, J. Q., He, Z. & Wang, Y. Automatic visual detection method of railway surface defects based on curvature filtering and improved GMM. *Chinese Journal of Scientific Instrument* **39,** 181–194 (2018).

165. Wang, L., Zhuang, L. & Zhang, Z. Automatic detection of rail surface cracks with a superpixel-based data-driven framework. *Journal of Computing in Civil Engineering* **33** (2019).

166. James, A. *et al. TrackNet - A Deep Learning Based Fault Detection for Railway Track Inspection* in *International Conference on Intelligent Rail Transportation (ICIRT)* (2019).

167. Hovad, E. *et al.* in *Intelligent Quality Assessment of Railway Switches and Crossings* 207–228 (2021).

168. Hory, C., Bouillaut, L. & Aknin, P. Time-frequency characterization of rail corrugation under a combined auto-regressive and matched filter scheme. *Mechanical Systems and Signal Processing* **29,** 174–186 (2012).

169. Molodova, M., Li, Z. & Dollevoet, R. Axle box acceleration: Measurement and simulation for detection of short track defects. *Wear* **271,** 349–356 (2011).

170. Wei, Z., Núñez, A., Li, Z. & Dollevoet, R. Evaluating degradation at railway crossings using axle box acceleration measurements. *Sensors* **17** (2017).

171. Bosso, N., Gugliotta, A. & Zampieri, N. Wheel flat detection algorithm for onboard diagnostic. *Measurement: Journal of the International Measurement Confederation* **123,** 193–202 (2018).

172. Boogaard, M. A., Li, Z. & Dollevoet, R. P. In situ measurements of the crossing vibrations of a railway turnout. *Measurement: Journal of the International Measurement Confederation* **125,** 313–324 (2018).

173. Bocz, P., Vinkó, Á. & Posgay, Z. A practical approach to tramway track condition monitoring: vertical track defects detection and identification using time-frequency processing technique. *Selected Scientific Papers - Journal of Civil Engineering* **13,** 135–146 (2018).

174. Sysyn, M., Gerber, U., Nabochenko, O., Li, Y. & Kovalchuk, V. Indicators for common crossing structural health monitoring with track-side inertial measurements. *Acta Polytechnica* **59,** 170–181 (2019).

175. Sysyn, M., Gruen, D., Gerber, U., Nabochenko, O. & Kovalchuk, V. Turnout monitoring with vehicle based inertial measurements of operational trains: A machine learning approach. *Communications - Scientific Letters of the University of Zilina* **21,** 42–48 (2019).

176. Li, J. & Shi, H. Rail corrugation detection of high-speed railway using wheel dynamic responses. *Shock and Vibration* (2019).

177. Ng, A. K., Martua, L. & Sun, G. *Dynamic modelling and acceleration signal analysis of rail surface defects for enhanced rail condition monitoring and diagnosis* in *Proceedings of the International Conference on Intelligent Transportation Engineering (ICITE)* (2019).

178. Carrigan, T. D., Fidler, P. R. & Talbot, J. P. *On the derivation of rail roughness spectra from axle-box vibration: Development of a new technique* in *Proceedings of the International Conference on Smart Infrastructure and Construction (ICSIC)* (2019).

179. Carrigan, T. D. & Talbot, J. P. Extracting Information from Axle-Box Acceleration: On the Derivation of Rail Roughness Spectra in the Presence of Wheel Roughness. *Notes on Numerical Fluid Mechanics and Multidisciplinary Design* **150,** 286–294 (2021).

180. Malekjafarian, A., OBrien, E., Quirke, P. & Bowe, C. Railway track monitoring using train measurements: An experimental case study. *Applied Sciences* **9** (2019).

181. Malekjafarian, A., Obrien, E. J., Quirke, P., Cantero, D. & Golpayegani, F. Railway track loss-of-stiffness detection using bogie filtered displacement data measured on a passing train. *Infrastructures* **6,** 1–17 (2021).

182. Baasch, B., Roth, M., Havrila, P. & Groos, J. C. *Detecting singular track defects by time-frequency signal separation of axle-box acceleration data* in *Proceedings of the World Congress on Railway Research (WCRR)* (2019).

183. Song, Y., Liang, L., Du, Y. & Sun, B. Railway polygonized wheel detection based on numerical time-frequency analysis of axle-box acceleration. *Applied Sciences* **10** (2020).

184. He, X., Yu, K., Cai, C. & Zou, Y. Dynamic responses of the metro train's bogie frames: Field tests and data analysis. *Shock and Vibration* (2020).

185. Ghosh, C., Verma, A. & Verma, P. Real time fault detection in railway tracks using Fast Fourier Transformation and Discrete Wavelet Transformation. *International Journal of Information Technology* (2021).

186. Chang, C. *et al.* Dynamic performance evaluation of an inspection wagon for urban railway tracks. *Measurement: Journal of the International Measurement Confederation* **170** (2021).

187. Chellaswamy, C., Krishnasamy, M., Balaji, L., Dhanalakshmi, A. & Ramesh, R. Optimized railway track health monitoring system based on dynamic differential evolution algorithm. *Measurement: Journal of the International Measurement Confederation* **152** (2020).

188. Heusel, J. *et al. Detecting corrugation defects in harbour railway networks using axle box acceleration data* in *International Conference on Condition Monitoring and Asset Management* (2021).

189. Sresakoolchai, J. & Kaewunruen, S. Detection and Severity Evaluation of Combined Rail Defects Using Deep Learning. *Vibration* **4,** 341–356 (2021).

190. Salvador, P., Villalba, I., Martínez-fernández, P. & Insa, R. *Application of time-frequency representations for the detection of railway track singularities* in *Proceedings of the Joint International Conference on Multibody System Dynamics* (2018).

191. Sresakoolchai, J. & Kaewunruen, S. Wheel flat detection and severity classification using deep learning techniques. *Insight-Non-Destructive Testing and Condition Monitoring* **63,** 393–402. ISSN: 17544904 (2021).

192. Zhang, B., Li, X., Ye, Y., Huang, Z. & Zhang, L. *Prototype Completion with Primitive Knowledge for Few-Shot Learning* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* (2020).

193. García, G. *et al.* Circumpapillary OCT-focused hybrid learning for glaucoma grading using tailored prototypical neural networks. *Artificial Intelligence in Medicine* **118** (2021).

194. Barredo Arrieta, A. *et al.* Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* **58,** 82–115 (2020).

195. Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U. & Langs, G. *Unsupervised anomaly detection with generative adversarial networks to guide marker discovery* in *Proceedings of the International Conference on Information Processing in Medical Imaging (IPMI)* (2017).
196. Andermatt, S., Horváth, A., Pezold, S. & Cattin, P. Pathology segmentation using distributional differences to images of healthy origin. *Medical Image Computing and Computer Assisted Intervention (MICCAI) - Brainlesion Workshop* (2019).
197. Ravanbakhsh, M., Sangineto, E., Nabi, M. & Sebe, N. Training adversarial discriminators for cross-channel abnormal event detection in crowds. *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)* **2** (2019).
198. Baur, C., Graf, R., Wiestler, B., Albarqouni, S. & Navab, N. *SteGANomaly: Inhibiting CycleGAN Steganography for Unsupervised Anomaly Detection in Brain MRI* in *Medical Image Computing and Computer Assisted Intervention (MICCAI)* (Springer International Publishing, 2020).
199. Sun, L. *et al.* An adversarial learning approach to medical image synthesis for lesion detection. *IEEE Journal of Biomedical and Health Informatics* **24,** 2303–2314 (2020).
200. Nick Pawlowski, M. C. L. *Unsupervised Lesion Detection in Brain CT using Bayesian Convolutional Autoencoders* in *Medical Imaging with Deep Learning (MIDL)* (2018).
201. Sabokrou, M. *et al. AVID: Adversarial Visual Irregularity Detection* in *Proceedings of the Asia Conference on Computer Vision (ACCV)* (2019).
202. Chen, X., You, S., Tezcan, K. C. & Konukoglu, E. Unsupervised lesion detection via image restoration with a normative prior. *Medical Image Analysis* **64** (2020).
203. Zimmerer, D., Isensee, F., Petersen, J., Kohl, S. & Maier-Hein, K. *Abstract: Unsupervised anomaly localization using variational autoencoders* in *Informatik aktuell* (2020).
204. Silva-Rodríguez, J., Naranjo, V. & Dolz, J. *Looking at the whole picture: constrained unsupervised anomaly segmentation* in *British Machine Vision Conference (BMVC)* (2021).
205. Goodfellow, I. *et al. Generative adversarial networks* in *Advances in Neural Information Processing Systems (NeurIPS)* **63** (2014).

206. Dehaene, D., Frigo, O., Combrexelle, S. & Eline, P. *Iterative energy-based projection on a normal data manifold for anomaly localization* in *Proceedings of the International Conference on Learning Representations (ICLR)* (2020).

207. Shi, Y., Yang, J. & Qi, Z. Unsupervised anomaly segmentation via deep feature reconstruction. *Neurocomputing* **424,** 9–22 (2021).

208. Zimmerer, D., Kohl, S., Petersen, J., Isensee, F. & Maier-Hein, K. *Context-encoding Variational Autoencoder for Unsupervised Anomaly Detection* in *Proceedings og the International Conference on Medical Imaging with Deep Learning (MIDL)* (2019).

209. Nguyen, B., Bethapudi, A., Jennings, A. & Willdocks, C. G. *Unsupervised region-based anomaly detection in brain MRI with adversarial image impainting* in *Proceedings of the IEEE International Symposium on-Biomedical Imaging (ISBI)* (2021).

210. Baur, C., Denner, S., Wiestler, B., Navab, N. & Albarqouni, S. Autoencoders for unsupervised anomaly segmentation in brain MR images: A comparative study. *Medical Image Analysis* **69,** 1–16 (2021).

211. Zimmerer, D. *et al.* MOOD 2020: A public Benchmark for Out-of-Distribution Detection and Localization on medical Images. *IEEE Transactions on Medical Imaging,* 1–11. ISSN: 1558254X (2022).

212. Meissen, F., Wiestler, B., Kaissis, G. & Rueckert, D. On the Pitfalls of Using the Residual Error as Anomaly Score. **1,** 1–15 (2022).

213. Zhang, Y., David, P. & Gong, B. *Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes* in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2017).

214. Pathak, D., Krahenbuhl, P. & Darrell, T. Constrained convolutional neural networks for weakly supervised segmentation. *Proceedings of the IEEE International Conference on Computer Vision (ICCV),* 1796–1804 (2015).

215. Peng, J. *et al.* Discretely-constrained deep network for weakly supervised segmentation. *Neural Networks* **130,** 297–308 (2020).

216. Boyd, S. & Vandenberghe, L. *Convex optimization* (Cambridge university press, 2004).

217. Kervadec, H. *et al.* Constrained-CNN losses for weakly supervised segmentation. *Medical Image Analysis* **54,** 88–99 (2019).

218. He, F. S., Liu, Y., Schwing, A. G. & Peng, J. *Learning to play in a day: Faster deep reinforcement learning by optimality tightening* in *Proceedings of the International Conference on Learning Representations (ICLR)* (2017), 1–13.

219. Fiacco, A. V. & McCormick, G. P. *Nonlinear programming: sequential unconstrained minimization techniques* (SIAM, 1990).

220. Luenberger, D. G. *Introduction to linear and nonlinear programming* (Addison-wesley Reading, MA, 1973).

221. Menze, B. H. *et al.* The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging* **34,** 1993–2024 (2015).

222. Bakas, S. *et al.* Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data* **4,** 1–13 (2017).

223. Bakas, S. *et al.* Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. *ArXiv Preprint* (2018).

224. Hssayeni, M. Computed Tomography Images for Intracranial Hemorrhage Detection and Segmentation. *PhysioNet* (2020).

225. Hssayeni, M. D. *et al.* Intracranial hemorrhage segmentation using a deep convolutional model. *Data* **5,** 1–18 (2020).

226. Goldberger, A. L. *et al.* PhysioBank, PhysionToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **101** (2000).

227. Meissen, F., Kaissis, G. & Rueckert, D. Challenging Current Semi-Supervised Anomaly Segmentation Methods for Brain MRI, 1–13 (2021).