Medical Image Analysis 95 (2024) 103191

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

A systematic comparison of deep learning methods for Gleason grading and scoring

Juan P. Dominguez-Morales [a,b,*,1], Lourdes Duran-Lopez [a,b,1], Niccolò Marini [c,d,1], Saturnino Vicente-Diaz [a,b], Alejandro Linares-Barranco [a,b], Manfredo Atzori [c,e], Henning Müller [c,f]

[a] Robotics and Technology of Computers Lab., ETSII-EPS, Universidad de Sevilla, Sevilla 41012, Spain
[b] SCORE Lab, I3US. Universidad de Sevilla, Spain
[c] Information Systems Institute, University of Applied Sciences Western Switzerland (HES-SO Valais), Technopôle 3, Sierre 3960, Switzerland
[d] Centre Universitaire d'Informatique, University of Geneva, Carouge 1227, Switzerland
[e] Department of Neuroscience, University of Padua, Via Giustiniani 2, Padua, 35128, Italy
[f] Medical faculty, University of Geneva, Geneva 1211, Switzerland

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
Computational pathology
Deep learning
Prostate cancer
Multiple-instance learning
Weak supervision
Full supervision
Semi-supervision

ABSTRACT

Prostate cancer is the second most frequent cancer in men worldwide after lung cancer. Its diagnosis is based on the identification of the Gleason score that evaluates the abnormality of cells in glands through the analysis of the different Gleason patterns within tissue samples. The recent advancements in computational pathology, a domain aiming at developing algorithms to automatically analyze digitized histopathology images, lead to a large variety and availability of datasets and algorithms for Gleason grading and scoring. However, there is no clear consensus on which methods are best suited for each problem in relation to the characteristics of data and labels. This paper provides a systematic comparison on nine datasets with state-of-the-art training approaches for deep neural networks (including fully-supervised learning, weakly-supervised learning, semi-supervised learning, Additive-MIL, Attention-Based MIL, Dual-Stream MIL, TransMIL and CLAM) applied to Gleason grading and scoring tasks. The nine datasets are collected from pathology institutes and openly accessible repositories.

The results show that the best methods for Gleason grading and Gleason scoring tasks are fully supervised learning and CLAM, respectively, guiding researchers to the best practice to adopt depending on the task to solve and the labels that are available.

1. Introduction

Histopathology has experienced a digital transformation over the past ten years, slowly moving away from the traditional workflow with microscopes and adopting a computerized approach. Information and communication technologies (ICT) have changed the way pathology is developing (Van der Laak et al., 2021). Advancements in slide image acquisition technology, software applications and high-speed networks allow integrating digital pathology into traditional workflow pipelines (Pallua et al., 2020). Digital pathology involves the acquisition, management, exchange and interpretation of pathology information, including images and pathology data in a digital environment (Niazi et al., 2019; Pallua et al., 2020). Digital slides are created during tissue sample acquisition, using devices called whole slide scanners, in order to obtain high-resolution digital images that can be digitally analyzed by medical experts or software tools.

Digitized slides enable the application of automatic algorithms and Artificial Intelligence (AI), as support tools for diagnosis, paving the road to computational pathology (Abels et al., 2019). In this regard, Computer-Aided Diagnosis (CAD) emerges with the purpose of assisting physicians in the interpretation of medical images, providing a second opinion to support the diagnosis process. The development of computational pathology CAD systems has become an important and challenging research topic, as it can lead to reduced diagnosis times and complement experts' decisions (Doi, 2007), opening unprecedented opportunities in healthcare and related markets.

AI, and, in particular, Deep Learning (DL) algorithms, have grown in popularity within the field of biomedical image analysis (Altaf et al., 2019; Razzak et al., 2018; Santos et al., 2019). These algorithms are able to learn relevant patterns from input images, using this information to perform a computer-based diagnosis. CNNs, which are currently

* Corresponding author at: Robotics and Technology of Computers Lab., ETSII-EPS, Universidad de Sevilla, Sevilla 41012, Spain.
E-mail address: jpdominguez@us.es (J.P. Dominguez-Morales).
[1] The first three authors have contributed equally to this paper.

https://doi.org/10.1016/j.media.2024.103191
Received 10 May 2022; Received in revised form 16 January 2024; Accepted 2 May 2024
Available online 4 May 2024

among the most popular neural networks in DL, are widely used for image analysis in several fields (Li et al., 2020), including biomedical image analysis (Anwar et al., 2018).

WSIs are high-resolution images, usually scanned with a spatial resolution in the order of μm per pixel or less (Sellaro et al., 2013). Modern scanners usually acquire the image with a spatial resolution of 0.23–0.25 μm, corresponding to an optical resolution (i.e., the magnification factor (×) of the lens used in the whole slide scanner (Sellaro et al., 2013)) of 40×. High-resolution leads to large images in terms of pixels, usually up to 200'000 × 200'000. Modern hardware barely manages to deal with this size of input data, forcing the splitting of the WSIs into small subimages, called patches.

CNNs require a large amount of labeled data during the network training in order to achieve robust results and to generalize on new unseen data (Madabhushi and Lee, 2016). In particular, fully-supervised algorithms, which show the highest performance in several tasks, such as classification, require pixel-level (or patch-level) annotations. The need of locally-labeled data is often a challenge in the field of computational pathology, since examining and annotating regions of interest in a large number of very large images is a time-consuming and expensive task for pathologists (Krupinski et al., 2013). As a consequence of the annotation process, which involves the participation of experienced pathologists, few datasets with patch-level annotations are publicly available. Although locally-labeled data are not always meaningful and needed, their lack can be partially alleviated using global image-level annotations. Global (or image-level) annotations include labels involving the whole image without identifying the regions of interest within it (Deng et al., 2020). Global annotations are much easier to collect, although they provide less information than patch-level labels.

Another aspect to deal with when working in histology is the heterogeneity of the data. The tissue samples obtained from a biopsy are processed in a laboratory using certain stains to enhance the contrast of biological structures. One of the most common staining methods used in diagnostic medicine is hematoxylin and eosin (H&E) stain Chan (2014). The lack of standardization of the staining procedure leads to color variations even from the same source (Marini et al., 2021a). On the other hand, the scanning device used to digitize samples also has a direct impact on the image color and texture, which often implies that these color differences become more evident when the images originate from different hospitals. Consequently, algorithms trained on a dataset from one source usually show a decrease in accuracy when tested on data from other sources (Ström et al., 2019; Tellez et al., 2019; Otálora et al., 2019). The high heterogeneity of clinical data, together with the fact that there are only a few publicly available datasets, hinders the generalization of DL models, and thus the development of universal CAD systems for specific cancers remains an unsolved challenge.

Prostate cancer is the second most frequently diagnosed cancer among men, with more than 1.2 million cases worldwide, and the fifth leading cause of cancer death, with around 350'000 deaths in 2018 (Rawla, 2019). A biopsy is the most reliable test to confirm the presence of prostate cancer (Borley and Feneley, 2009). The samples obtained from a biopsy are processed and viewed under a microscope or scanned resulting in digital images. These are usually gigapixel-resolution WSIs or tissue microarray (TMA) cores, depending on how they were acquired. WSIs are obtained after scanning the glass slide containing the whole biopsy sample, generating a large digital image (Farahani et al., 2015). TMA cores are tissue cylinders of 0.6–2.00 mm diameter extracted from the biopsy sample (Eskaros et al., 2017).

The aggressiveness of prostate cancer is evaluated through a scoring system called the Gleason Grading System (GGS) (Matoso and Epstein, 2016). This system allows pathologists to assign a score to a tissue sample based on its microscopic appearance and the patterns of the cancer cells. The GGS determines the cellular differentiation degree of prostate tumors considering 5 Gleason Pattern (GP) (1 to 5) (Amin and Tickoo, 2016). Pathologists examine the image and assign a lower or higher pattern depending on the tissue appearance. GP 1 is assigned to areas of the tissue containing cells that resemble normal prostate cells, while in GP 5, cancer cells greatly differ from normal prostate cells. The higher the pattern, the higher the aggressiveness of the cancer and the lower the differentiation between cancer cells. Fig. 1 shows a few examples obtained from H&E-stained tissue images that include benign cases and the three most commonly used GPs: 3, 4 and 5. The two most predominant GPs in an image are summed up to assign the corresponding Gleason score (GS), which ranges from 2 to 10. However, GS 2 to 5 are almost never present, since, in these cases, biopsies are usually not taken until the tumor has advanced (Chen and Zhou, 2016). Therefore, a GS of 6 (3+3) is usually the lowest score, scores of 7 (3+4/4+3) and 8 (4+4) correspond to a mid-grade cancer, and a score of 9–10 (4+5/5+4/5+5) corresponds to a high-grade cancer. Lower-grade cancers grow more slowly and there is a lower risk of spreading compared to high-grade cancers. The GS allows pathologists to determine the tumor status and aggressiveness and predict the biological behavior of the tumor to plan treatments. The most appropriate therapy for the patient is determined based on this score. Although the GGS is currently the most widely used grading system for prostate cancer, many studies have found a high inter-observer variability among pathologists when diagnosing prostate cancer based on this system (Lessells et al., 1997; McLean et al., 1997), reporting more than 30% in terms of differences in the GS (Arvaniti et al., 2018; Salmo, 2015).

With the recent advancements in DL techniques for histopathology image classification, several methods have been developed to exploit the characteristics of the data, such as fully-supervised (Arvaniti et al., 2018; Ström et al., 2019; Nagpal et al., 2019; Campanella et al., 2019; Duran-Lopez et al., 2020), weakly-supervised (Campanella et al., 2019; van der Laak et al., 2019; del Toro et al., 2017; Arvaniti and Claassen, 2018; Otálora et al., 2020a, 2021; Ilse et al., 2018; Lu et al., 2021; Chikontwe et al., 2020; Li et al., 2021b; Yao et al., 2020) and semi-supervised methods (Bulten et al., 2020; Marini et al., 2021c; Otálora et al., 2020b; Shaw et al., 2020; Pulido et al., 2020; Tolkach et al., 2020; Schmidt et al., 2022; Lai et al., 2021). Several datasets have been released, such as the Prostate cANcer graDe Assessment (PANDA) Challenge dataset, The Cancer Genome Atlas-PRostate ADenocarcinoma (TCGA-PRAD), the Gleason 2019 Challenge from MICCAI, SICAPV2 and Diagset. In particular, the PANDA dataset (Bulten et al., 2022), which was released during a competition hosted in MICCAI 2020, still represents the largest publicly available dataset with local annotations in the computational pathology domain, with over 10'000 WSIs that are pixel-wise annotated.

This paper presents a systematic comparison of different state-of-the-art training approaches for Gleason grading (patch-level) and Gleason scoring (WSI-level and TMA core-level). These methods are fully-supervised learning, weakly-supervised learning, semi-supervised learning and MIL. A total of 9 heterogeneous datasets from various sources were used for training, validating and testing the models, including around 13'000 WSIs and 1'100 TMA cores, in order to evaluate the performance and the generalization capability for each of the methods considered.

1.1. Related work

Several approaches were developed for training deep CNNs for the classification of prostate Gleason grading and scoring, including full supervision, weak supervision, semi-supervision and MIL.

1.1.1. Fully-supervised learning

Fully-supervised learning approaches include methods developed to train machine learning models using patch-level (or local) annotations (Arvaniti et al., 2018; Ström et al., 2019; Nagpal et al., 2019; Campanella et al., 2019; Duran-Lopez et al., 2020). Patch-level annotations include information about pixel-wise regions of the image. These

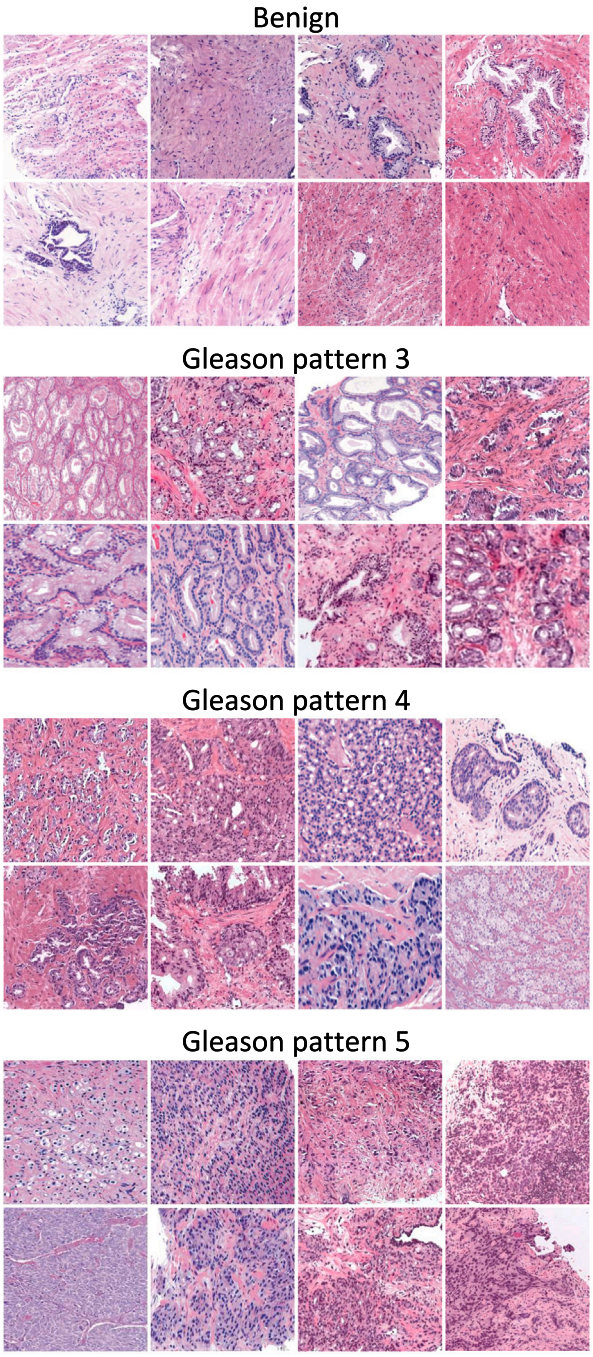


Fig. 1. Examples of benign cases and Gleason patterns 3–5 from different tissue images.

correspond to regions of interest that medical experts have manually annotated and associated with a specific label, based on the identified finding. Arvaniti et al. (2018) showed that fully-supervised algorithms lead to Gleason scoring results on TMA cores that are at the same level as inter-pathologist agreement. Accurate patch-level classification not only allows for high performance on Gleason grading, but also for the development of multi-task models in order to precisely identify malignant regions within the image (Li et al., 2018). However, these types of annotations are hard to collect, as the labeling process is time-consuming and expensive for experts. For this reason, only a few of the publicly-available datasets include patch-level annotations. On the other hand, since these annotations are obtained directly from well-defined regions of interest with no noise, the amount needed to reach high performance is lower than that required by other state-of-the-art methods (Arvaniti et al., 2018; Otálora et al., 2021; Campanella et al., 2019). Particularly in Otálora et al. (2021), an incremental fully-supervised algorithm was used to evaluate the performance of the model depending on the amount of patch-level annotations used.

1.1.2. Weakly-supervised learning

Weakly-supervised learning approaches include methods developed to train machine learning models using image-level (or global) annotations, when patch-level annotations are not available (Campanella et al., 2019; Otálora et al., 2020a; van der Laak et al., 2019). Image-level annotations include information about the whole image, usually focusing on a dangerous disease, such as cancer. The exploitation of these labels is not trivial, since image-level annotations lead to incorrectly-labeled data: usually the findings related to conditions involve small tissue regions and the labels do not include any detail about the regions of interest where the findings are identified. This fact has a dramatic drawback: large datasets are required. For example, Campanella et al. (2019) showed that almost perfect performance in binary tasks (cancer vs non-cancer) can be reached using over 10'000 WSIs. A line of research on this topic includes approaches to train CNNs at the patch-level, through the labeling of patches with the image-level annotations referring to the image where the patches come from, as shown in del Toro et al. (2017), Arvaniti and Claassen (2018), Otálora et al. (2020a, 2021). Since image-level annotations include incorrectly-labeled data, the combination of image-level and patch-level annotations may help to increase the performance of the models (Arvaniti and Claassen, 2018; Otálora et al., 2021). However, since CNNs are usually developed to work at the patch-level, it is not completely clear how to aggregate the predictions made on the single patches to have a global prediction. Currently, the weakly-supervised algorithms that show the highest performance in computational pathology tasks are based on the MIL framework (Campanella et al., 2019; Ilse et al., 2018; Lu et al., 2021; Chikontwe et al., 2020; Li et al., 2021b; Ilse et al., 2020; Yao et al., 2020; Marini et al., 2021b; Li et al., 2021a; Shao et al., 2021). MIL allows modeling data as a bag of instances, aggregated to have a global prediction. This formulation fits well with how WSIs are usually currently treated in computational pathology experiments, where WSIs (bags) are split in patches (instances). The instance aggregation can be made on the instance embeddings (embedding-based) (Lu et al., 2021; Chikontwe et al., 2020; Li et al., 2021b; Ilse et al., 2018), or on the instance predictions (instance-based) (Campanella et al., 2019; Schmidt et al., 2022; Marini et al., 2022a; Javed et al., 2022). Embedding-based approaches, through the creation of an embedding representing the WSI, reach higher performance on the task involving predictions at the image-level, although they do not produce predictions on the single instances. On the other hand, instance-based approaches enable predictions on the single instances and at the image-level (Javed et al., 2022). Despite the fact that several aggregation solutions have been developed (such as max pooling and average pooling) (Wang et al., 2019), the state-of-the-art algorithm used to aggregate the instances is currently based on attention networks, as shown in Ilse et al. (2018), Yao et al. (2020), Wang et al. (2019), Hashimoto et al. (2020). Attention-based Multiple Instance Learning (AB-MIL) (Ilse et al., 2018) is a MIL framework aimed to weigh the single instances in order to obtain a global prediction. The framework includes a pooling aggregator, called attention network, that weighs the contribution of every instance (patch) within a bag (WSI), aiming to be permutation-invariant. Their success relies on the fact that they are learnable functions, opposed to other solutions, such as max pooling, which are less flexible to input data, and on the fact that they guarantee interpretability (it is possible to use the attention values to generate heatmaps, highlighting where models are focusing) on the model outcomes. AB-MIL is designed to work on binary classification problems; therefore, a single attention channel is used in the attention pooling layer. Additive-MIL (Javed et al., 2022) is a MIL framework aiming to

solve the problems related to AB-MIL, especially in multiclass scenarios. In AB-MIL models, high attention weights, linked to a patch, do not necessarily imply that the patch is responsible for the global image-level prediction. This characteristic raises a problem in multiclass problems, where attention scores do not provide information about the class-wise importance of a patch, since the network does not differentiate between positive and negative contributions of patches to image-level predictions. To solve this limitation, Additive-MIL includes an attention pooling layer with an attention channel for every output class. This contribution guarantees more interpretability in multiclass networks, providing a heatmap for every class. Clustering-constrained Attention Multiple Instance Learning (CLAM) (Lu et al., 2021) is a MIL framework exploiting an attention-based network to aggregate single WSI patches. The attention mechanism aims to highlight relevant sub-regions of WSIs to improve the global image prediction. Furthermore, CLAM adopts a cluster mechanism at instance-level to aggregate and refine representative regions, aiming to enrich the WSI-representation. Dual-Stream MIL (Li et al., 2021a) is a MIL framework aiming to both produce patch-level and image-level predictions and to enrich the WSI-level representation of a self-supervised learning algorithm. The instance-level predictions are optimized using a max-pooling mechanism on the single instance predictions. An attention mechanism is used to aggregate the single instances to have an image-level embedding, which is used to classify the WSI. Furthermore, the DS-MIL framework exploits a contrastive self-supervision learning algorithm to improve the WSI representation used to classify images, and it combines data collected from multiple magnification levels, in order to enrich the image-level representation. TransMIL (Shao et al., 2021) is a MIL framework aiming to exploit spatial and morphological information included within WSIs. The framework aims to overcome the attention network mechanism, which does not take into account the spatial relationship among input instances, exploiting the Transformer architectures (Vaswani et al., 2017). Transformer architecture models data as a sequence of tokens (patches) aiming to highlight relationships among single instances.

1.1.3. Semi-supervised learning

Semi-supervised approaches include methods developed to train machine learning models using automatically labeled data (Bulten et al., 2020; Marini et al., 2021c; Otálora et al., 2020b; Shaw et al., 2020; Pulido et al., 2020). The approaches are based on the development of an automatic algorithm that annotates unlabeled data, reducing the effort needed by medical experts for the annotations (Schmidt et al., 2022; Marini et al., 2021c; Tolkach et al., 2020). The collection of unlabeled data in medical domains is a relatively cheap task, since no medical experts are required to make annotations, and due to the increasing number of publicly released datasets (Marini et al., 2021c; Peikari et al., 2018; Foucart et al., 2019). Although the combination of patch-level annotated data and automatically-labeled data have been shown to improve the generalization performance of machine learning algorithms, automatically-annotated data are noisy, due to possible errors or biases within the algorithm (Zhang et al., 2021). Semi-supervised approaches usually involve two roles: the annotator and the model that exploits the automatically-labeled data. These roles may be interpreted by the same model or by several models. In the former case, the same model is trained with annotated data, annotates unlabeled data, and then exploits them to be fine-tuned (Tolkach et al., 2020; Schmidt et al., 2022; Lai et al., 2021). In the latter and more conventional case, two or more models are involved. An example of semi-supervised paradigm including two models is the Teacher/Student method (Zhou et al., 2020; Marini et al., 2021c; Bulten et al., 2020). In this paradigm, the teacher annotates unlabeled data, which are subsequently used to train the student. The teacher may be a larger model than the student (Marini et al., 2021c) (in terms of parameters) or it can be built with the same architecture of the student (Ke et al., 2019). Finally, it is possible to have a chain of models (Shaw et al., 2020), where each of them is trained with an increasing amount of automatically-labeled data and then annotates new unlabeled data for the next one in the chain.

1.1.4. Self-supervised learning

Self-supervised learning framework aims to learn a relevant data representation from unlabeled data, that can be re-used, after a fine-tuning, to perform downstream (specialized) tasks. The framework objective is to limit the need for experts to annotate large datasets, exploiting the increasing availability of data (Liu et al., 2021; Chen et al., 2020; He et al., 2020). Self-supervision is reaching increasing success in domains such as computational pathology, where manual annotations are expensive to collect. Typically, CNN backbones are pre-trained on ImageNet (Deng et al., 2009) data. However, the dataset includes natural images; thus, pre-trained weights learnt features that might not be suited to be used on solving computational pathology tasks. Even if the learnt representation is strong, the CNN must be fine-tuned afterwards, to perform peculiar tasks, since it cannot solve any task, such as classification or segmentation. Among self-supervised algorithms, contrastive learning algorithms are currently the most widely adopted solutions, such as MoCo (He et al., 2020) and simCLR (Chen et al., 2020), in contrast with generative algorithms. These algorithms show similar characteristics, since they were both developed to learn a representation of data clustering from similar samples: both algorithms minimize the distance between the feature vectors representing similar samples and maximize the distance between the feature vectors representing dissimilar samples. Couples of similar and dissimilar examples are generated using data augmentation: a sample is similar to its augmented version and dissimilar to other samples in a batch. The samples are also named queries. In the computational pathology domain, the adoption and development of self-supervised algorithms is constantly gaining ground (Dehaene et al., 2020; Ciga et al., 2022; Srinidhi et al., 2022; Wang et al., 2022). Dehaene et al. (2020) showed that the pre-training of a CNN with a contrastive learning algorithm helps to learn stronger features, which can be re-used for training a MIL CNN, comparing the performance with other MIL CNN pre-trained on ImageNet. Ciga et al. (2022) exploited 57 datasets to build a more effective histopathology representation. The CNN, after a fine-tuning, showed more accurate performance and feature robustness, compared with ImageNet weights or random weights, to classify WSIs. Furthermore, the paper shows that the representation is effective, since it allows to cluster patches according to tissue morphologies. Srinidhi et al. (2022) presented a self-supervised algorithm combined with a semi-supervised algorithm: firstly, the model was trained to learn a robust data representation (exploiting the multi-scale structure of WSIs as downstream tasks); then, it was fine-tuned on a limited amount of data in order to learn how to transfer the self-supervised representation to peculiar tasks; finally, the downstream task performance was improved by combining a small amount of labeled data (used for downstream task training) and a large amount of unlabeled data, pseudo-labeled by the model. The algorithm was tested on the classification of tumor metastasis, on the classification of tissue type and on the regression of tumor cellularity quantification, showing an increase in performance. Wang et al. (2022) presented CTransPath, which is a hybrid model that combines a Convolutional Neural Network (CNN) with a multi-scale Swin Transformer architecture. This model was pretrained on unlabeled histopathological images and was evaluated on various downstream tasks, including patch retrieval, patch classification, weakly-supervised whole-slide image classification, mitosis detection, and colorectal adenocarcinoma gland segmentation. The results showed state-of-the-art performance, while also outperforming other self-supervised methods in terms of robustness and transferability.

1.2. Main contributions

The main contributions of this work include the following:

- A systematic comparison of state-of-the-art training approaches on Gleason grading and Gleason scoring tasks (including fully-supervised learning, weakly-supervised learning, semi-supervised learning, Additive-MIL, AB-MIL, DS-MIL, TransMIL, CLAM and self-supervised learning).

Table 1
Summary of the datasets used in this work. The number of patients in Gleason Challenge and Diagset datasets are not specified in their corresponding publications.

Dataset	Number of patients	Number of images	Pixel-wise annotations	Image-level annotations
TMAZ	886	886 cores	Yes	Yes
Gleason Challenge	–	237 cores	Yes	Yes
SICAPv2	95	119 WSIs	Yes	Yes
Valme	199	938 WSIs	Yes	Yes
PANDA Challenge	2'113	10'516 WSIs	Yes[a]	Yes
Diagset	–	375 WSIs	Yes	No
Subset of TCGA-PRAD	300	300 WSIs	No	Yes
Clinic	43	221 WSIs	No	Yes
Puerta del Mar	18	144 WSIs	No	Yes

[a] Pixel-level annotations were not performed by expert pathologists but by following a semi-automatic procedure.

- Training state-of-the-art models on highly heterogeneous data sources related to Gleason grading and scoring, providing a benchmark for heterogeneous data training. The combination of a total of nine different datasets (six of which are publicly available) was used for training, validating and testing the trained models. This allowed evaluating the performance and the generalization of the methods over many datasets.
- Testing state-of-the-art models using as external test-set the largest publicly-available dataset currently available (PANDA Challenge dataset), providing a reference for the evaluation of models in the literature.
- Fully-supervised learning shows the highest performance in patch-level classification tasks.
- MIL methods obtain the highest performance in image-level classification tasks, particularly CLAM.

The rest of the paper is structured as follows: Section 2 presents the different materials and methods used in this work, including detailed information regarding the datasets (Section 2.1), the way in which patches were extracted from the images, preprocessed and augmented (Section 2.2), the different training approaches evaluated (Section 2.3), the deep learning framework and hyperparameters used (Section 2.4), and the metrics used to evaluate the performance of the models (Section 2.5). In Section 3, the results of the trained models are reported, dividing between patch-level results (Section 3.1) and image-level results (Section 3.2). The performance results obtained for each of the evaluated models are discussed in Section 4, where they are compared in detail. Finally, the conclusions of the paper are presented in Section 5.

2. Materials and methods

2.1. Datasets

Nine datasets from different sources were used to train, validate and test the CNN models. These are the Tissue MicroArray dataset Zurich (TMAZ), the Gleason 2019 Challenge from MICCAI, SICAPv2, Diagset, a subset of TCGA-PRAD, PANDA Challenge, Valme, Clinic and Puerta del Mar datasets. All these datasets are publicly available, except Valme, Clinic and Puerta del Mar, which are not public yet, but they are expected to be in the near future. Table 1 summarizes the number of images (WSIs or TMA cores) and patients used from each dataset and whether each of them includes pixel-wise and image-level annotations reported by expert pathologists. Table 2 shows a detailed distribution of the image-level annotations for those datasets that include them.

- TMAZ (Arvaniti et al., 2018) is a public dataset that includes 886 prostate TMA cores from the University Hospital of Zurich corresponding to 886 different patients. These cores were scanned with the NanoZoomer-XR Digital slide scanner (Hamamatsu) at magnification 40× (0.23 μm per pixel). Each image is 3'100 pixels. TMAZ combines both image-level and pixel-wise annotations from different pathologists.
- Gleason 2019 Challenge from MICCAI (Nir et al., 2018) consists of a set of TMA cores annotated in detail by several expert pathologists. It includes 237 cores of 5'120 pixels digitized at magnification 40× (0.25 μm per pixel) using a SCN400 Slide Scanner (Leica Microsystems, Wetzlar, Germany). Both pixel-wise and image-level annotations are provided for malignant images.
- SICAPv2 (Silva-Rodríguez et al., 2020) consists of 119 WSIs from 95 different patients. The images were obtained at magnification 40× (0.25 μm per pixel) with a Ventana iScan Coreo scanner from Roche. SICAPv2 does not include benign WSIs. For malignant cases, both global GS and patch-level GP annotations are provided.
- Diagset (Koziarski et al., 2021) includes 375 WSIs digitized using a Hamamatsu C12000-22 (0.25 μm per pixel). This dataset provides both normal and malignant cases, reporting only pixel-wise annotations given by a group of pathologists. No image-level annotations are available.
- Valme includes 938 WSIs from 199 different patients obtained from Virgen de Valme University Hospital in Seville (Spain). These were digitized at magnification 40× (0.25 μm per pixel) with a VENTANA iScan HT scanner from Roche Diagnostics. Valme contains normal and malignant WSIs with image-level annotations. A total of 70 out of the 938 WSIs were pixel-wise annotated by a pathologist.
- TCGA-PRAD (Zuley et al., 2016; Clark et al., 2013) is a repository with 500 WSIs obtained from different centers. The dataset includes adenocarcinomas, cystic, mucinous and serous tumors and ductal and lobular neoplasms. These WSIs were scanned at magnification 40× (no information regarding the scanners employed to digitize the images is reported). Together with the WSIs, the pathologist reports with the diagnosis information are provided. However, since neither slide-level nor patch-level annotations are included, a subset of 300 WSIs from 300 different patients was selected and manually labeled at the image level based on the full medical diagnosis reported.
- Clinic is composed of 221 WSIs from 43 different patients scanned at Clinic Hospital from Barcelona (Spain) using a VENTANA iScan HT scanner (Roche Diagnostics) at magnification 40× (0.25 μm per pixel). Only image-level annotations are provided without any pixel-wise report. It contains both normal and malignant WSIs.
- Puerta del Mar contains 144 WSIs from 18 different patients obtained from Puerta del Mar University Hospital in Cádiz (Spain). These images were scanned at magnification 40× (0.2431 μm per pixel) using a MIRAX SCAN from Zeiss. Only malignant and normal image-level annotations are reported, without specifying the corresponding GS for each malignant WSI. In addition, this dataset does not include pixel-wise annotations. For normal cases, 79 WSIs were provided: 33 of them were obtained from needle core biopsy and the remaining 46 images are whole mount (from the entire specimen after surgery). On the other hand, 65 WSIs were labeled as malignant, 26 of which were obtained from needle core biopsy and 39 are whole mount.
- PANDA Challenge (Bulten et al., 2022) consists of 10'516 digitized WSIs from 2'113 patients with their corresponding ground truth of primary and secondary GPs originating from two different sources (Radboud University Medical Center and Karolinska Institutet). This makes PANDA the largest publicly-available WSI dataset at the moment. The authors used different scanners with slightly different maximum microscope resolutions and worked with different pathologists for the labeling process. WSIs of the biopsies were obtained using four different scanner models:

Table 2
Dataset distribution of the image-level annotated TMA cores and WSIs used.

Dataset	Benign	GS6	GS7=3+4	GS7=4+3	GS8	GS9-10
TMAZ	115 cores	272 cores	89 cores	52 cores	218 cores	140 cores
Gleason Challenge	17 cores	63 cores	31 cores	21 cores	100 cores	5 cores
SICAPv2	–	14 WSIs	22 WSIs	23 WSIs	18 WSIs	42 WSIs
Subset of TCGA-PRAD	–	38 WSIs	58 WSIs	54 WSIs	62 WSIs	88 WSIs
Valme	281 WSIs	285 WSIs	139 WSIs	119 WSIs	56 WSIs	58 WSIs
PANDA Challenge	2'873 WSIs	2'616 WSIs	1'340 WSIs	1'227 WSIs	1'245 WSIs	1'215 WSIs
Clinic	142 WSIs	42 WSIs	13 WSIs	8 WSIs	7 WSIs	9 WSIs
Puerta del Mar	79 WSIs		65 malignant WSIs with no GS information			
Total	132 cores	335 cores	120 cores	73 cores	318 cores	145 cores
	3'375 WSIs	2'995 WSIs	1'572 WSIs	1'431 WSIs	1'388 WSIs	1'412 WSIs
			+ 65 malignant WSIs with no GS information			

3DHISTECH Pannoramic Flash II 250 (0.24 μm per pixel at magnification 40×), Leica Aperio AT2 (0.50 μm per pixel at magnification 20×, and 0.25 μm at magnification 40×), Hamamatsu C9600-12 (0.45 μm per pixel at magnification 20×) and Hamamatsu C13220-01 (0.46 μm per pixel at magnification 20×). Among all the WSIs that the dataset contains, a subset of 5'060 (those obtained from Radboud University Medical Center) were also pixel-wise annotated. However, these annotations were not performed by expert pathologists but by following a semi-automatic procedure (Bulten et al., 2022). Therefore, pixel-wise annotations from this dataset were not used in this work.

2.2. Image preprocessing

As was introduced in Section 1, current hardware cannot work with gigapixel-size images as input to CNNs due to limited memory. A widely-known approach to overcome this issue is to tile the images into smaller subimages called patches, which can be handled by the CNN. All the images from the different datasets that were presented in Section 2.1 were preprocessed following the same pipeline: firstly, the background regions from the image are obtained and removed; then, the image is tiled into a set of patches; and, finally, the patches are extracted and selected, using Multi_Scale_tools library (Marini et al., 2022b).

In this work, all the patches were extracted at 40× magnification and with 750 × 750 pixel size, except for some datasets, such as PANDA and TCGA, in which part of the samples were scanned at 20× magnification. For these cases, patches were extracted at 20× magnification and with 375 × 375 pixel size. 40× magnification was selected based on the fact that some datasets, such as TMAZ, only provided this magnification level for the samples. On the other hand, the aforementioned patch size was selected considering that previous studies also used the same size (Arvaniti et al., 2018; Arvaniti and Claassen, 2018; Marini et al., 2021c). The extracted patches were then downsampled (resized) to 224 × 224 pixels, since it is the required input size used for the two different pre-trained backbones that were used in this work (DenseNet121 and Resnext50_32x4d, as detailed in Section 2.4).

As was previously mentioned, before extracting and selecting the patches from the datasets, the background was first removed, since it does not provide any useful information for the GP and GSs classification tasks. This process was performed in two different ways, depending on the nature of the dataset and the images that compose it:

- For TMA cores (TMAZ and Gleason challenge datasets), the background of each core was extracted following the masks provided by the expert pathologists' annotations. 750 × 750 pixel-size patches were densely extracted (without overlapping) from the cores if they contained at least 60% tissue, and then downsampled to 224 × 224 pixels. For the TMA cores that contain pixel-wise annotations made by pathologists (cores from TMAZ and Gleason challenge), 30 patches with possible overlapping were randomly extracted from each, taking into account that patches had to overlap at least by 70% with the annotation. The label of each patch was directly assigned based on the label of the annotated region from which it was extracted.
- For WSIs (Valme, Clinic, Puerta del Mar, Gleason Challenge, TCGA-PRAD, PANDA Challenge and Diagset datasets), the background was removed using masks that were generated by means of the HistoQC[2] software tool (Janowczyk et al., 2019), which is an open-source quality control application for the automated assessment of digital pathology slides. The masks were generated in a way to exclude not only the background of the slide, but also unwanted areas that do not correspond to tissue, such as pen marks and other external agents. After applying the mask to the original WSIs, 750 × 750 pixel-size patches were densely extracted (without overlapping) from them and downsampled to 224 × 224 pixels. The amount of patches extracted per WSI is not a fixed number, and depends on the amount of tissue that the slide contains. For the WSIs that contain pixel-wise annotations made by pathologists (slides from Valme, Gleason challenge and Diagset), patches were densely-extracted only from the annotated regions (with at least 90% overlap with the annotation). The label of these patches was directly assigned based on the label that the pathologist associated with the region from which the patch was extracted. On the other hand, for WSIs with no specific annotations where only the global Gleason grading was present (the two most frequent GPs in the slide), the patches were densely-extracted from the whole slide and labeled with the primary GP of the WSI. These image-level labels are not ground truth, since not all the tissue within the slide corresponds to the primary GP, and most of it could even correspond to benign tissue. In order to reduce noise, a subset of the patches was selected for each WSI using the Blue Ratio (BR) method described in Chang et al. (2012). BR assigns a value to each patch depending on the amount of blue that is present on the patch. Therefore, patches with a denser nuclei concentration will have a higher BR value. Among all the patches extracted from a single WSI, only the first 20% top-ranked samples ordered by decreasing BR value were selected, up to a total of 500 patches per slide.

2.2.1. Data partitions

The aforementioned pipeline was followed to extract patch-level and image-level annotated patches from the WSIs and TMA cores from the datasets presented in Section 2.1 in order to train, validate and test the proposed methods (see Section 2.3).

Table 3 presents the distribution of the patches with patch-level annotations extracted from Valme, TMAZ, SICAPv2, Gleason challenge and Diagset datasets. The number of patches per class (benign, GP3, GP4 and GP5) and per training subset (train, validation and test) is reported, which results in a total of 96'370 patches. The class-wise

[2] https://github.com/choosehappy/HistoQC Retrieved May 11, 2024.

Table 3
Distribution of the patches extracted from pixel-wise annotations from Valme, TMAZ, Gleason challenge, SICAPv2 and Diagset datasets. The partitioning was performed by patient, meaning that patches obtained from slides from the same patient were only present in a single partition.

Class	Train	Validation	Test	*Total*
Benign	20'486	4'737	12'848	38'071
GP3	14'034	2'995	9'401	26'430
GP4	10'937	2'228	10'731	23'896
GP5	5'471	766	1'736	7'973
Total	50'928	10'726	34'716	**96'370**

Table 4
Dataset distribution of the patches with patch-level annotations used for each of the datasets.

Dataset	Benign	GP 3	GP 4	GP 5
TMAZ	3'487	8'946	7'424	3'610
Gleason challenge	1'080	2'431	3'649	100
SICAPv2	11'069	10'784	2'979	2'767
Valme	13'652	3'026	5'510	800
Diagset	8'783	1'243	4'334	696
Total	38'071	26'430	23'896	7'973

representation for each of the datasets is presented in Table 4. Gleason challenge and Diagset datasets were only represented in the test subset. The division of the patches in the three training subsets was done considering that patches from the same patient were only present in a single subset. Thus, no patient was represented in more than one subset at the same time.

Regarding the image-level annotations of TMA cores and WSIs, Table 5 presents the distribution of the images in the 6 different classes (benign, GS6, GS7=3+4, GS7=4+3, GS8, GS9-10) and the three training subsets. These images correspond to Valme, Clinic, TCGA-PRAD, TMAZ, SICAPv2, PANDA Challenge and Gleason challenge datasets. Both TCGA-PRAD and Gleason challenge were only represented in the test subset. From the 12'817 images (including TMA cores and WSIs) that were used, a total of 2'515'327 patches were extracted from image-level annotations, whose distribution is presented in Table 6. As was done for the patch-level annotations, patches were divided into the train, validation and test subsets following a distribution where the patches from the same patient were only represented in a single subset.

The external test set was created to be highly heterogeneous, including half of the biggest dataset available, PANDA (those WSIs sourced from Radboud University Medical Center), the whole TCGA-PRAD subset, and several other datasets (Diagset and Gleason challenge). The latter three aforementioned datasets were not used for training nor evaluating any of the training methods considered. In the case of PANDA, since we did not have information regarding the WSIs that correspond to each patient, image-level annotated patches obtained from images sourced from Karolinska Institutet were used in the training set, while those from Radboud University Medical Center were only present in the test set, with none of them being considered for the validation set.

2.2.2. Data augmentation

Three different types of operations were applied to augment the training datasets and, thus, to increase the variability of the data and avoid overfitting. These three operations are rotations, flips and color augmentation. Only three rotations are considered in this augmentation: 90, 180 and 270 degrees. These rotations were selected based on the fact that they allow avoiding the need of filling the empty space in the corners that appears when not rotating images by steps of 90 degrees. Regarding flips, both horizontal and/or vertical flips were applied to the input image. The color augmentation was performed by defining a range of values for each of the parameters in the HSV representation (hue, saturation and value). The hue shift limit was set between −15 and 8, the saturation shift limit was set between −20 and 10, and the value shift limit was set between −8 and 8. Each of the three aforementioned operations used to augment the variability of the training dataset was applied to every patch with a probability of 0.5. The open-source Albumentations library (Buslaev et al., 2020) was used to perform all these operations automatically every time a patch was loaded into memory.

Table 5
Distribution of the images (WSIs and TMA cores) with image-level annotations from Valme, TMAZ, Gleason challenge, SICAPv2, TCGA-PRAD, Clinic and PANDA challenge datasets. The partitioning was performed by patient, meaning that images from the same patient were only present in a single partition.

Class	Train	Validation	Test	*Total*
Benign	2'222	131	1'075	3'428
GS6	2'200	116	1'014	3'330
GS7=3+4	765	25	902	1'692
GS7=4+3	381	22	1'101	1'504
GS8	652	28	1'025	1'705
GS9–10	400	23	1'135	1'558
Total	6'220	345	6'252	**12'817**

Table 6
Distribution of the patches extracted from image-level annotations (WSIs and TMA cores) from Valme, TMAZ, Gleason challenge, SICAPv2, TCGA-PRAD, Clinic and PANDA challenge datasets. The partitioning was performed by patient, meaning that patches obtained from slides from the same patient were only present in a single partition.

Class	Train	Validation	Test	*Total*
Benign	484'336	35'945	148'070	668'381
GP3	610'631	27'710	345'728	984'069
GP4	259'805	9'789	490'432	760'026
GP5	22'806	2'529	77'516	102'851
Total	1'377'608	75'973	1'061'746	**2'515'327**

2.3. Training approaches

2.3.1. Fully-supervised learning

In fully-supervised learning, training is performed with datasets where each sample has a label associated (patch-level annotations). In this regard, each patch that is used to train the CNN model comes from specific regions within the image that pathologists have manually annotated and associated with one of the four classes considered in this study: benign, GP3, GP4 or GP5.

Therefore, for this learning approach, Valme, SICAPv2 and TMAZ datasets were used, since these contain patch-level annotations among the ones that were considered in this work (Table 1). A total of 50'928 patches from these datasets were considered for the training subset, and 10'726 for the validation subset from the same datasets. Other datasets, such as Diagset and Gleason challenge, also contain patch-level annotations. These were not used in the training and the validation partitions, but on the test partition together with the test sets from Valme, SICAPv2 and TMAZ in order to test the generalization capability of this learning approach. A diagram of the workflow for the fully-supervised learning approach is presented in Fig. 2.

2.3.2. Weakly-supervised learning and hybrid approaches based on weak and strong supervision

Three different variants of weakly-supervised learning algorithms were proposed. In the first variant, the network is trained using datasets with image-level annotations, labeling each patch with the global label corresponding to the image from where the patches were extracted. In the second variant, the network is trained using datasets with image-level annotations (as described for the previous variant) and then fine-tuned with datasets that contain patch-level annotations (those that were used for the fully-supervised learning, where each sample has a specific label associated). In the third variant, the patches obtained

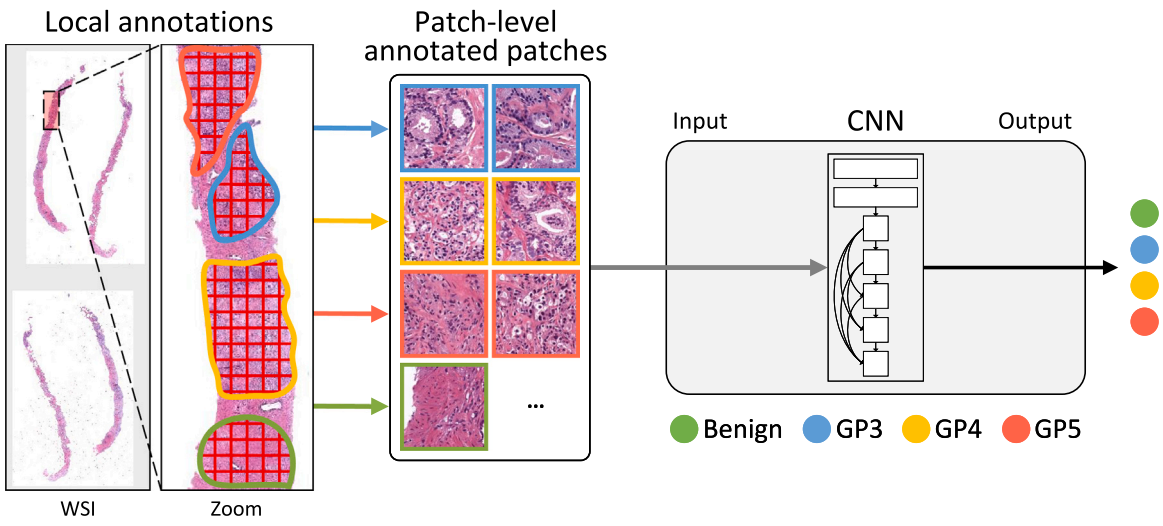


Fig. 2. Diagram of the fully-supervised learning approach. Patches are densely-extracted from local annotations, and then used to train the CNN.

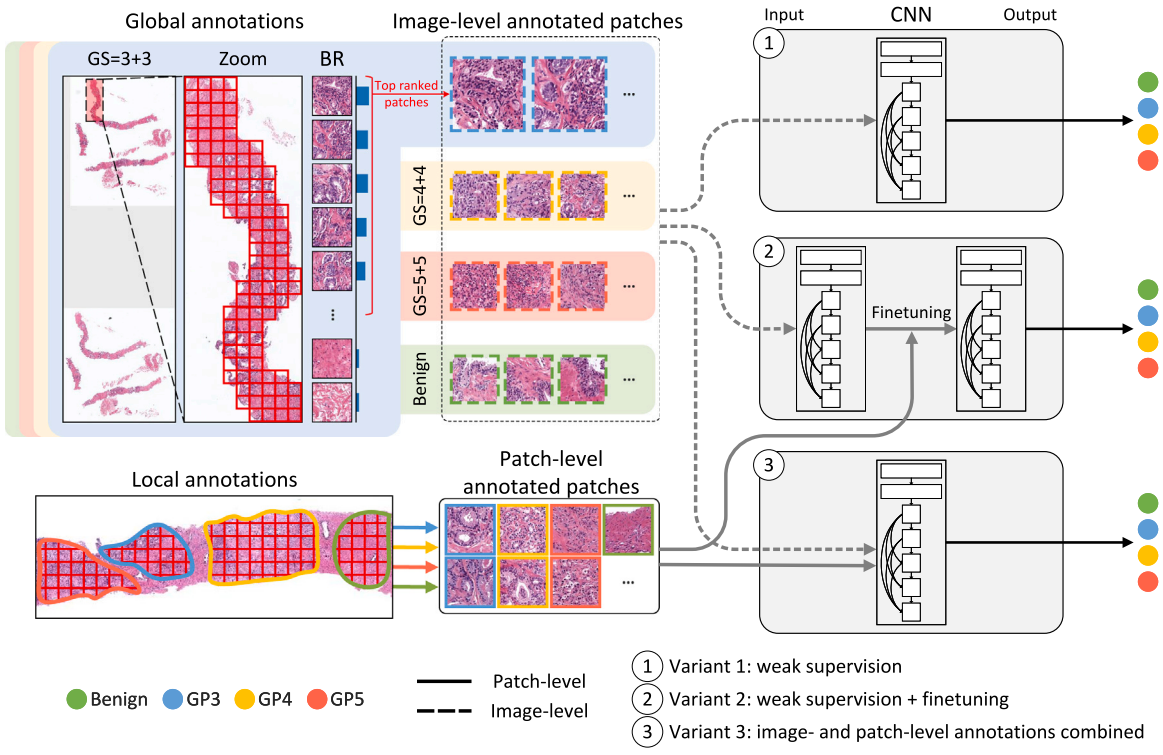


Fig. 3. Diagram of the 3 different weakly-supervised learning approach variants. Patches are densely-extracted from images with global annotations after selecting the top ranked patches based on descending BR. Patch-level annotated patches are obtained from images with local annotations. Variant 1 is trained only with the former. For variant 2, the network is trained with image-level annotated patches and then fine-tuned with patch-level-annotated patches. In variant 3, a combination of image-level and patch-level annotations is used to train the model.

from image-level annotations and patch-level annotations are combined to train the network. The overall workflow of the three different variants is summarized in the block diagram presented in Fig. 3.

Variant 1: Weak supervision. The first variant consisted of training the CNN with image-level annotated data (unsupervised learning). To this end, the densely-extracted patches from the images in the train and validation partitions from Valme, TMAZ, SICAPv2, PANDA challenge and Clinic datasets were used for training and validating the network. A total of 6'220 images (WSIs and TMA cores) were used to train the network, which correspond to 1'377'608 patches. For validating the network, 75'973 patches from 345 images were used.

Variant 2: Transfer learning. This variant consists of two different steps in the learning process: the model is first trained with patches obtained from image-level annotations, as in the previous variant. After training the network in an unsupervised manner, the second step consisted in fine-tuning the aforementioned network. To this end, the weights of the model were transferred and then fine-tuned with patches obtained from images with patch-level annotations (the same ones that were used in the fully-supervised learning approach, i.e., patches from Valme, SICAPv2 and TMAZ datasets) in order to improve its performance. In this way, the CNN builds on top of the generalization capability achieved with the first step, and the pixel-wise accuracy is improved with the patch-level annotated data used in the second step.

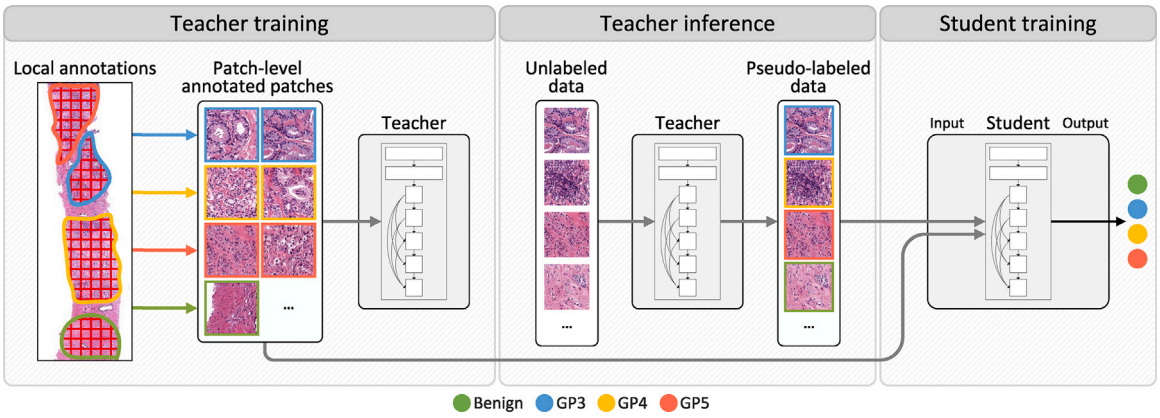


Fig. 4. Diagram of the workflow considered for the semi-supervised learning approach. Patch-level annotations are used to train the teacher model. Then, the trained teacher is used to label patches extracted from unlabeled images. These pseudo-labeled patches, together with the patch-level annotated patches used in the first step, are used to train the student model.

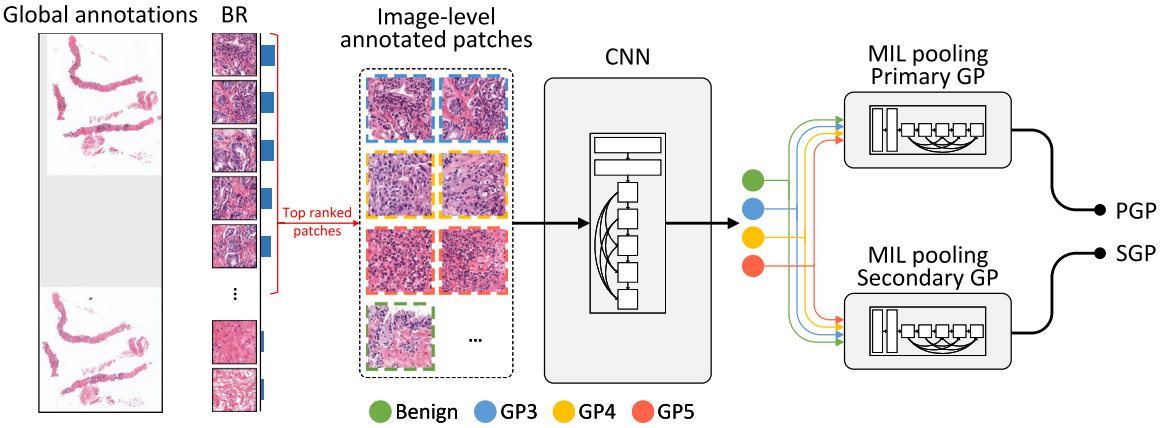


Fig. 5. Diagram of instance-based MIL learning approaches. Patches are densely-extracted from images with global annotations after selecting the top ranked patches based on descending BR. The CNN performs predictions at the patch level, which are aggregated by the implemented aggregation model, reporting both image-level and patch-level predictions.

Variant 3: Combination of image-level and patch-level. annotations

The third variant of the weakly-supervised learning method that was considered in this work consists of a single step, where both image-level and patch-level annotated patches (the same patches that were used in the previous variant) were combined and used for training and validating the network. Therefore, a total of 1'428'536 patches for training the model, and 86'699 for validating it were used.

The idea behind this approach was to reduce the training time compared to the two-steps variant (unsupervised learning plus fine-tuning), while also benefiting from the broader generalization provided by training with both image-level and patch-level annotated samples.

2.3.3. Semi-supervised learning

The semi-supervised learning approach presented involves two models: a teacher model and a student model (Marini et al., 2021c). The teacher model is trained with patch-level annotated data from SICAPv2 and TMAZ (same partitions as used in the fully-supervised learning approach). The teacher is a CNN with a large number of parameters (more than the student) to capture and learn robust features from the input data. The role of the teacher is to annotate unlabeled patches, collected from Valme, Clinic, TCGA-PRAD, Diagset and Puerta del Mar datasets, generating the so-called pseudo labels. The student model is trained with the pseudo-labeled data and with a small portion of patch-level annotated data (the same samples used to train the teacher model). Fig. 4 presents a block diagram with the workflow followed for the teacher/student learning approach.

2.3.4. Multiple-instance learning

The paper presents seven different MIL CNNs, trained to output the Primary and the Secondary GPs included in the images. These are Additive-MIL, AB-MIL, TransMIL, DS-MIL and CLAM. For Additive-MIL and AB-MIL, both instance-based and embedding-based approaches were considered, with the rest of the methods being embedding-based. The instance-based CNN aggregates the predictions on the single instances (patches) to have an image-level prediction (see Fig. 5); while the embedding-based CNN aggregates the embeddings of the single instances to have an image embedding that is used to obtain an image-level prediction (see Fig. 6). In both frameworks, the aggregation of predictions and embeddings is conducted by exploiting a specific pooling layer (Ilse et al., 2018; Javed et al., 2022; Lu et al., 2021; Li et al., 2021a; Shao et al., 2021), which depends on the MIL implementation. This pooling network is a learnable function, which guarantees more flexibility compared to other pooling layers (such as max pooling or average pooling). The choice of using both frameworks allows showing their advantages and disadvantages: while embedding-based MIL is supposed to reach higher performance on image-level classification, avoiding any possibility to have predictions at the patch-level, instance-based framework guarantees predictions on the single patches, usually reaching lower performance at image-level.

2.3.5. Self-supervised learning

This paper presents two setups for the CNN starting weights, using ImageNet pre-trained weights and self-supervision solutions. CNNs are not usually trained from scratch, due to the limited number of data and

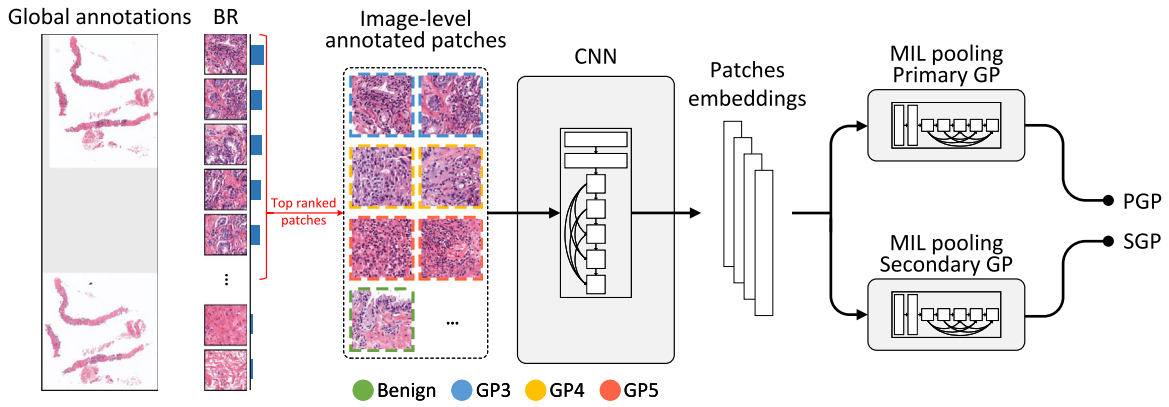


Fig. 6. Diagram of embedding-based MIL learning approaches. Patches are densely-extracted from images with global annotations after selecting the top ranked patches based on descending BR. The CNN generates embeddings, which are aggregated by the implemented aggregation model, reporting image-level predictions.

to their large number of parameters to tune: CNN are commonly pre-trained on large datasets and then trained on particular data to solve a task. The first setup (i.e., ImageNet weights) is a common trend in deep learning algorithms: even if the original dataset includes natural images, the features learnt from these data include patterns such as curves or lines that can match some patterns in histopathology data. However, the features learnt on ImageNet may not catch all peculiar features of histopathology data. The second setup presented in the paper aims to overcome this problem: the CNN is pre-trained using simCLR (Chen et al., 2020), which is a self-supervised algorithm aiming to learn similarity and dissimilarity between input data (in this case between patches).

2.4. Deep learning framework

PyTorch (Paszke et al., 2019) was used to train and evaluate all the different models and to perform all the experiments presented. The CNN models that were used for the fully-supervised learning, the three variants of the weakly-supervised learning and the student from the semi-supervised learning rely on DenseNet121 (Huang et al., 2017). On the other hand, a Resnext50_32x4d (Xie et al., 2017) model implemented by Yalniz et al. (2019) was used for the teacher. Both architectures were modified using a classifier with only four output nodes in the last layer, which correspond to the number of GPs, and thus, to the number of output classes to be classified.

Adam was used as optimizer, with a learning rate of 10^{-3}, and a decay weight of 0. Since classes were not represented equally in terms of the number of patches per class (see Section 2.2.1), scikit-learn's compute_class_weight function was used. This allows estimating the weight of each class in unbalanced datasets in order to weight the loss function during training, avoiding the overfitting of the model on the most represented class. The batch size used when training each of the models was set to 32, and the number of epochs was set to 15, since it was observed that the loss function was not improving after the first 10–12 epochs. These values were used for all the proposed learning approaches except for the transfer learning (second step of the weakly-supervised variant 2), in which only 5 epochs were used to fine-tune the weakly-supervised models (which were previously trained for 15 epochs) with patch-level annotated data. The loss function was evaluated on the validation partition at the end of each epoch, and the weights of the model were only saved if its value was lower than the one achieved in the previous epoch.

2.5. Evaluation metrics

In order to evaluate the performance of the models, the quadratic weighted Cohen's Kappa Score (κ) (Cohen, 1960) was used. This metric measures the agreement or disagreement between the ground truth and the predicted value, where $\kappa = 1$ means perfect agreement between both, while $\kappa = 0$ means that the degree of agreement is the same as would be expected by chance. In this case, the quadratic weighted version of this coefficient was used, since it penalizes predicted values far from their actual class to a greater extent (i.e., in the Gleason grading task, predicting a benign patch as GP3 would penalize less than predicting it as GP4 or GP5).

Since the proposed models (except for MIL) perform the GP classification on input patches, a majority voting mechanism was used to aggregate the patch-level predictions into a GS value. With the predictions performed over all the patches in a single WSI, this method selects the primary and the secondary GPs based on the two most frequent GPs in the image. This method is limited for images where the first and the secondary GPs are the same. A common approach to handle this limitation is considering the image to have the same primary and secondary GPs if it has at least twice the amount of patches predicted with the most predominant GP as with the second one.

3. Results

The model performance is evaluated in κ-score, reporting the average and standard deviation of the performance for each of the learning approaches presented in Section 2.3. For each variant, ten models were trained, providing more realistic results and helping to provide more realistic estimates. Therefore, a total of 130 models were trained and evaluated: 10 fully-supervised, 10 weakly-supervised (variant 1), 10 fine-tuned weakly-supervised with transfer learning (variant 2), 10 weakly-supervised combining image-level and patch-level annotations (variant 3), 10 teachers, 10 students (the teacher that achieved the best performance was used in the training of the students), 10 instance-based Additive-MIL, 10 embedding-based Additive-MIL, 10 instance-based AB-MIL, 10 embedding-based AB-MIL, 10 DS-MIL, 10 TransMIL and 10 CLAM models.

Moreover, in order to measure the impact of different weight initialization approaches, five alternatives were considered: using pre-trained weights from ImageNet and fine-tuning all the layers, using pre-trained weights from ImageNet and fine-tuning only the classification layers (freezing the weights of all layers but the last ones), using self-supervised weights and fine-tuning all the layers, using self-supervised weights and fine-tuning only the classification layers, and using random weights and fine-tuning all the layers. Ten models were trained for each training approach and each of these weight initialization alternatives (except for the different MIL methods, in which only frozen layers with random, ImageNet and self-supervised weights were considered due to the complexity of the models). Therefore, considering all these different experiments, a total of 510 models were trained.

The results for each of the training approaches are divided into patch-level (Gleason grading) and image-level (Gleason scoring) results, which are presented in Sections 3.1 and 3.2.

Table 7
Results obtained for each training approach at the patch-level (Gleason grading). The performance, evaluated in κ, is the average over ten different trained models. The best results for each training approach are highlighted in bold.

Training approach	Weight init.	Dataset					
		Valme	TMAZ	SICAPv2	Diagset	Gleason challenge	*Combined*
Full supervision	ImageNet	**0,7345 ± 0,0324**	0,5677 ± 0,0268	**0,7491 ± 0,0301**	**0,6238 ± 0,0351**	0,5410 ± 0,0685	*0,6432 ± 0,0415*
	Self-supervision	0,7251 ± 0,0336	**0,5841 ± 0,0220**	0,7352 ± 0,0302	0,6007 ± 0,0531	**0,5873 ± 0,0272**	***0,6465 ± 0,0349***
	Random	0,6138 ± 0,0335	0,3555 ± 0,0515	0,6487 ± 0,0322	0,2304 ± 0,0767	0,2666 ± 0,0474	*0,4230 ± 0,0509*
	ImageNet frozen	0,5025 ± 0,0187	0,1745 ± 0,0115	0,4791 ± 0,0661	0,1056 ± 0,0902	0,2464 ± 0,0272	*0,3016 ± 0,0524*
	Self-sup. frozen	0,3179 ± 0,0202	0,3216 ± 0,0159	0,5170 ± 0,0129	0,5267 ± 0,0099	0,4460 ± 0,0138	*0,4258 ± 0,0149*
Weak supervision	ImageNet	0,1926 ± 0,0847	0,4966 ± 0,0697	**0,2742 ± 0,0683**	0,4758 ± 0,0881	0,2931 ± 0,0979	*0,3465 ± 0,0825*
	Self-supervision	**0,2969 ± 0,0890**	**0,5457 ± 0,0502**	0,2297 ± 0,1152	**0,5512 ± 0,0627**	**0,4704 ± 0,0273**	***0,4188 ± 0,0754***
	Random	0,1462 ± 0,0709	0,2052 ± 0,0538	−0,1145 ± 0,1288	0,2356 ± 0,0816	0,2320 ± 0,0409	*0,1409 ± 0,0810*
	ImageNet frozen	0,1661 ± 0,0205	0,4741 ± 0,0196	0,2035 ± 0,0187	0,2569 ± 0,0368	0,2215 ± 0,0534	*0,2644 ± 0,0327*
	Self-sup. frozen	0,1834 ± 0,0316	0,3229 ± 0,0139	0,0617 ± 0,0528	0,3642 ± 0,0177	0,3933 ± 0,0226	*0,2651 ± 0,0310*
Weak supervision + fine-tuning	ImageNet	**0,7038 ± 0,0414**	0,5901 ± 0,0188	**0,7569 ± 0,0334**	**0,6057 ± 0,0799**	0,5202 ± 0,0269	*0,6353 ± 0,0454*
	Self-supervision	0,7018 ± 0,0437	**0,6013 ± 0,0209**	0,7563 ± 0,0240	0,5974 ± 0,0977	**0,5826 ± 0,0588**	***0,6479 ± 0,0565***
	Random	0,4613 ± 0,0409	0,3113 ± 0,0245	0,4793 ± 0,0722	0,2315 ± 0,0449	0,1394 ± 0,1027	*0,3246 ± 0,0633*
	ImageNet frozen	0,6037 ± 0,0120	0,4980 ± 0,0056	0,6774 ± 0,0139	0,5271 ± 0,0190	0,3186 ± 0,0262	*0,5250 ± 0,0168*
	Self-sup. frozen	0,4171 ± 0,0229	0,3612 ± 0,0291	0,5504 ± 0,0143	0,5626 ± 0,0079	0,4889 ± 0,0166	*0,4761 ± 0,0196*
Image- + patch-level annot.	ImageNet	0,4477 ± 0,0961	**0,5596 ± 0,0463**	0,6737 ± 0,0614	0,5054 ± 0,0551	0,2417 ± 0,0703	*0,4856 ± 0,0680*
	Self-supervision	**0,5241 ± 0,0690**	0,5321 ± 0,0361	**0,7295 ± 0,0393**	**0,5304 ± 0,0531**	**0,2625 ± 0,0804**	***0,5157 ± 0,0581***
	Random	0,2483 ± 0,0338	0,1297 ± 0,0826	0,1310 ± 0,1092	0,1709 ± 0,0431	0,2162 ± 0,0302	*0,1792 ± 0,0673*
	ImageNet frozen	0,3271 ± 0,0150	0,4378 ± 0,0199	0,4801 ± 0,0232	0,2854 ± 0,0296	0,1945 ± 0,0277	*0,3450 ± 0,0237*
	Self-sup. frozen	0,3173 ± 0,0200	0,4440 ± 0,0213	0,4757 ± 0,0272	0,2933 ± 0,0225	0,2018 ± 0,0240	*0,3464 ± 0,0231*
Semi-supervision	ImageNet	0,6326 ± 0,0735	0,4614 ± 0,0449	0,7143 ± 0,0501	**0,3968 ± 0,0808**	0,4468 ± 0,0486	*0,5304 ± 0,0613*
	Self-supervision	**0,6384 ± 0,0520**	**0,4760 ± 0,0232**	**0,7400 ± 0,0643**	0,3645 ± 0,1441	**0,4471 ± 0,0534**	***0,5332 ± 0,0787***
	Random	0,4164 ± 0,0238	0,1574 ± 0,0299	0,4835 ± 0,0505	0,1044 ± 0,0389	0,0470 ± 0,0261	*0,2417 ± 0,0352*
	ImageNet frozen	0,5875 ± 0,0296	0,4549 ± 0,0151	0,6712 ± 0,0299	0,0833 ± 0,0405	0,1392 ± 0,0317	*0,3872 ± 0,0305*
	Self-sup. frozen	0,5865 ± 0,0260	0,4584 ± 0,0195	0,6870 ± 0,0166	0,0779 ± 0,0307	0,1286 ± 0,0254	*0,3877 ± 0,0242*
Instance-based Additive-MIL	Random frozen	0,0087 ± 0,0103	0,0127 ± 0,0168	−0,0082 ± 0,0368	0,0081 ± 0,0173	−0,0058 ± 0,0323	*0,0031 ± 0,0248*
	ImageNet frozen	**0,0868 ± 0,0336**	**0,4428 ± 0,0123**	**0,2679 ± 0,0251**	0,3105 ± 0,0283	**0,4862 ± 0,0072**	***0,3188 ± 0,0235***
	Self-sup. frozen	0,0364 ± 0,0166	0,4234 ± 0,0153	0,2602 ± 0,0148	**0,4175 ± 0,0410**	0,4510 ± 0,0137	*0,3177 ± 0,0228*

3.1. Patch-level results

Fully-supervised learning reaches the highest performance on the Gleason grading task. Gleason grading is evaluated considering the patch-level classification performance on the test set including a total of 34'716 patches from Valme, TMAZ, SICAPv2, Gleason challenge and Diagset datasets.

Table 7 presents the average and the standard deviation of the κ of the models for each dataset and weight initialization. In this table, the results for each of the datasets is specified individually, together with a last column showing the combined results, i.e., the κ score achieved on average by the ten models of each training approach. The combined κ was obtained by calculating the average of the κ scores achieved on the five datasets, while the combined standard deviation was obtained by calculating the root mean square of the standard deviations achieved on each of the datasets. As can be observed, most of the different MIL approaches except for instance-based Additive-MIL are not present in the aforementioned table, since, as was explained in Section 1, those training approaches cannot yield a patch-level classification.

Among the 280 models represented in the tables, the confusion matrix and the ROC curves of the one that achieved the best performance (which was trained with the fully-supervised approach and pre-trained with self-supervision) are shown in Fig. 7. In Fig. 7(a), the confusion matrix is presented, which reports the difference in the predictions performed by the pathologists and by the model per class, achieving a combined κ of 0.7185. On the other hand, Fig. 7(b) shows the ROC curves for each of the classes together with their Area Under Curve (AUC) value, which illustrate the diagnostic ability of each class when its discrimination threshold is varied.

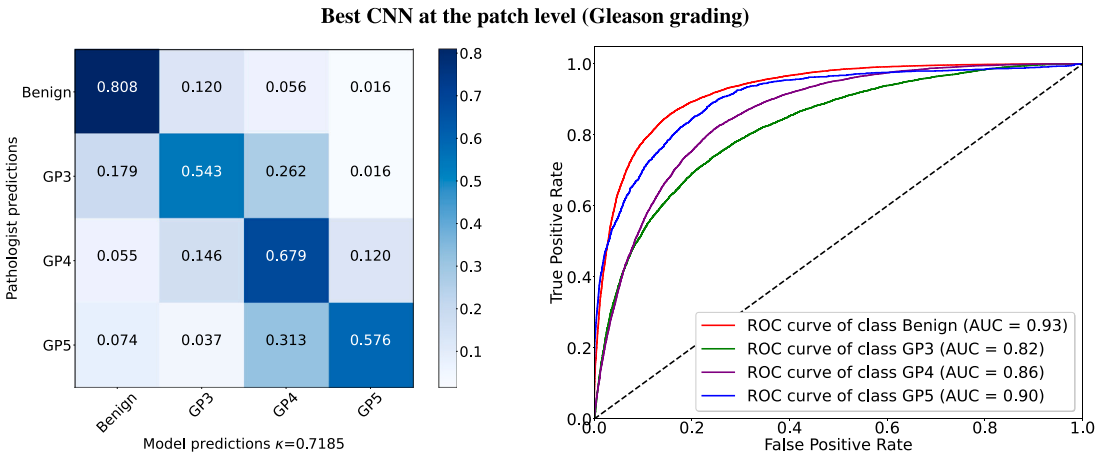
3.2. Image-level results

Clustering-constrained Attention Multiple Instance Learning (CLAM) reaches the highest performance on the Gleason scoring task. Gleason scoring is evaluated considering the image-level classification on the test, including a total of 6'252 images (cores and WSIs).

Table 8 reports the average and the standard deviation of the κ of the models for each dataset and weight initialization. In this table, the results for each of the datasets is specified individually, together with a last column showing the combined results, i.e., the κ score achieved on average by the ten models of each training approach. The combined κ was obtained by calculating the average of the κ scores achieved on the five datasets, while the combined standard deviation was obtained by calculating the root mean square of the standard deviations achieved on each of the datasets. Fig. 8 shows the confusion matrix of the model that achieved the best performance at the image-level among the 460 models that were evaluated. It corresponds to a model that was trained with the CLAM approach and self-supervised weights, which achieved a combined κ of 0.6493.

4. Discussion

Gleason grading and Gleason scoring are still an open challenge in computational pathology, and, in particular, in prostate cancer classification, due to the need of large annotated datasets and of data heterogeneity. This aspect hinders the generalization of deep learning models (the state-of-the-art algorithm in the computational pathology domain) when training, moving away from the idea of having universal CAD systems for specific malignancies. The amount of public datasets including prostate WSIs is yearly increasing, but most of them are only globally-annotated, without any local annotation. Obtaining local (or patch-level) annotations is not easy, since it requires expert pathologists to analyze and locally annotate digitized images, which is time consuming and outside of the scope of their routine work. In this work, we collected several publicly-available prostate cancer histology datasets (TMAZ, SICAPv2, TCGA-PRAD, Gleason challenge, Diagset and PANDA challenge) in combination with three private (but expected to be publicly released in the near future) datasets (Valme,



(a) Normalized confusion matrix.

(b) ROC curve.

Fig. 7. Confusion matrix (left) and ROC curves (right) obtained for the CNN model achieving the best patch-level results (Gleason grading). The results correspond to a fully-supervised model that was pre-trained with self-supervision. The confusion matrix is normalized, and represents all the patches with patch-level annotations in the test set, achieving a combined $\kappa = 0,7185$.

Table 8
Results obtained for each training approach at the image level (Gleason scoring). The performance, evaluated in κ, is the average over ten different trained models. The best results for each training approach are highlighted in bold.

Training approach	Weight init.	Dataset							
		Valme	TMAZ	SICAPv2	Gleason challenge	TCGA	Clinic	PANDA	*Combined*
Full supervision	ImageNet	0,4864 ± 0,0485	0,5377 ± 0,0585	**0,1758 ± 0,0301**	0,5168 ± 0,0497	0,2519 ± 0,0427	0,1743 ± 0,1495	0,1480 ± 0,0564	*0,3273 ± 0,0753*
	Self-supervision	**0,5143 ± 0,0712**	**0,6345 ± 0,0567**	0,1574 ± 0,0306	**0,5853 ± 0,0667**	0,3066 ± 0,0321	0,2232 ± 0,1393	**0,1617 ± 0,0546**	***0,3690 ± 0,0693***
	Random	0,2134 ± 0,0397	0,2757 ± 0,0350	0,0432 ± 0,0289	0,1429 ± 0,0290	−0,0069 ± 0,0463	0,4269 ± 0,0500	0,0117 ± 0,0108	*0,1581 ± 0,0363*
	ImageNet frozen	0,2110 ± 0,0345	0,2800 ± 0,0321	0,0457 ± 0,0297	0,1391 ± 0,0429	−0,0041 ± 0,0486	**0,4499 ± 0,0390**	0,1446 ± 0,0664	*0,1809 ± 0,0414*
	Self-sup. frozen	0,4257 ± 0,0152	0,5408 ± 0,0322	0,0950 ± 0,0182	0,4736 ± 0,0147	**0,4516 ± 0,0217**	−0,1923 ± 0,1337	0,1603 ± 0,0521	*0,2792 ± 0,0575*
Weak supervision	ImageNet	0,4813 ± 0,0707	0,5856 ± 0,1052	0,4259 ± 0,0142	0,4413 ± 0,0920	0,2874 ± 0,0869	0,0018 ± 0,0217	0,3425 ± 0,1335	*0,3665 ± 0,1017*
	Self-supervision	**0,5165 ± 0,0712**	0,5678 ± 0,0957	0,4063 ± 0,1220	**0,5263 ± 0,0814**	**0,3680 ± 0,0786**	**0,1019 ± 0,1458**	0,3342 ± 0,1047	***0,4030 ± 0,1061***
	Random	0,2534 ± 0,0475	0,2794 ± 0,1052	0,0168 + 0,1778	0,0692 ± 0,0652	0,1488 ± 0,0599	0,0181 ± 0,0544	0,0492 ± 0,0508	*0,1193 ± 0,0913*
	ImageNet frozen	0,4197 ± 0,0148	**0,6391 ± 0,0401**	0,4296 ± 0,1508	0,4703 ± 0,0930	0,2100 ± 0,0466	−0,0126 ± 0,0186	0,1455 ± 0,1203	*0,3288 ± 0,0847*
	Self-sup. frozen	0,3070 ± 0,0317	0,4058 ± 0,0230	**0,5408 ± 0,0632**	0,2848 ± 0,0537	0,2865 ± 0,0352	−0,0769 ± 0,0957	**0,3550 ± 0,0635**	*0,3004 ± 0,0571*
Weak supervision + fine-tuning	ImageNet	0,4869 ± 0,0427	0,6361 ± 0,0586	**0,1661 ± 0,0631**	0,5843 ± 0,0474	0,2617 ± 0,0302	0,0499 ± 0,0557	0,1855 ± 0,0725	0,3386 ± 0,0555
	Self-supervision	**0,5347 ± 0,0401**	**0,6662 ± 0,0581**	0,1641 ± 0,0356	**0,6289 ± 0,0388**	0,3564 ± 0,0469	0,3436 ± 0,1153	0,2154 ± 0,0845	***0,4156 ± 0,0694***
	Random	0,2523 ± 0,0354	0,4093 ± 0,0329	0,1098 ± 0,0620	0,0664 ± 0,0649	0,1427 ± 0,0607	**0,3909 ± 0,1416**	0,0497 ± 0,0499	*0,2030 ± 0,0723*
	ImageNet frozen	0,4115 ± 0,0151	0,5948 ± 0,0979	0,1056 ± 0,1720	0,3125 ± 0,0464	0,2089 ± 0,0476	0,2715 ± 0,0679	0,1455 ± 0,1203	*0,2929 ± 0,0948*
	Self-sup. frozen	0,4287 ± 0,0298	0,5845 ± 0,0283	0,1089 ± 0,0100	0,5293 ± 0,0208	**0,4421 ± 0,0179**	−0,1132 ± 0,0632	**0,3209 ± 0,0178**	*0,3287 ± 0,0313*
Image- + patch-level annot.	ImageNet	0,4954 ± 0,0325	**0,6542 ± 0,1048**	0,3433 ± 0,1518	0,5088 ± 0,111	0,3840 ± 0,0834	0,3421 ± 0,1026	0,2592 ± 0,0847	*0,4267 ± 0,0998*
	Self-supervision	**0,5297 ± 0,0195**	0,6215 ± 0,0725	**0,4529 ± 0,1595**	0,4525 ± 0,0618	**0,3771 ± 0,0849**	**0,3431 ± 0,1014**	0,3550 ± 0,0634	***0,4474 ± 0,0937***
	Random	0,3313 ± 0,0814	0,1028 ± 0,0925	0,1324 ± 0,0710	0,0943 ± 0,1308	0,0778 ± 0,0471	0,1509 ± 0,0849	0,2652 ± 0,0702	*0,1650 ± 0,0859*
	ImageNet frozen	0,5231 ± 0,0271	0,5025 ± 0,0482	0,2927 ± 0,1047	0,5268 ± 0,0328	0,1910 ± 0,0213	0,0462 ± 0,0765	**0,4507 ± 0,0198**	*0,3618 ± 0,0558*
	Self-sup. frozen	0,4963 ± 0,0373	0,5397 ± 0,0615	0,2984 ± 0,0739	**0,5628 ± 0,0406**	0,1962 ± 0,0316	0,0225 ± 0,0144	0,4410 ± 0,0180	*0,3653 ± 0,0444*
Semi-supervision	ImageNet	0,3824 ± 0,0698	0,5998 ± 0,0667	0,1838 ± 0,0537	0,5361 ± 0,1010	**0,3522 ± 0,0441**	**0,0545 ± 0,0698**	0,3065 ± 0,0903	***0,3451 ± 0,0673***
	Self-supervision	0,4577 ± 0,0451	**0,6689 ± 0,0380**	0,1467 ± 0,0425	**0,5939 ± 0,0676**	0,3340 ± 0,0623	−0,1239 ± 0,1157	**0,3236 ± 0,0875**	*0,3430 ± 0,0710*
	Random	0,2345 ± 0,0405	0,2384 ± 0,0589	0,1269 ± 0,0882	0,0351 ± 0,0419	−0,0244 ± 0,0360	−0,0625 ± 0,0855	0,1057 ± 0,0454	*0,0934 ± 0,0601*
	ImageNet frozen	0,4777 ± 0,0235	0,6437 ± 0,0226	**0,3267 ± 0,0789**	0,3906 ± 0,0978	0,2031 ± 0,0472	0,0044 ± 0,0531	0,2845 ± 0,0590	*0,3329 ± 0,0602*
	Self-sup. frozen	**0,4800 ± 0,0365**	0,6662 ± 0,0204	0,3159 ± 0,0860	0,3546 ± 0,0574	0,1450 ± 0,0504	0,0097 ± 0,0565	0,3051 ± 0,0668	*0,3252 ± 0,0569*
Instance-based Additive-MIL	Random frozen	0,1768 ± 0,0634	0,0277 ± 0,0199	0,3243 ± 0,1850	0,0850 ± 0,0668	0,0522 ± 0,0283	0,0598 ± 0,0435	0,1800 ± 0,0928	*0,1294 ± 0,0882*
	ImageNet frozen	**0,6281 ± 0,0161**	0,6001 ± 0,0380	**0,7307 ± 0,0434**	**0,6576 ± 0,0199**	**0,5806 ± 0,0203**	**0,4941 ± 0,1810**	0,4208 ± 0,0378	***0,5874 ± 0,0742***
	Self-sup. frozen	0,5839 ± 0,0303	**0,6174 ± 0,0325**	0,6893 ± 0,0774	0,6516 ± 0,0304	0,5770 ± 0,0244	0,3437 ± 0,1196	**0,5292 ± 0,0118**	*0,5703 ± 0,0584*
Embedding-based Additive-MIL	Random frozen	0,1086 ± 0,0493	0,0170 ± 0,0125	0,1324 ± 0,1420	0,0610 ± 0,0469	0,0558 ± 0,0662	0,0426 ± 0,0194	0,0973 ± 0,0875	*0,0735 ± 0,0731*
	ImageNet frozen	**0,6342 ± 0,0168**	**0,6481 ± 0,0447**	**0,7406 ± 0,0426**	**0,6361 ± 0,0364**	0,5762 ± 0,0190	**0,5752 ± 0,0362**	0,4123 ± 0,0248	***0,6033 ± 0,0332***
	Self-sup. frozen	0,5985 ± 0,0148	0,5827 ± 0,0179	0,7154 ± 0,0464	0,6326 ± 0,0296	**0,5869 ± 0,0299**	0,4400 ± 0,0950	**0,5393 ± 0,0192**	*0,5850 ± 0,0445*
Instance-based AB-MIL	Random frozen	0,0859 ± 0,0600	0,0086 ± 0,0075	0,2237 ± 0,1328	0,0754 ± 0,0535	0,0551 ± 0,0320	0,0659 ± 0,0428	0,0985 ± 0,0833	*0,0876 ± 0,0696*
	ImageNet frozen	**0,6134 ± 0,0276**	0,5921 ± 0,0554	**0,6402 ± 0,1083**	0,6112 ± 0,0307	**0,5902 ± 0,0171**	**0,5414 ± 0,0364**	0,5297 ± 0,0309	***0,5883 ± 0,0522***
	Self-sup. frozen	0,5929 ± 0,0236	**0,6085 ± 0,0151**	0,5998 + 0,0208	**0,6889 ± 0,0256**	0,5184 ± 0,0222	0,3349 ± 0,1075	**0,6080 ± 0,0158**	*0,5645 ± 0,0450*
Embedding-based AB-MIL	Random frozen	0,1541 ± 0,0569	0,0056 ± 0,0137	0,3247 ± 0,1274	0,1269 ± 0,0617	0,0406 ± 0,0369	0,1034 ± 0,0471	0,1823 ± 0,0961	*0,1339 ± 0,0720*
	ImageNet frozen	**0,6090 ± 0,0233**	0,6305 ± 0,0328	**0,6836 ± 0,0955**	0,5724 ± 0,0429	**0,5910 ± 0,0163**	**0,5592 ± 0,0254**	0,5622 ± 0,0360	***0,6011 ± 0,0460***
	Self-sup. frozen	0,5831 ± 0,0248	**0,6309 ± 0,0155**	0,6286 ± 0,0396	**0,6949 ± 0,0162**	0,5464 ± 0,0255	0,3569 ± 0,1046	**0,6087 ± 0,0172**	*0,5785 ± 0,0456*
DS-MIL	Random frozen	0,1476 ± 0,0856	0,0244 ± 0,0254	0,3491 ± 0,1816	0,1521 ± 0,0718	0,0437 ± 0,0518	0,1268 ± 0,0863	0,1408 ± 0,0963	*0,1406 ± 0,0967*
	ImageNet frozen	0,4990 ± 0,0617	0,5197 ± 0,0675	0,5787 ± 0,1699	0,6389 ± 0,0374	0,5431 ± 0,0430	**0,5848 ± 0,0510**	0,5451 ± 0,0279	*0,5585 ± 0,0792*
	Self-sup. frozen	**0,5921 ± 0,0581**	**0,6295 ± 0,0200**	**0,7037 ± 0,0746**	**0,6471 ± 0,0164**	**0,5475 ± 0,0537**	0,2530 ± 0,1303	**0,6163 ± 0,0255**	***0,5699 ± 0,0656***
TransMIL	Random frozen	0,1195 ± 0,0689	0,0406 ± 0,0531	0,2361 ± 0,1482	0,0945 ± 0,0472	−0,0135 ± 0,0578	0,2302 ± 0,1605	0,1514 ± 0,0388	*0,1227 ± 0,0944*
	ImageNet frozen	**0,6308 ± 0,0150**	**0,6119 ± 0,0321**	**0,7132 ± 0,0820**	0,6250 ± 0,0270	**0,5687 ± 0,0351**	**0,5304 ± 0,0860**	0,5491 ± 0,0364	***0,6042 ± 0,0516***
	Self-sup. frozen	0,6100 ± 0,0231	0,5915 ± 0,0207	0,6765 ± 0,1056	**0,6479 ± 0,0256**	0,5530 ± 0,0263	0,4291 ± 0,0837	**0,6238 ± 0,0143**	*0,5903 ± 0,0543*
CLAM	Random frozen	0,1470 ± 0,0607	0,0664 ± 0,0924	0,2694 ± 0,1818	0,1064 ± 0,0680	0,0785 ± 0,0488	0,0887 ± 0,0435	0,1364 ± 0,1138	*0,1275 ± 0,0979*
	ImageNet frozen	**0,5514 ± 0,0515**	0,5912 ± 0,0758	0,7156 ± 0,1166	0,6402 ± 0,0374	0,5686 ± 0,0321	**0,6307 ± 0,0604**	0,5614 ± 0,0295	***0,6084 ± 0,0643***
	Self-sup. frozen	0,5495 ± 0,0264	**0,6204 ± 0,0214**	**0,7204 ± 0,0828**	**0,7004 ± 0,0219**	**0,5725 ± 0,0231**	0,3224 ± 0,1043	**0,6316 ± 0,0104**	*0,5882 ± 0,0535*

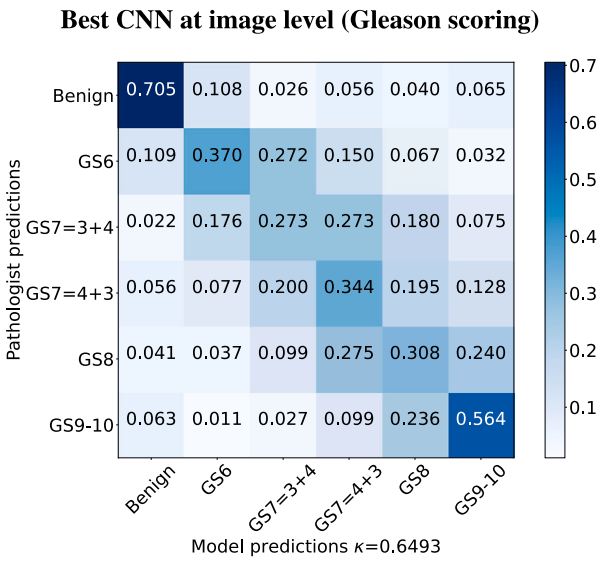


Fig. 8. Normalized confusion matrix obtained for the CNN model achieving the best image-level results (Gleason scoring). The results correspond to a CLAM model trained with self-supervised weights. The confusion matrix represents all the cores and WSIs in the test set, achieving a combined $\kappa = 0,6493$.

Clinic and Puerta del Mar). With this amount of both image-level and patch-level annotated data, different training methods were applied and evaluated with the main purpose of analyzing their performance on Gleason grading (patch-level) and Gleason scoring (image-level) and their generalization over a large test set obtained from many different sources.

Fully-supervised learning directly depends on pixel-wise annotations for training CNNs. As was previously mentioned, it is difficult to collect large heterogeneous datasets from different sources with local annotations. Thus, achieving high performance with this method is not easy, since the training cannot benefit from larger amounts of data with image-level annotations, as the rest of the methods do. On the other hand, in fully-supervised learning, training is less time-consuming and needs an inferior amount of resources due to this aspect. When analyzing the patch-level results (see Tables 7), it can be observed that this method achieves the highest performance on average across all the datasets in the test set, outperforming the rest of the training approaches. Moreover, as another positive note to these results, they were obtained with models that were only trained with patch-level annotated data (a total of 50'928 patches, as presented in Table 3), which is less than 4% of the data used to train the next simplest model among those evaluated (the models trained with the weakly-supervised learning approach used 1'377'608 image-level annotated patches in the training subset, as presented in Table 6). Therefore, fewer patches are needed to achieve even better results at the patch level compared to other methods. However, these have to be obtained from pixel-wise annotated data. When looking at the image-level results (see Table 8), the disadvantages of this method become clear. Fully-supervised learning has the lowest performances at the image level. There are only two cases in which this approach performs well, which is for TMA cores (TMAZ and Gleason challenge datasets) and Valme. Since the method achieves high performance at the patch level and only a few patches are densely-extracted from TMA cores due to their size, it makes sense that the method also performs well at the core level on these datasets. On the other hand, Valme was one of the datasets used for training the fully-supervised models, which could explain why the results of these models at the image level on Valme are not as low when compared to the rest of the datasets. The image-level results achieved for datasets such as SICAPv2 and TCGA-PRAD are poor, showing that this approach is not good at generalizing on many different datasets at this level.

As opposed to full supervision, weakly-supervised models were trained with image-level labels only. This training approach allows exploiting image-level annotations, which is an advantage over full supervision, since images obtained from public datasets commonly have image-level labels instead of pixel-wise annotations, which are more difficult to obtain. Therefore, in the weakly-supervised approach, we can expect to have many more data than in the fully-supervised approach, which is an advantage when training CNNs. As a counterpart, image-level data are not ground truth, since the label assigned to each patch is inherited from the most predominant GP in the TMA core or WSI. In order to filter the noise, different heuristics must be used, such as BR. This process does not increase the complexity of the model architecture itself, which is as simple as the one used for the full supervision, but adds an extra layer in the image preprocessing step, which is completely necessary for training the CNN with relevant data only. This, together with the fact that a larger amount of patches were used to train the network, makes this approach more time consuming than the previous one. In terms of patch-level results, weakly-supervised learning is one of the methods that performs the worst (together with instance-based MIL). This behavior is completely expected taking into account that the CNNs were trained with a large amount of incorrectly-labeled patches. Nevertheless, as can be seen in the image-level results, the higher heterogeneity of the data used to train the weakly-supervised models leads to a better generalization over all the datasets in the test subset, except for Clinic.

The second variant of the weakly-supervised learning approach consisted in transferring the weights from the first and fine-tuning the model with patch-level annotated data. This variant is more time-consuming than the previous ones, since a two-step training process is needed. Benefiting from both patch-level and image-level annotations is more convenient, since both can be exploited at the same time, using more data for training the models. An expected behavior prior to analyzing the results would be for the CNNs to learn the image-level generalization achieved with weak supervision while also acquiring part of the high-performance on Gleason grading learnt with full supervision after fine-tuning. However, this is not exactly what happens. As can be seen from the patch-level and image-level results, the transfer learning approach performs almost the same as the fully-supervised approach, being slightly worse at the patch level and slightly better at the image level. Although the former are high, the latter are not as good as in the weak supervision, meaning that the network mostly forgot how to perform the first task when it was fine-tuned for learning the second. This behavior is known as catastrophic forgetting (Goodfellow et al., 2013; Ramasesh et al., 2021).

The third and last variant of the weakly-supervised learning consisted in training CNNs with all the image-level and patch-level labels in the training set together at the same time. As opposed to the previous variant, this variant does not require a two-step learning process in which the network is first trained with image-level annotations and then fine-tuned with patch-level annotations. Instead, all these patches are combined and used to train the third best of all the evaluated approaches at the image level. In this case, this method used a total of 1'428'536 patches (where around 4% correspond to patch-level annotations) for training the models. This proportion could define the trade-off between performance at the image level and at the patch level, since increasing the amount of patch-level annotations would make this approach closer to full supervision, while the other way around would make it closer to the first weakly-supervised variant.

Among the different training approaches evaluated, semi-supervision with the teacher/student paradigm is the most complex in terms of resources and time needed. Two different models have to be trained: firstly, the teacher, with patch-level annotations only, which is used to predict unlabeled or image-level annotated data and generate new labels based on the prediction; and then the student, which is trained

on the predictions made by the teacher together with patch-level annotations. This whole process is slow and depends on many steps. However, as an advantage of this method when compared to the rest, it is the only method that can exploit unlabeled datasets (the teacher can make predictions on patches extracted from unlabeled images and assign a label to them). Semi-supervised learning achieves good performance at the patch level, being the third best among the six different approaches. At the image level, this approach is close to full supervision (slightly better), although it can still generalize better on some of the most complex datasets in the test set, such as TCGA-PRAD and PANDA.

All the discussed training approaches (full supervision, the three weak supervision variants and semi-supervision) share a particular disadvantage when performing image-level predictions. All of them are trained at the patch level without using the global annotation of the image in the process. Therefore they need an extra processing layer on top of the output of the network that aggregates all the patch-level predictions of an image into a single GS value. As is presented in Section 2.5, in this work we used majority voting for this purpose, which is not perfect and could lead to errors in the GS prediction, especially for those cases where the first and the second most predominant GPs in the image are the same. In our implementation, this limitation is partially solved by considering the image to have the same first and second GPs if the most represented pattern has at least twice the amount of patches as the second pattern. This threshold introduces errors in the prediction, which result in a lower performance of these methods at the image level. Although majority voting is not the best solution, as this problem could be addressed by means of some other complex AI-based approaches (Duran-Lopez et al., 2021; Campanella et al., 2019), it is the simplest.

Instance-based and embedding-based MIL approaches do not make use of the majority voting algorithm, since the architecture used consists of an attention model that aggregates patch-level predictions and embeddings, respectively. As a counterpart, the architecture of the models is not trivial, and the training step is not as straight-forward as in the rest of the methods. Since the GS label of the images is used in the training step, hundreds of images are needed to achieve robust and high-performance results. This is one of the main drawbacks of MIL, since the rest of the methods can exploit individual patch annotations when training, requiring a smaller amount of images in total. When analyzing the results at the patch level, it can be observed that the instance-based Additive-MIL models do not perform very well (they achieve the worst results at the patch level). Furthermore, embedding-based MIL models (including Additive-MIL, AB-MIL, DS-MIL, TransMIL and CLAM) are not even capable of reporting patch-level results due to their architecture. At the image level, the different MIL methods clearly achieve the best performance and generalization. Except for TMAZ, regarding which we already mentioned why full supervision, transfer learning, semi-supervision and even the third weakly-supervised variant achieved higher κ than the rest, the different MIL approaches outperform the rest of the methods in most of the datasets. Particularly, CLAM stands out when looking at Gleason, Clinic and PANDA results. This behavior was expected, since MIL approaches are optimized for good performance at the image level, unlike fully-supervised learning, which is optimized for good performance at the patch level.

While CLAM was the method that achieved the best performance at the image level, TransMIL, embedding-based AB-MIL and embedding-based Additive-MIL performed similarly on average. The difference between the performance of these 4 methods does not seem to be representative in this case. However, when looking at the performance on the PANDA dataset (the largest in the test set among the ones considered), CLAM seems to outperform the rest, particularly taking into account that the results reported are the average of 10 models and, thus, outliers are reduced.

The evaluation conducted both at the patch level and at the image level for each of the trained models consisted of a vast number of samples obtained from different datasets. Some of these datasets were used as part of an external test set in which samples from the same datasets were not used in the train and the validation partitions. This applies to Gleason challenge and Diagset datasets in the case of Gleason grading, which correspond to a total of 22'316 patches, representing around 65% of the test set and around 25% of the whole amount of patch-level annotations. For the Gleason scoring task, Gleason challenge, TCGA-PRAD and PANDA (images sourced from Radboud University Medical Center) datasets were used as external test set (half of PANDA was part of the training set, but only those images sourced from Karolinska Institutet and not from Radboud University Medical Center), which correspond to a total of 5'597 images, representing around 90% of the test set and around 45% of the whole amount of images with image-level annotations. Training with part of this external test set would have definitely improved the results and the generalization of all the different approaches evaluated in this work, but we preferred to test their limitations by having a vast test partition as external test set in order to have an unbiased evaluation of the different approaches. However, the great heterogeneity of the training partition allowed the different approaches to achieve similar results on the datasets that are part of the external test set compared to the rest of the test partition.

Initializing the models with random weights leads to the worst results among the different weight initialization alternatives considered, both at patch-level and at image-level results. Self-supervision generally improves both Gleason grading and Gleason scoring results on average for most of the methods considered. Particularly, this can clearly be seen when looking at image-level results on the PANDA dataset, which represent the largest amount of images in the test set. Therefore, self-supervised learning should definitely be used, instead of using pre-trained weights from ImageNet, in order to achieve higher performance on Gleason grading and scoring tasks.

5. Conclusions

In this work, a systematic comparison between different state-of-the-art methods for both Gleason grading and Gleason scoring classification is presented. These methods, which include fully-supervised, weakly-supervised, semi-supervised, Additive-MIL, AB-MIL, DS-MIL, TransMIL and CLAM learning approaches, were trained and evaluated using nine datasets from different sources collected from pathology workflows and publicly available repositories. The performance of the methods was analyzed, highlighting those reaching higher κ scores at Gleason grading and Gleason scoring, together with their advantages and drawbacks and their generalization capability over many different datasets. In particular, regarding Gleason grading, the models trained using the fully-supervised approach achieved the best performance, with their main limitation being the need for locally-annotated data, which are commonly scarce in publicly-available datasets. However, full supervision is the less time-consuming training approach, since the architecture is simple and it requires fewer images to train in order to achieve similar or better results than the other methods evaluated. On the other hand, in terms of Gleason scoring, models trained using MIL methods and, particularly, CLAM, reach higher performance and better generalization than the rest. MIL methods can exploit image-level annotations without the need for a patch-aggregation algorithm, since an attention model included in their complex architecture is dedicated to this task. As a counterpart, they need more images and a longer training process than the rest of the methods. Full supervision limits the use of heterogeneous data for training, reducing the generalization of the model in comparison to weak supervision and MIL. However, the former can reach higher performance with fewer images on internal data in patch-level classification. The impact of using models pre-trained with self-supervision showed a general improvement over those pre-trained with weights from ImageNet. The results presented in this work could guide researchers working on the automatic analysis of digitized histopathology images on which practices to adopt depending on the task to solve and the heterogeneity of the data.

CRediT authorship contribution statement

Juan P. Dominguez-Morales: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, visualization, Writing – original draft, Writing – review & editing, Resources. **Lourdes Duran-Lopez:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing, Resources. **Niccolò Marini:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Saturnino Vicente-Diaz:** Funding acquisition, Project administration, Supervision, Writing – review & editing, Resources. **Alejandro Linares-Barranco:** Funding acquisition, Project administration, Supervision, Writing – review & editing, Resources. **Manfredo Atzori:** Funding acquisition, Project administration, Resources, Supervision, Validation, Writing – review & editing. **Henning Müller:** Funding acquisition, Project administration, Resources, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was partially supported by PROMETEO (AT17_5410_USE) and DAFNE (US-1381619) projects, funded by the Junta de Andalucía, and by the MIND-ROB (PID2019-105556GB-C33) project, funded by the Spanish Ministry of Science, Innovation and Universities. L. Duran-Lopez and Juan P. Dominguez-Morales would like to thank Henning Müller and his group for hosting them during a two-month internship between 28th July 2021 and 28th September 2021, during which the idea of this paper was originated and most of the results presented in this work were obtained. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825292 "ExaMode".

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2024.103191.

References

Abels, E., Pantanowitz, L., Aeffner, F., Zarella, M.D., van der Laak, J., Bui, M.M., Vemuri, V.N., Parwani, A.V., Gibbs, J., Agosto-Arroyo, E., et al., 2019. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. J. Pathol. 249 (3), 286–294.

Altaf, F., Islam, S.M., Akhtar, N., Janjua, N.K., 2019. Going deep in medical image analysis: concepts, methods, challenges, and future directions. IEEE Access 7, 99540–99572.

Amin, M.B., Tickoo, S.K., 2016. Diagnostic Pathology: Genitourinary E-Book. Elsevier Health Sciences.

Anwar, S.M., Majid, M., Qayyum, A., Awais, M., Alnowami, M., Khan, M.K., 2018. Medical image analysis using convolutional neural networks: a review. J. Med. Syst. 42 (11), 1–13.

Arvaniti, E., Claassen, M., 2018. Coupling weak and strong supervision for classification of prostate cancer histopathology images. arXiv preprint arXiv:1811.07013.

Arvaniti, E., Fricker, K.S., Moret, M., Rupp, N., Hermanns, T., Fankhauser, C., Wey, N., Wild, P.J., Rueschoff, J.H., Claassen, M., 2018. Automated gleason grading of prostate cancer tissue microarrays via deep learning. Sci. Rep. 8 (1), 1–11.

Borley, N., Feneley, M.R., 2009. Prostate cancer: diagnosis and staging. Asian J. Androl. 11 (1), 74.

Bulten, W., Kartasalo, K., Chen, P.H.C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D.F., van Boven, H., Vink, R., et al., 2022. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. Nat. Med. 1–10.

Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., van der Laak, J., Hulsbergen-van de Kaa, C., Litjens, G., 2020. Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study. Lancet Oncol. 21 (2), 233–241.

Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A., 2020. Albumentations: Fast and flexible image augmentations. Information 11 (2), http://dx.doi.org/10.3390/info11020125, URL https://www.mdpi.com/2078-2489/11/2/125.

Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. 25 (8), 1301–1309.

Chan, J.K., 2014. The wonderful colors of the hematoxylin–eosin stain in diagnostic surgical pathology. Int. J. Surg. Pathol. 22 (1), 12–32.

Chang, H., Loss, L.A., Parvin, B., 2012. Nuclear segmentation in H&E sections via multi-reference graph cut (MRGC). In: International Symposium Biomedical Imaging.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. PMLR, pp. 1597–1607.

Chen, N., Zhou, Q., 2016. The evolving gleason grading system. Chin. J. Cancer Res. 28 (1), 58.

Chikontwe, P., Kim, M., Nam, S.J., Go, H., Park, S.H., 2020. Multiple instance learning with center embeddings for histopathology classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 519–528.

Ciga, O., Xu, T., Martel, A.L., 2022. Self supervised contrastive learning for digital histopathology. Mach. Learn. Appl. 7, 100198.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al., 2013. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J. Digit. Imaging 26 (6), 1045–1057.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20 (1), 37–46.

Dehaene, O., Camara, A., Moindrot, O., de Lavergne, A., Courtiol, P., 2020. Self-supervision closes the gap between weak and strong supervision in histology. arXiv preprint arXiv:2012.03583.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 248–255.

Deng, S., Zhang, X., Yan, W., Chang, E.I., Fan, Y., Lai, M., Xu, Y., et al., 2020. Deep learning in digital pathology image analysis: a survey. Front. Med. 14 (4), 470–487.

Doi, K., 2007. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Comput. Med. Imaging Graph. 31 (4–5), 198–211.

Duran-Lopez, L., Dominguez-Morales, J.P., Conde-Martin, A.F., Vicente-Diaz, S., Linares-Barranco, A., 2020. PROMETEO: A CNN-based computer-aided diagnosis system for WSI prostate cancer detection. IEEE Access 8, 128613–128628.

Duran-Lopez, L., Dominguez-Morales, J.P., Gutierrez-Galan, D., Rios-Navarro, A., Jimenez-Fernandez, A., Vicente-Diaz, S., Linares-Barranco, A., 2021. Wide & deep neural network model for patch aggregation in CNN-based prostate cancer detection systems. Comput. Biol. Med. 136, 104743.

Eskaros, A.R., Egloff, S.A.A., Boyd, K.L., Richardson, J.E., Hyndman, M.E., Zijlstra, A., 2017. Larger core size has superior technical and analytical accuracy in bladder tissue microarray. Lab. Invest. 97 (3), 335–342.

Farahani, N., Parwani, A.V., Pantanowitz, L., et al., 2015. Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. Pathol. Lab. Med. Int. 7 (23–33), 4321.

Foucart, A., Debeir, O., Decaestecker, C., 2019. SNOW: Semi-supervised, noisy and/or weak data for deep learning in digital pathology. In: 2019 IEEE 16th International Symposium on Biomedical Imaging. ISBI 2019, IEEE, pp. 1869–1872.

Goodfellow, I.J., Mirza, M., Xiao, D., Courville, A., Bengio, Y., 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks. arXiv preprint arXiv:1312.6211.

Hashimoto, N., Fukushima, D., Koga, R., Takagi, Y., Ko, K., Kohno, K., Nakaguro, M., Nakamura, S., Hontani, H., Takeuchi, I., 2020. Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3852–3861.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708.

Ilse, M., Tomczak, J., Welling, M., 2018. Attention-based deep multiple instance learning. In: International Conference on Machine Learning. PMLR, pp. 2127–2136.

Ilse, M., Tomczak, J.M., Welling, M., 2020. Deep multiple instance learning for digital histopathology. In: Handbook of Medical Image Computing and Computer Assisted Intervention. Elsevier, pp. 521–546.

Janowczyk, A., Zuo, R., Gilmore, H., Feldman, M., Madabhushi, A., 2019. HistoQC: an open-source quality control tool for digital pathology slides. JCO Clin. Cancer Inform. 3, 1–7.

Javed, S.A., Juyal, D., Padigela, H., Taylor-Weiner, A., Yu, L., Prakash, A., 2022. Additive mil: Intrinsically interpretable multiple instance learning for pathology. Adv. Neural Inf. Process. Syst. 35, 20689–20702.

Ke, Z., Wang, D., Yan, Q., Ren, J., Lau, R.W., 2019. Dual student: Breaking the limits of the teacher in semi-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6728–6736.

Koziarski, M., Cyganek, B., et al., 2021. DiagSet: a dataset for prostate cancer histopathological image classification. arXiv preprint arXiv:2105.04014.

Krupinski, E.A., Graham, A.R., Weinstein, R.S., 2013. Characterizing the development of visual search expertise in pathology residents viewing whole slide images. Hum. Pathol. 44 (3), 357–364.

Lai, Z., Wang, C., Oliveira, L.C., Dugger, B.N., Cheung, S.C., Chuah, C.N., 2021. Joint semi-supervised and active learning for segmentation of gigapixel pathology images with cost-effective labeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 591–600.

Lessells, A.M., Burnett, R.A., Howatson, S.R., Lang, S., Lee, F.D., McLaren, K.M., Nairn, E.R., Ogston, S.A., Robertson, A.J., Simpson, J.G., et al., 1997. Observer variability in the histopathological reporting of needle biopsy specimens of the prostate. Hum. Pathol. 28 (6), 646–649.

Li, B., Li, Y., Eliceiri, K.W., 2021a. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14318–14328.

Li, W., Li, J., Sarma, K.V., Ho, K.C., Shen, S., Knudsen, B.S., Gertych, A., Arnold, C.W., 2018. Path R-CNN for prostate cancer diagnosis and gleason grading of histological images. IEEE Trans. Med. Imaging 38 (4), 945–954.

Li, J., Li, W., Sisk, A., Ye, H., Wallace, W.D., Speier, W., Arnold, C.W., 2021b. A multi-resolution model for histopathology image classification and localization with multiple instance learning. Comput. Biol. Med. 131, 104253.

Li, Z., Yang, W., Peng, S., Liu, F., 2020. A survey of convolutional neural networks: analysis, applications, and prospects. arXiv preprint arXiv:2004.02806.

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., Tang, J., 2021. Self-supervised learning: Generative or contrastive. IEEE Trans. Knowl. Data Eng..

Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F., 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat. Biomed. Eng. 5 (6), 555–570.

Madabhushi, A., Lee, G., 2016. Image analysis and machine learning in digital pathology: Challenges and opportunities. Med. Image Anal. 33, 170–175.

Marini, N., Atzori, M., Otálora, S., Marchand-Maillet, S., Müller, H., 2021a. H&E-adversarial network: a convolutional neural network to learn stain-invariant features through Hematoxylin & Eosin regression. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 601–610.

Marini, N., Marchesin, S., Otálora, S., Wodzinski, M., Caputo, A., Van Rijthoven, M., Aswolinskiy, W., Bokhorst, J.M., Podareanu, D., Petters, E., et al., 2022a. Unleashing the potential of digital pathology data by training computer-aided diagnosis models without human annotations. NPJ Digit. Med. 5 (1), 1–18.

Marini, N., Otálora, S., Ciompi, F., Silvello, G., Marchesin, S., Vatrano, S., Buttafuoco, G., Atzori, M., Müller, H., 2021b. Multi-scale task multiple instance learning for the classification of digital pathology images with global annotations. In: MICCAI Workshop on Computational Pathology. PMLR, pp. 170–181.

Marini, N., Otálora, S., Müller, H., Atzori, M., 2021c. Semi-supervised training of deep convolutional neural networks with heterogeneous data and few local annotations: An experiment on prostate histopathology image classification. Med. Image Anal. 73, 102165.

Marini, N., Otálora, S., Podareanu, D., van Rijthoven, M., van der Laak, J., Ciompi, F., Müller, H., Atzori, M., 2022b. Multi_scale_tools: a python library to exploit multi-scale whole slide images. In: Data-Enabled Intelligence for Medical Technology Innovation, Volume I. Frontiers Media SA.

Matoso, A., Epstein, J.I., 2016. Grading of prostate cancer: past, present, and future. Curr. Urol. Rep. 17 (3), 1–6.

McLean, M., Srigley, J., Banerjee, D., Warde, P., Hao, Y., 1997. Interobserver variation in prostate cancer gleason scoring: are there implications for the design of clinical trials and treatment strategies? Clin. Oncol. 9 (4), 222–225.

Nagpal, K., Foote, D., Liu, Y., Chen, P.H.C., Wulczyn, E., Tan, F., Olson, N., Smith, J.L., Mohtashamian, A., Wren, J.H., et al., 2019. Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. NPJ Digit. Med. 2 (1), 1–10.

Niazi, M.K.K., Parwani, A.V., Gurcan, M.N., 2019. Digital pathology and artificial intelligence. Lancet Oncol. 20 (5), e253–e261.

Nir, G., Hor, S., Karimi, D., Fazli, L., Skinnider, B.F., Tavassoli, P., Turbin, D., Villamil, C.F., Wang, G., Wilson, R.S., et al., 2018. Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. Med. Image Anal. 50, 167–180.

Otálora, S., Atzori, M., Andrearczyk, V., Khan, A., Müller, H., 2019. Staining invariant features for improving generalization of deep convolutional neural networks in computational pathology. Front. Bioeng. Biotechnol. 198.

Otálora, S., Atzori, M., Khan, A., Jimenez-del Toro, O., Andrearczyk, V., Müller, H., 2020a. Systematic comparison of deep learning strategies for weakly supervised gleason grading. In: Medical Imaging 2020: Digital Pathology, vol. 11320, International Society for Optics and Photonics, p. 113200L.

Otálora, S., Marini, N., Müller, H., Atzori, M., 2020b. 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. Springer, pp. 193–203.

Otálora, S., Marini, N., Müller, H., Atzori, M., 2021. Combining weakly and strongly supervised learning improves strong supervision in gleason pattern classification. BMC Med. Imag. 21 (1), 1–14.

Pallua, J., Brunner, A., Zelger, B., Schirmer, M., Haybaeck, J., 2020. The future of pathology is digital. Pathol. Res. Pract. 216 (9), 153040.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32.

Peikari, M., Salama, S., Nofech-Mozes, S., Martel, A.L., 2018. A cluster-then-label semi-supervised learning approach for pathology image classification. Sci. Rep. 8 (1), 1–13.

Pulido, J.V., Guleria, S., Ehsan, L., Fasullo, M., Lippman, R., Mutha, P., Shah, T., Syed, S., Brown, D.E., 2020. Semi-supervised classification of noisy, gigapixel histology images. In: 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering. BIBE, IEEE, pp. 563–568.

Ramasesh, V.V., Lewkowycz, A., Dyer, E., 2021. Effect of scale on catastrophic forgetting in neural networks. In: International Conference on Learning Representations.

Rawla, P., 2019. Epidemiology of prostate cancer. World J. Oncol. 10 (2), 63.

Razzak, M.I., Naz, S., Zaib, A., 2018. Deep learning for medical image processing: Overview, challenges and the future. Classif. BioApps 323–350.

Salmo, E.N., 2015. An audit of inter-observer variability in Gleason grading of prostate cancer biopsies: The experience of central pathology review in the North West of England. Integr. Cancer Sci. Ther. 2 (2), 104–106.

Santos, M.K., Ferreira, J.R., Wada, D.T., Tenório, A.P.M., Barbosa, M.H.N., Marques, P.M.d., 2019. Artificial intelligence, machine learning, computer-aided diagnosis, and radiomics: advances in imaging towards to precision medicine. Radiol. Brasileira 52, 387–396.

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

Sellaro, T.L., Filkins, R., Hoffman, C., Fine, J.L., Ho, J., Parwani, A.V., Pantanowitz, L., Montalto, M., 2013. Relationship between magnification and resolution in digital pathology systems. J. Pathol. Inform. 4.

Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al., 2021. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Adv. Neural Inf. Process. Syst. 34, 2136–2147.

Shaw, S., Pajak, M., Lisowska, A., Tsaftaris, S.A., O'Neil, A.Q., 2020. Teacher-student chain for efficient semi-supervised histology image classification. arXiv preprint arXiv:2003.08797.

Silva-Rodríguez, J., Colomer, A., Sales, M.A., Molina, R., Naranjo, V., 2020. Going deeper through the Gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. Comput. Methods Programs Biomed. 195, 105637.

Srinidhi, C.L., Kim, S.W., Chen, F.D., Martel, A.L., 2022. Self-supervised driven consistency training for annotation efficient histopathology image analysis. Med. Image Anal. 75, 102256.

Ström, P., Kartasalo, K., Olsson, H., Solorzano, L., Delahunt, B., Berney, D.M., Bostwick, D.G., Evans, A.J., Grignon, D.J., Humphrey, P.A., et al., 2019. Pathologist-level grading of prostate biopsies with artificial intelligence. arXiv preprint arXiv:1907.01368.

Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.M., Ciompi, F., Van Der Laak, J., 2019. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Med. Image Anal. 58, 101544.

Tolkach, Y., Dohmgörgen, T., Toma, M., Kristiansen, G., 2020. High-accuracy prostate cancer pathology using deep learning. Nat. Mach. Intell. 2 (7), 411–418.

del Toro, O.J., Atzori, M., Otálora, S., Andersson, M., Eurén, K., Hedlund, M., Rönnquist, P., Müller, H., 2017. Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade gleason score. In: Medical Imaging 2017: Digital Pathology, vol. 10140, International Society for Optics and Photonics, p. 101400O.

van der Laak, J., Ciompi, F., Litjens, G., 2019. No pixel-level annotations needed. Nat. Biomed. Eng. 3 (11), 855–856.

Van der Laak, J., Litjens, G., Ciompi, F., 2021. Deep learning in histopathology: the path to the clinic. Nat. Med. 27 (5), 775–784.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wang, Y., Li, J., Metze, F., 2019. A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 31–35.

Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X., 2022. Transformer-based unsupervised contrastive learning for histopathological image classification. Med. Image Anal. 81, 102559.

Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1492–1500.

Yalniz, I.Z., Jégou, H., Chen, K., Paluri, M., Mahajan, D., 2019. Billion-scale semi-supervised learning for image classification. arXiv preprint arXiv:1905.00546.

Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., Huang, J., 2020. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. Med. Image Anal. 65, 101789.

Zhang, L., Amgad, M., Cooper, L.A., 2021. A histopathology study comparing contrastive semi-supervised and fully supervised learning. arXiv preprint arXiv:2111.05882.

Zhou, Y., Chen, H., Lin, H., Heng, P.A., 2020. Deep semi-supervised knowledge distillation for overlapping cervical cell instance segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 521–531.

Zuley, M.L., Jarosz, R., Drake, B.F., Rancilio, D., Klim, A., Rieger-Christ, K., Lemmerman, J., 2016. Radiology data from the cancer genome atlas prostate adenocarcinoma [tcga-prad] collection. Cancer Imag. Arch 9.