Computers in Biology and Medicine 159 (2023) 106856

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

A comparative study of the inter-observer variability on Gleason grading against Deep Learning-based approaches for prostate cancer

José M. Marrón-Esquivel [a,b,c,*], L. Duran-Lopez [a,b,c,d], A. Linares-Barranco [a,b,c,d], Juan P. Dominguez-Morales [a,b,c,d]

[a] Robotics and Tech. of Computers Lab., Universidad de Sevilla, 41012 Seville, Spain
[b] Escuela Técnica Superior de Ingeniería Informática (ETSII), Avenida de Reina Mercedes s/n, Universidad de Sevilla, 41012 Seville, Spain
[c] Escuela Politécnica Superior (EPS), Universidad de Sevilla, 41011 Seville, Spain
[d] Smart Computer Systems Research and Engineering Lab (SCORE), Research Institute of Computer Engineering (I3US), Universidad de Sevilla, 41012 Seville, Spain

ARTICLE INFO

Keywords:
Prostate cancer
Computational pathology
Deep Learning
Convolutional neural networks
Inter-observer variability
Medical image analysis

ABSTRACT

Background : Among all the cancers known today, prostate cancer is one of the most commonly diagnosed in men. With modern advances in medicine, its mortality has been considerably reduced. However, it is still a leading type of cancer in terms of deaths. The diagnosis of prostate cancer is mainly conducted by biopsy test. From this test, Whole Slide Images are obtained, from which pathologists diagnose the cancer according to the Gleason scale. Within this scale from 1 to 5, grade 3 and above is considered malignant tissue. Several studies have shown an inter-observer discrepancy between pathologists in assigning the value of the Gleason scale. Due to the recent advances in artificial intelligence, its application to the computational pathology field with the aim of supporting and providing a second opinion to the professional is of great interest.

Method: In this work, the inter-observer variability of a local dataset of 80 whole-slide images annotated by a team of 5 pathologists from the same group was analyzed at both area and label level. Four approaches were followed to train six different Convolutional Neural Network architectures, which were evaluated on the same dataset on which the inter-observer variability was analyzed.

Results : An inter-observer variability of 0.6946 κ was obtained, with 46% discrepancy in terms of area size of the annotations performed by the pathologists. The best trained models achieved $0.826 \pm 0.014\kappa$ on the test set when trained with data from the same source.

Conclusions: The obtained results show that deep learning-based automatic diagnosis systems could help reduce the widely-known inter-observer variability that is present among pathologists and support them in their decision, serving as a second opinion or as a triage tool for medical centers.

1. Introduction

Among all the pathologies that affect society, cancer is one of those in which the number of cases has increased the most. In 2020, 1.41 million new cases were diagnosed, representing the cause of around 375,000 deaths worldwide. It is known that prostate cancer is one of the most aggressive type of cancers that can be diagnosed [1]. According to GLOBOCAN [1], in countries with higher Human Development Index (HDI), life expectancy is higher and, consequently, higher incidences of cancer are recorded. This explains why Europe, with 9% of the world's population, accounts for 23% of the world's cancer cases [1]. Late diagnosis is a negative factor for the patient's prognosis, while early diagnosis greatly favors recovery and overcoming the pathology. The stages of cancer depend on the size of the tumor and how far it has spread throughout the rest of the body.

Cancer can be diagnosed on the basis of different medical tests performed on the patient. The following imaging tests may be requested to establish this diagnosis: Computed Tomography (CTI), Magnetic Resonance Imaging (MRI), Nuclear Tomography, Bone Scan, Positron Tomography (PET), Ultrasound, X-ray or Biopsy.

Histology is the branch of biology that studies the composition, structure and characteristics of the organic tissues of living beings. From a biopsy extracted from a patient, a pathologist can perform a histological study of the tissue and, based on its structure, report the details of the diagnosis. Prostate biopsy consists in obtaining samples of prostate tissue by means of a needle that punctures a region determined by a transrectal ultrasound process. These tissue samples are then processed in a laboratory and scanned, resulting in very high

* Corresponding author at: Robotics and Tech. of Computers Lab., Universidad de Sevilla, 41012 Seville, Spain.
E-mail address: jmarron@us.es (J.M. Marrón-Esquivel).

https://doi.org/10.1016/j.compbiomed.2023.106856
Received 30 November 2022; Received in revised form 7 February 2023; Accepted 30 March 2023
Available online 6 April 2023

resolution Whole-Slide Images (WSIs), which are subsequently analyzed and inspected by pathologists.

The aggressiveness of prostate cancer can be determined by a scoring system called Gleason Grading System (GGS) [2]. GGS scores a prostate cancer based on its histological appearance considering five different malignant cell patterns called Gleason patterns (GPs), which range from 1 to 5. Pathologists examine the structure of the cells in WSIs and assign a lower or higher GP to different malignant areas depending on how much they differ from healthy or normal tissue. The two most predominant patterns are summed up to assign the Gleason score (GS), which ranges from 2 to 10. However, scores of 2-5 are almost never used, since Gleason patterns 1 and 2 are very uncommon [3]. This score is subsequently used by the physician to determine the most appropriate treatment for the patient. However, many studies have reported inter-observer variability between pathologists in the process of labeling cancerous sections of tissue (more than 30% of discrepancy in Gleason scoring) [4–6].

Numerous research centers and hospitals have studied different approaches with the purpose of reducing the inter-observer variability among pathologists. In this regard, artificial intelligence has demonstrated potential to be used as a supporting service in diagnostic imaging tasks, such as radiology, dermatology and histopathology [7–9], among others. These systems, called Computer-Aided Diagnosis (CAD) systems, are automatic or semi-automatic algorithms with the purpose of supporting the professional when making a diagnosis or interpreting an image.

One of the most widely-used algorithms in artificial intelligence are Artificial Neural Networks (ANNs). ANNs are inspired by the operations performed by the human brain. These networks, like the human brain, receive information from the environment through a training process, where the synaptic weights store the acquired knowledge. Different types of ANNs can be found based on the type and degree of connections, as well as on the number of layers. Convolutional Neural Networks (CNNs) are a type of ANN that has become popular in recent years, since they are very effective for machine vision tasks, such as image classification and segmentation, among other applications. Recently, many researchers have studied the application of CNNs in the diagnosis of numerous types of diseases that involve image interpretation. Some works, such as [10–14], have demonstrated the potential of this kind of deep learning algorithms in many different visual pattern classification problems involving medical imaging.

Campanella et al. [15] developed a deep learning-based system to distinguish between cancer and non-cancer slides using more than 44000 WSIs from breast, skin and prostate tissue without pixel-wise annotations. In that paper, the authors presented a novel framework based on the multiple instance learning approach, which generates a semantically rich feature representation. A Recurrent Neural Network (RNN) is used to integrate the extracted information in order to report the final classification result, reporting an Area Under the Receiver Operating Characteristic Curve (AUC) of 0.986 for prostate cancer detection [15]. In [16], a CNN architecture is presented to distinguish between low (GS6-GS7) and high (GS8 to GS10) Gleason scores using 895 Tissue Microarrays (TMA) images. A total of 641 TMAs were used for training the CNN, which was then evaluated on an independent test set consisting of 245 TMAs annotated by two different pathologists. The authors report agreements of 0.75 and 0.71 (in terms of Cohen's quadratic kappa statistic) between the system and each pathologist, respectively, which were comparable with the agreement obtained between the pathologists (0.71). Strom et al. [17] presented two CNNs ensembles (each consisting of 30 InceptionV3 [18] models pre-trained on ImageNet [19]) to perform binary classification (benign or tumor) and GPs prediction. The authors used 6682 WSIs for training the system, 1631 WSIs for an independent test and 330 WSIs for an external validation. The system achieved an AUC of 0.997 (independent test) and 0.986 (external validation) for the binary task. Regarding the Gleason grading, the authors obtained a mean pairwise kappa of 0.62, which was within the range of the inter-observer variation between 23 pathologists (0.60-0.73).

In this work, the inter-observer variability of a group of 5 pathologists that annotated a dataset containing slides from Clinic Hospital in Barcelona was analyzed at different levels. Different deep learning architectures were trained using Prostate cANcer graDe Assessment (PANDA) [20], the largest publicly available dataset, and Clinic dataset in four different approaches, comparing the aforementioned variability with the performance of the models.

The main contributions of this work include the following:

- A study of the inter-observer variability of a team of pathologists from the same hospital where the WSIs were sourced was performed.
- A set of deep learning architectures were trained with four different methods including the largest publicly-available prostate cancer dataset and evaluated on the same dataset of which the inter-observer variability was analyzed.
- A total of 240 CNN models were evaluated and compared, including a broad discussion of the inter-observer variability analysis and the performance obtained by the neural networks.
- The best results were obtained with DenseNet121 models, which achieve a higher quadratic Cohen's kappa score (0.826 ± 0.014) than the inter-observer variability (0.6946), proving the viability of deep learning-based systems for supporting pathologists in the diagnosis.

The rest of the paper is structured as follows: Section 2 presents the materials and methods used, including the dataset (Section 2.1), the pre-processing applied to the images (Section 2.2), an introduction of the different CNN models used (Section 2.4), a brief description of the experiments performed (Section 2.5) and the metrics that were considered to evaluate the trained models (Section 2.6). Then, in Section 3, the results are presented, dividing them into those related to the inter-observer variability analysis (Section 3.1) and those related to the CNN models (Section 3.2). In Section 4, the results obtained are discussed, and, finally, the conclusions of this work are presented in Section 5.

2. Materials and methods

2.1. Dataset

In this work, a local dataset of pathological biopsy images obtained from prostate cancer patients from Clinic Hospital (Barcelona, Spain) was used. These cases consisted of different samples obtained by means of needle core biopsy and prepared with haematoxylin and eosin (H&R) stain in the laboratory. The samples were then digitized using a VENTANA iScan HT (Roche Diagnostics) scanner at 40× magnification (0.25 μm per pixel). A total of 80 different WSIs were obtained, which were then pixel-wise annotated by a team of five pathologists from the same hospital from which the samples were acquired. Since annotating WSIs with that level of detail is very time consuming, each of the pathologists annotated only around a third part of the total amount of images (each WSI was annotated by, at least, two pathologists). Pixel-wise annotations (also called strong annotations) were performed using digital graphic tablets and the QuPath software [21], labeling malignant tissue regions with Gleason patterns 3, 4 and 5. Although the five pathologists that annotated the WSIs work in the same team, they did not ask each other nor shared any information regarding the annotation process in order not to bias the inter-observer variability study performed in this work. Thus, this dataset was first used to analyze this aspect in a quantitative and statistical manner, and then for training different deep learning architectures in order to compare the results.

Since deep learning algorithms require a highly heterogeneous training dataset in order to generalize well on unseen data, another

Table 1
Distribution of the WSIs used from Clinic Hospital and PANDA challenge datasets.

Dataset	GS6	GS7 = 3+4	GS7 = 4+3	GS8	GS9-10	Total
Clinic Hospital	42	12	10	7	9	**80**
PANDA challenge	802	673	909	764	964	**4112**

dataset was used in combination with the one obtained from Clinic Hospital for this purpose. This was the Prostate cANcer graDe Assessment (PANDA) Challenge dataset [20], which is public and contains 11000 WSIs of digitized H&R-stained biopsies from Radboud University Medical Center (Nijmegen, Netherlands) and Karolinska Institutet (Stockholm, Sweden), of which 5060 images are pixel-wise annotated. This makes PANDA the largest publicly-available prostate cancer digital pathology dataset at present.

Table 1 summarizes the WSIs used from each dataset, specifying their corresponding ground-truth Gleason score.

2.2. Image pre-processing

WSIs are gigapixel-resolution images whose size can be greater than 1 GB. Current GPUs and neural networks are unable to process these images due to their memory limitations. Therefore, a pre-processing step was applied. A widely-known solution to overcome this problem is patch-sampling the WSIs, which consists in extracting smaller subimages, called patches, from the source images, thus allowing them to be used as input for neural networks. This process is the current and most widely-used method to work with WSIs in deep learning, and has been used in previous publications, such as [15,16,22–24], among many others. In this work, the size of the patches extracted were set to 750 × 750 pixels at 40× magnification, since it was previously used in other studies [16]. The patches were densely extracted, which means that no overlapping between them was set. Then, these were subsampled to 224 × 224 pixels in order to reduce computation and also due to the fact that it is the default input size in the pre-trained CNN models that were used (see Section 2.4).

Pathologists' annotations were used to delimit malignant areas within WSIs. Patches were only extracted from these areas, since they contained labeled regions of tissue corresponding to the three Gleason patterns considered in this work (GP 3-5). An 80% overlapping threshold with the annotations was set when extracting the patches from the WSIs, meaning that a patch had to overlap at least by that amount with an annotation in order to be considered for the dataset, discarding those with high background content and avoiding the addition of noisy information in the dataset. As a result, 17632 patches were obtained from the Clinic dataset, and 87824 from the PANDA dataset after applying the patch-sampling process.

The patches obtained were used to train and validate different CNN models, leaving part of the patches from Clinic to test them and compare the results with the inter-observer variability measured among the pathologists. Tables 2 and 3 show the training, validation and testing partitions used for the Clinic and the PANDA dataset, respectively, with their corresponding GP distribution. The partitions were carried out taking into account that all the patches obtained from the same patient were only involved in a single set.

The lack of standardization in the H&E staining process leads to color variations not only between images from different medical centers or digitized with different scanners, but also from the same source due to possible variations that may occur in the image preparation process [25,26]. Therefore, there is a tendency to alleviate this problem by means of stain normalization and color augmentation techniques. These techniques help deep learning algorithms focus on the relevant features of the images during the training step, while also homogenizing color variations that may be present among them. This is particularly important when working with images from different centers and scanners where a different H&E staining process was followed.

Table 2
Patch distribution used in the training, validation and test subsets with their corresponding GP from the Clinic dataset.

	GP3	GP4	GP5	*Total*
Train	3794	5424	2047	*11265*
Validation	1309	1718	728	*3755*
Test	657	833	1122	*2612*
Total	*5760*	*7975*	*3897*	*17632*

Table 3
Patch distribution used in the training and validation subsets with their corresponding GP from the PANDA challenge dataset.

	GP3	GP4	GP5	*Total*
Train	19954	38995	6019	*65868*
Validation	6604	13357	1995	*21956*
Total	26558	53252	8014	*87304*

Different techniques, such as Histogram Equalization (HE), Color Space Transformation (CST) and Color Deconvolution (CD) can be found in the literature.

2.2.1. Histogram equalization

The basic idea behind HE is to transform the intensity values of the pixels in an image so that the resulting image has a uniform distribution of intensities. The transformation is achieved by computing the cumulative distribution function (CDF) of the pixel intensities in the image and using it to map the original intensity values to new ones. The new intensity values are chosen such that the CDF of the new intensities is a linear function [27].

The result of this transformation is an image where the intensity values are spread out over a wider range, increasing the contrast of the image. However, histogram equalization can also result in the over-amplification of noise in the image, so it is important to use the technique with caution. To solve this problem, there are several adaptations of the method as seen in [28], where the use of Adaptive Histogram Equalization (AHE) and Contrast Limited Adaptive Histogram Equalization (CLAHE) is proposed.

2.2.2. Color space transformation

CST is based on changing the color space of an image, such as RGB to grayscale or HSV. After that, filters are applied to these color spaces and transformed back to RGB [29]. In [30], the authors developed a method to transform a source image into a target image in the *Lab* color space. This was achieved by calculating the mean and standard deviation for each channel. After completing the transformation, the normalized image was then converted back to the original RGB color space.

2.2.3. Color Deconvolution

CD is a technique that separates the contributions of different dyes or stains used in histological images. The goal of color deconvolution is to isolate the individual color channels in an image so that each component can be analyzed and processed independently.

Histological images are often stained with multiple dyes in order to highlight different structures within the tissue. For example, one dye may be used to stain the nuclei, while another is used to stain the cytoplasm. By separating the contributions of the different dyes, it is possible to better visualize and analyze the tissue.

Stain Color Adaptive Normalization Algorithm (SCAN) [31] is based on CD. It has been proposed as a solution to enhance the contrast between the histological tissue and the background while preserving the local structures in the image. This is achieved without altering the color of the lumen and the background.

In [32], an Adaptive Color Deconvolution algorithm is proposed for stain separation and color normalization of H&E-stained samples.

The process of normalization is accomplished using a uniform color transformation that maps pixels from the source image to the template image. This approach does not require the classification of stains. Instead, the parameters for color normalization are determined through an integrated optimization process that takes into account the distribution of pixel values. This results in the preservation of the structural information present in histological images.

On the other hand, it is possible to find models that use more than one technique, as in the case of [33], where a retinex model is designed that applies first the color space transformation and then the color deconvolution.

2.3. Data augmentation

Increasing the number of images and the heterogeneity of the dataset is a very relevant aspect to consider when training a CNN, since it makes the system more robust, improves its generalization over unseen data and prevent overfitting [34]. The relevance of applying data augmentation in computational pathology has been studied and proved in the literature [35]. Thus, data augmentation techniques have been applied to increase the number of images and the heterogeneity of the dataset during the training process. Different transformations were performed to the original patches, thus, for each training patch, horizontal and vertical flips were applied, along with 90 degrees rotations.

In order to tackle the problem of image and stain variability, training patches were augmented in color in their HSV (hue, saturation, value) representation within a specific limited range ([−15, 8] for the hue, [−20, 10] for the saturation and [−8, 8] for the value). Soft color augmentation has proven to be one of the most optimal approaches for tackling the stain variability problem [22,36]. Rotations, flips and color augmentation were performed automatically at training time with 50% probability for each of the mentioned processes. To this end, the open-source Albumentations library [37] was used.

2.4. Convolutional neural network models

Among all the existing types of neural networks, CNNs have proven to be one of the most accurate and successful algorithms for image analysis [38]. By means of convolution layers, the network is able to extract the main features of the images, which are then fed to a set of fully-connected layers in order to perform the classification. In addition to convolutions, CNNs consist of other types of layers that improve and speed up the learning and inference steps by reducing the amount of processed information.

In this work, different CNN models were trained and evaluated in order to compare their results with the inter-observer variability between pathologists in the Clinic dataset. Among them, the widely-known VGG16 [39], DenseNet121 [40] and InceptionV3 [18] were used. All these models were trained based on pre-trained weights from the Imagenet dataset [19]. Along with these, a Grid Search algorithm [41] was performed, in which different custom models containing from one convolution stage (convolution + pooling + activation layers) and a fully-connected layer up to a total of 10 convolutional stages were explored and evaluated, varying the number and size of the convolution filters. Custom models have shown improved performance for some specific cases compared to pre-trained ones in the literature [42].

The Adam optimizer [43] was used when training all of the models, considering different learning rates ranging from 1×10^{-3} to 1×10^{-6}, which varied depending on the model. This optimizer was selected based on the evolution of the training and validation losses.

In this work, TensorFlow[1] [44] version 2.2.0, which is a well-known Deep Learning Python library that allows designing, training and evaluating deep neural networks, was used for that purpose.

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

As can be observed in Tables 2 and 3, the patch distribution between the three different GPs is not balanced. This could make the training process focus on classes with a larger number of images when updating the weights of the network. This potential problem was avoided by using the *class_weights* parameter in TensorFlow, which makes the backpropagation algorithm to compensate classes during the training step based on the number of occurrences.

2.5. Training strategy and experiments

Different experiments were carried out in order to, firstly, analyze the inter-observer variability on the Clinic dataset and, then, compare the results with the performance of different CNN models.

Regarding the inter-observer variability analysis, two different experiments were carried out. The first, called Overlapping Annotated Area Analysis (OAAA), consisted in measuring the overlapping area of annotations by different pathologists corresponding to the same region of the slide. To this end, only those annotations that overlap on the same WSI were analyzed. The second experiment, called Labeling Discrepancy Analysis (LDA) follows an approach that is similar to that performed in the previous one. However, instead of measuring the inter-observer variability in terms of the area size of the annotated malignant tissue regions, an analysis regarding the label that was set for each of the annotations was performed.

Regarding the CNN experiments that were proposed and performed, the following different training approaches were considered:

- Experiment 1: Training and validating the models using Clinic dataset only.
- Experiment 2: Training and validating the models using PANDA dataset only.
- Experiment 3: Training and validating the models with PANDA dataset, and then fine-tuning the models using Clinic dataset (after applying transfer learning).
- Experiment 4: Training and validating using both PANDA and Clinic datasets combined.

The test partition in each of the experiments consisted of patches extracted from the Clinic dataset. For each of the experiments, six different CNN architectures were considered: three of them were obtained using Grid Search (in order to achieve faster and less complex models [42]), while the other three correspond to widely-known pre-trained networks, including VGG16, DenseNet121 and InceptionV3. Fig. 1 shows a block diagram explaining the experiments performed.

Transfer learning [45] is a well-known Deep Learning technique in which the feature-extraction layers (convolutional layers) of a previously-trained model are frozen, and the weights of the last layers are updated by training them with a different dataset (fine-tune), allowing the network to be adapted to a new dataset [46].

The partitions used in all the experiments are presented in Tables 2 and 3. As was previously mentioned, the test of the models was performed using the corresponding partition of Clinic dataset. Since PANDA was only considered to train and validate the models due to its size and heterogeneity, no testing partition can be seen in Table 3.

2.6. Evaluation metrics

Different evaluation metrics can be used to determine the effectiveness of neural networks. Among them, accuracy, specificity and sensitivity are some of the most used ones. The former reports a global idea of how the network performs, although it has a main drawback: it treats all classes as equal. This means that, in terms of accuracy, there is no difference on classifying a cheetah as a cat or as a dolphin (both would be considered as a misclassification of the network). Sensitivity and specificity are widely used in medical applications, but they are mainly useful for binary classification problems.

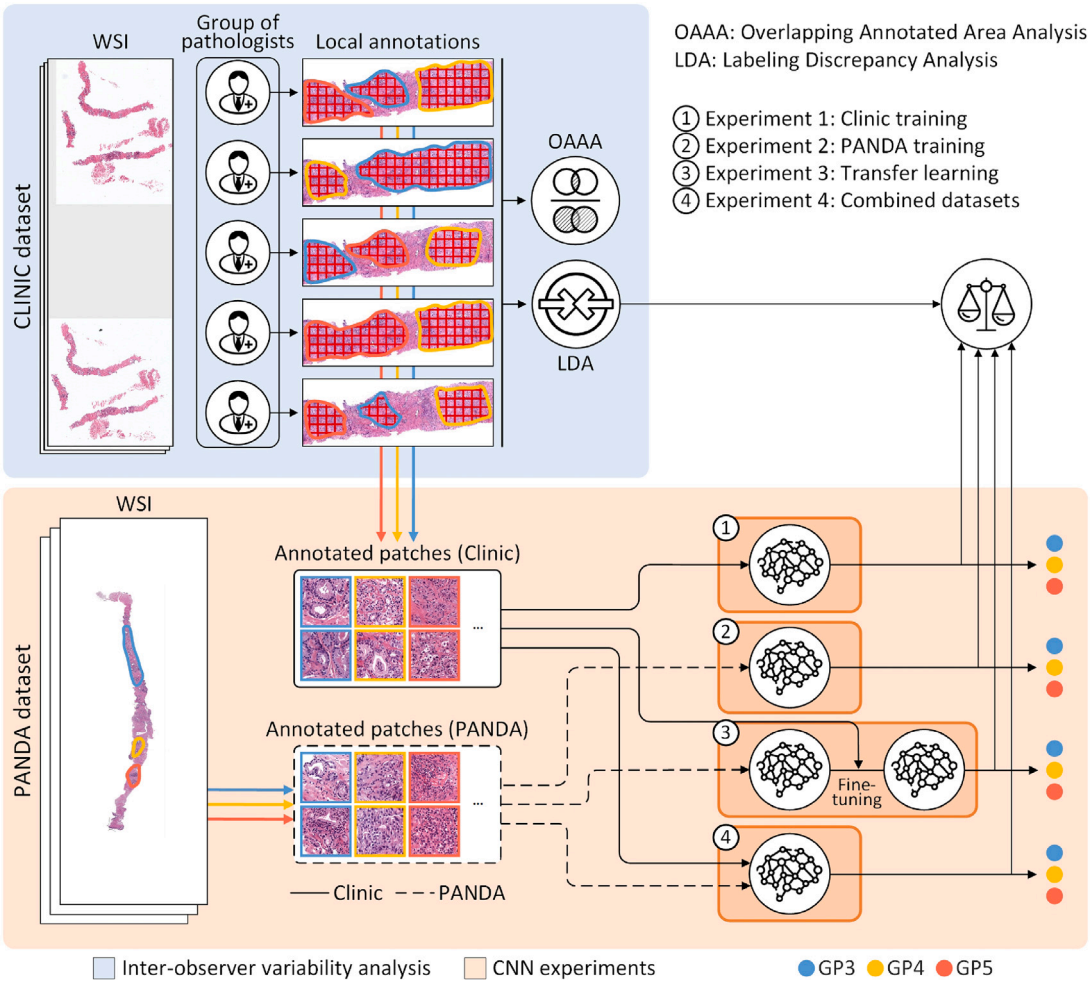


Fig. 1. Block diagram of the experiments performed. The upper part (blue box) shows the analysis of the annotations provided by 5 pathologists from the Clinic dataset, distinguishing between the overlapped annotated area analysis and the labeling discrepancy analysis. On the bottom part (orange box), four different training strategies performed with deep learning algorithms are depicted, using patches extracted from the annotated regions of the WSIs, obtained both from PANDA and Clinic datasets.

Therefore, to evaluate the performance of the models trained in this work, Cohen's kappa coefficient (κ) [47] was used. This metric measures the agreement or disagreement between two raters (which, in our case, are the annotations from pathologists and the predicted class from the network, or even the annotations that two pathologists assigned to the same tissue region). A value of 1 means a complete agreement of both raters. On the other hand, a score of 0 represents a random agreement. The quadratic version of the score was used (see Eq. (1)), which penalizes disagreements between the two raters depending on the class difference in a quadratic manner (instead of linearly, which would be the default κ). This score has been extensively used in previous works in the field of computational pathology [16,17,22,48,49]. That way, a disagreement between GP3 and GP5 would result in a stronger penalization in κ than that of a disagreement between GP3 and GP4.

$$\kappa = 1 - \frac{\sum_{i,j}^{k} w_{i,j} O_{i,j}}{\sum_{i,j}^{k} w_{i,j} E_{i,j}}, w_{i,j} = \frac{(i-j)^2}{(N-1)^2} \quad (1)$$

In Eq. (1), i and j are the CNN output classes, ranging from 0 to 2 for GP classification (0: GP 3, 1: GP 4, 2: GP 5; $N = 3$). $O_{i,j}$ is the multiclass confusion matrix, which is an $N \times N$ histogram representing the number of images that were classified with a specific pattern i by the first evaluator and j by the second. $E_{i,j}$ is an $N \times N$ matrix of expected results, i.e., a histogram with the expected number of images classified as i by the first evaluator and as j by the second. The weighted matrix, $w_{i,j}$, is calculated as a function of the difference between the true and predicted class, and it is used to penalize predictions more strongly the more different they are from the true value. In this matrix, the main diagonal is always 0, while the outer values of the anti-diagonal are 1. More information regarding quadratic weighted kappa can be found in scikit-learn's cohen_kappa_score function.[2] and in Data Science Bowl 2019 Evaluation page[3]

3. Results

The results of the experiments performed are divided into two main subsections: firstly, the inter-observer variability among pathologists of the Clinic dataset is studied and evaluated in Section 3.1; then, different deep learning models were trained and evaluated on the same dataset used in 3.1, and the results obtained using the aforementioned metrics (see Section 2.6) are presented in Section 3.2.

3.1. Inter-observer variability analysis

The inter-observer variability of the Clinic dataset was analyzed at two different levels. Firstly, the overlapping area of the annotated regions from the different pathologists was measured. Then, a comparison

[2] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html

[3] https://www.kaggle.com/c/data-science-bowl-2019/overview/evaluation

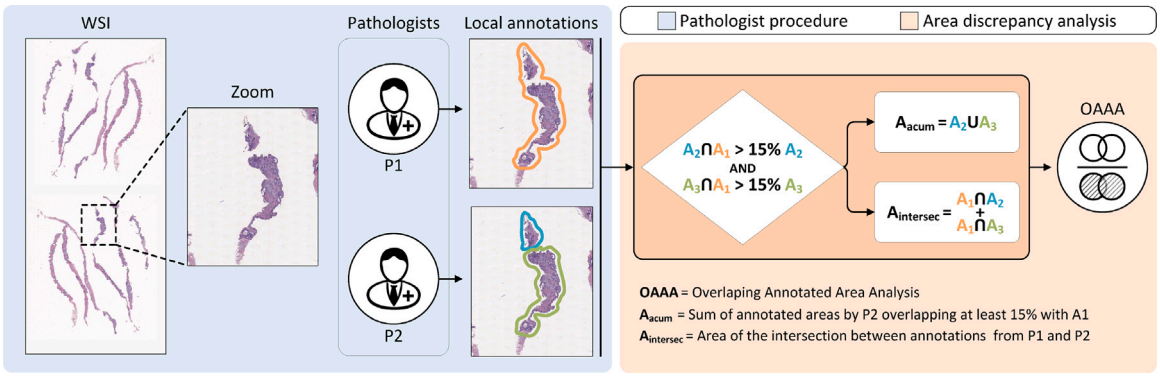


Fig. 2. Overlapping Annotated Area Analysis procedure. Annotations from two different pathologists (P1 and P2) that intersect at a minimum of 15% between them are used for the calculation (see Eq. (2)).

of the label of intersecting annotations from different pathologists was performed.

3.1.1. Overlapping annotated area analysis (OAAA)

As mentioned in Section 2.5, this experiment consisted in measuring the overlapping area of annotations by different pathologists corresponding to the same region of the slide. To this end, only those annotations that overlap on the same WSI were analyzed.

Some problematic cases can appear, such as those in which the same malignant region is annotated by two pathologists in complete different ways (e.g., one of them selects a large tissue region as GP3, while the other annotates smaller subregions inside the large one with the same label). In order to prevent these cases from affecting the results, the area of all the small subregions were summed up and then compared to the larger one, instead of performing a simple pairwise comparison. Moreover, a 15% overlap between annotation areas was set to avoid outliers.

The overlapping area ratio between two annotations was obtained by means of the intersection over union, taking into account the aforementioned (see Eq. (2)).

$$OAAA = (1 - \frac{A_{acum}}{A_1 + A_{acum} - A_{intersec}}) * 100 \quad (2)$$

Where A_{acum} is the sum of the areas annotated by pathologist P1 that intersect at a minimum of 15% with A_1 (a larger annotation performed by pathologist P2), and $A_{intersec}$ is the area of the intersection between the annotations. A total of 145 pairs of intersecting annotations were analyzed. Among these, only 3 pairs did not exceed the 15% overlapping threshold set. For each pair of annotations, expression (2) was applied. Fig. 2 shows the whole procedure followed for the OAAA calculation. As a result, a mean pairwise area discrepancy of 46% was obtained on overlapping annotations by different pathologists.

3.1.2. Labeling discrepancy analysis (LDA)

As mentioned in Section 2.5, an analysis regarding the label that was set for each of the annotations was performed. To this end, Cohen's kappa score was used (see Section 2.6). Two lists of GPs were defined to obtain the result, where index i refers to a tissue area annotated as $GP1[i]$ by one pathologist and as $GP2[i]$ by another pathologist.

The amount of pairs of annotations from the same tissue region that share the same label was calculated, which resulted in 74.34% agreement (25.66% discrepancy). The ground-truth confusion matrix obtained from the two aforementioned lists of annotations can be seen in Fig. 3. A total of 2116 pairs of annotations were analyzed, of which 543 (25.66%) did not share the same pattern. Consequently, a quadratic Cohen's kappa of 0.6946 was obtained.

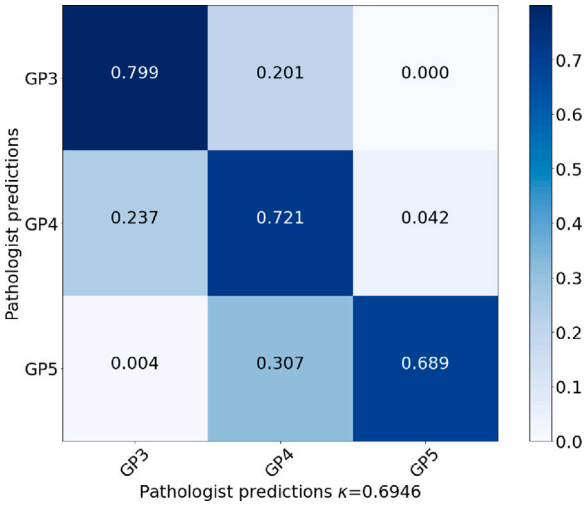


Fig. 3. Confusion matrix and kappa score of the ground-truth annotations provided by the group of pathologists from Clinic Hospital that annotated the slides.

Table 4
Test accuracy (*test_acc*) and Cohen's quadratic Kappa score (κ) of the 10 models designed by means of the Grid Search algorithm using Clinic dataset for training, validating and testing.

	test_acc	κ
IT 1	0.445	0.367
IT 2	0.501	0.429
IT 3	0.570	0.617
IT 4	0.487	0.609
IT 5	0.522	0.604
IT 6	0.636	0.727
IT 7	0.605	0.713
IT 8	0.658	0.739
IT 9	0.319	0
IT 10	0.319	0

3.2. CNN experiments results

As introduced in Section 2.5, different CNN models were trained using both datasets described in Section 2.1 in order to compare the results with the discrepancy of the group of pathologists that was evaluated in the previous experiment.

In order to report robust results, each of the networks was trained 10 times each. The results from each network architecture are reported as the mean and the standard deviation of the Cohen's quadratic kappa statistic (see Section 2.6).

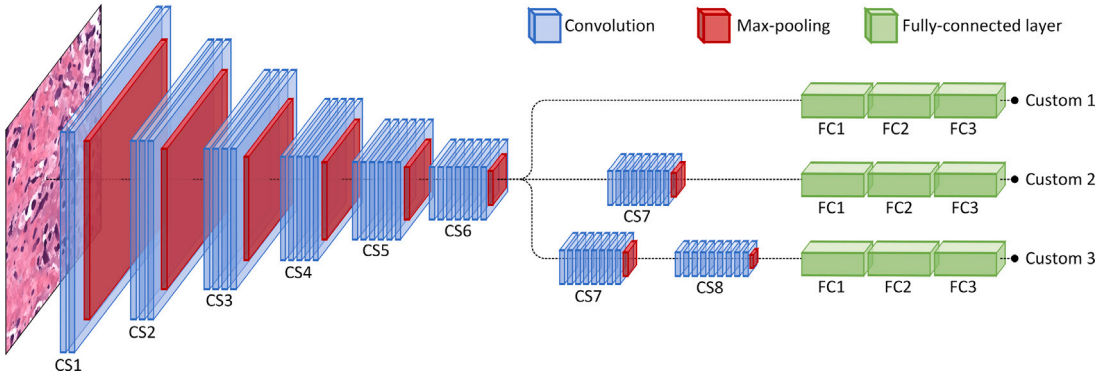


Fig. 4. Diagram of the architecture of the three custom models (Custom 1, Custom 2 and Custom 3) used. *CS* stands for Convolutional Stage, which consists of a convolution layer, a max-pooling layer and an activation layer.

Table 5
Kappa score achieved on the Clinic test set for the 10 models trained for each CNN architecture with the Clinic dataset. The mean and standard deviation are also reported for each architecture.

	VGG16	InceptionV3	DenseNet121	Custom 1	Custom 2	Custom 3
Model 1	0.680	0.779	0.845	0.727	0.669	0.735
Model 2	0.708	0.772	0.838	0.675	0.632	0.702
Model 3	0.692	0.808	0.833	0.779	0.733	0.749
Model 4	0.698	0.767	0.817	0.710	0.658	0.753
Model 5	0.653	0.813	0.811	0.698	0.751	0.714
Model 6	0.586	0.805	0.830	0.747	0.717	0.664
Model 7	0.795	0.818	0.796	0.644	0.626	0.741
Model 8	0.773	0.800	0.841	0.708	0.717	0.705
Model 9	0.680	0.794	0.832	0.746	0.694	0.736
Model 10	0.732	0.810	0.821	0.662	0.752	0.747
Mean	**0.700 ± 0.056**	**0.797 ± 0.017**	**0.826 ± 0.014**	**0.709 ± 0.040**	**0.695 ± 0.044**	**0.725 ± 0.027**

3.2.1. Clinic training

As was explained in Section 2.5, the first experiment regarding the use of CNNs consisted in training different architectures with the Clinic dataset and then evaluating them with an external partition of the same dataset. Table 4 presents the results obtained with the Grid Search algorithm. In each iteration, a convolution layer was added together with a batch normalization layer and a 2D max pooling layer. The best results were obtained in iterations 6 (0.727 κ), 7 (0.713 κ) and 8 (0.739 κ), which, from now on, will be referred to as Custom 1, Custom 2 and Custom 3 models, respectively. Fig. 4 shows a representation of the architecture of these three custom models.

These three models, together with VGG16, InceptionV3 and DenseNet121 were trained ten times each, and the results for each of them can be seen in Table 5.

Fig. 5 presents the confusion matrices for the best models of each of the trained architectures. As can be observed, all the models achieve very high accuracy on GP3 and GP5, which is not the case for GP4, as it is very often confounded with GP3.

3.2.2. PANDA training

The second experiment consisted in training each of the architectures proposed with the PANDA dataset, which were then tested on the Clinic dataset. Table 6 presents the results for each of the ten models trained for each architecture, along with the mean κ and standard deviation. As expected, the models achieved a lower performance with respect to the previous experiment, since the training and validation of the networks was performed on a dataset (PANDA) different than the one used for the final evaluation with which the metrics were obtained (Clinic).

co 6 presents the confusion matrices of the model that achieved the best results for each of the architectures considered. As can be seen, the models tend to classify GP3 as GP4 in most of the cases. It should also be mentioned that these matrices are obtained from the best models, which correspond to extreme positive outliers not representing the average case, as can be seen in Table 6.

3.2.3. Transfer learning

Since the results obtained in the previous experiment were not as good as expected, the same models were used in a transfer learning experiment in order to improve the results. To this end, the weights of the feature extraction layers of the trained models (trained only with the PANDA dataset) where frozen, and the fully-connected layers where fine-tuned with the train and validation partitions of the Clinic dataset.

The results obtained after testing the fine-tuned models with the test partition of the Clinic dataset are presented in Table 7. The best overall result was achieved by the VGG16 models, which obtained an average of $0.746 \pm 0.030\kappa$. Among them, the best model was able to achieve a κ of 0.789. A clear improvement can be observed on the results obtained for each of the architectures compared to the previous experiment.

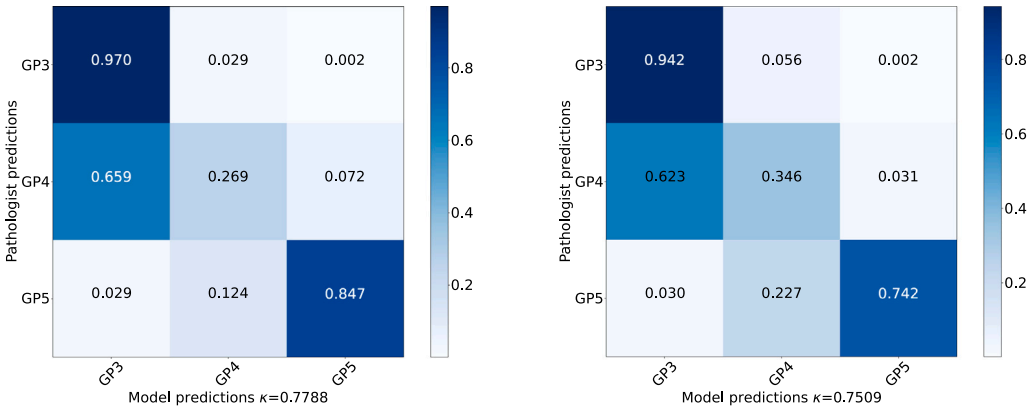
Fig. 7 presents the confusion matrices obtained with the best model of each of the evaluated architectures. A behavior similar to that of the first experiment is observed, in which the models tend to classify GP4 as GP3, even for those cases with the highest κ.

3.2.4. Clinic and PANDA datasets combined

In this subsection, the models have been trained with both datasets in order to improve the notorious confusion between GP3 and GP4 found in the predictions of the models. In this case, no layers were frozen, and all of them were trained from scratch with both datasets at the same time (see Fig. 8 and Table 8).

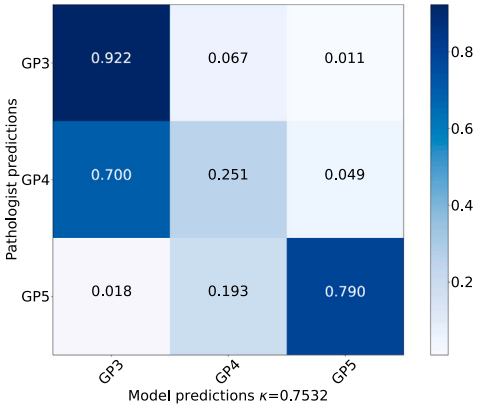
4. Discussion

The inter-pathologist variability is widely-known in the computational pathology field and, particularly, in prostate cancer classification. The high heterogeneity of the digitized tissue samples, the lack of very precise rules to follow when choosing the specific GP to assign to a tissue region (the pattern is assigned based on the extent to which the tissue resembles native tissue) and the subjectivity of the pathologists that perform the annotations are some of the main factors that increase

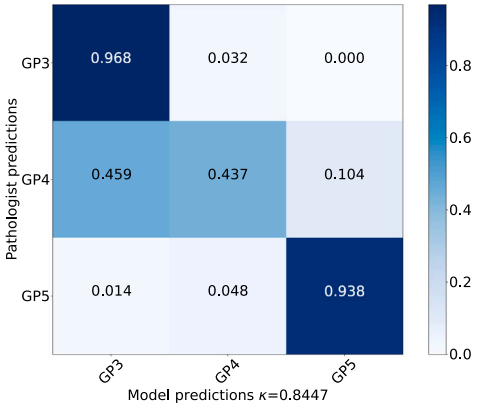


(a) Custom 1 model.

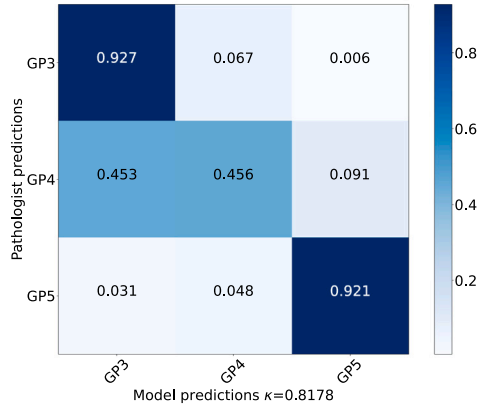
(b) Custom 2 model.



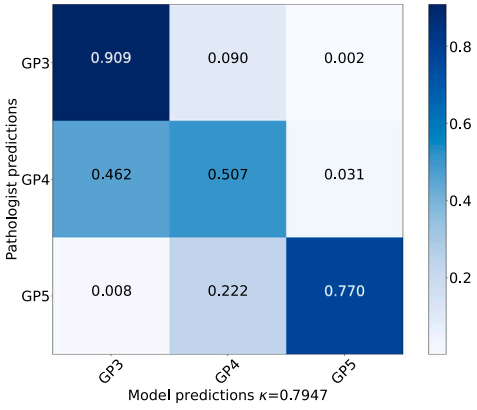
(c) Custom 3 model.



(d) DenseNet121 model.



(e) InceptionV3 model.



(f) VGG16 model.

Fig. 5. Confusion matrices of the best models trained with Clinic dataset.

the aforementioned variability. This, together with the recent advances in artificial intelligence, suggests that deep learning algorithms could be used as a support system for pathologists in order to help them in the analysis task and to reduce the inter-observer variability.

In this work, different experiments were performed in order to analyze the inter-observer variability of a local dataset from Clinic Hospital (Barcelona, Spain) that was annotated by a group of 5 pathologists, which was then compared to four different deep learning-based training approaches including 6 different CNN architectures.

In Section 3.1.1, the size of the tissue areas annotated by the pathologists was analyzed. To this end, a pairwise comparison between coincident annotated regions was performed, resulting on an average discrepancy of 46% in size. Such high discrepancy can be attributed to the subjectivity of annotating WSIs by hand in a manual process, in which a pathologist may have been very strict in making the annotations while others may have annotated in a more general way without being too specific. This also includes cases where a pathologist may have marked a whole tissue region as a region of interest, while others may have annotated a set of smaller tissue areas within that same region. These cases are considered in the OAAA calculation, since smaller annotations from a pathologist that overlap with a larger one from a different pathologist are summed up before calculating the

Table 6
Kappa score achieved on the Clinic test set for the 10 models trained for each CNN architecture with the PANDA dataset. The mean and standard deviation are also reported for each architecture.

	VGG16	InceptionV3	DenseNet121	Custom 1	Custom 2	Custom 3
Model 1	0.454	0.526	0.389	0.499	0.427	0.082
Model 2	0.450	0.552	0.516	0.195	0.021	0.084
Model 3	0.501	0.483	0.483	0.162	0.101	0.186
Model 4	0.512	0.385	0.446	0.424	0.213	0.046
Model 5	0.539	0.432	0.563	0.142	0.236	0.409
Model 6	0.584	0.456	0.420	0.062	0.372	0.225
Model 7	0.484	0.204	0.469	0.020	0.158	0.018
Model 8	0.470	0.576	0.530	0.333	0.288	0.381
Model 9	0.529	0.579	0.420	0.389	0.358	0.336
Model 10	0.634	0.560	0.529	0.298	0.276	0.083
Mean	**0.516 ± 0.055**	**0.475 ± 0.109**	**0.476 ± 0.055**	**0.252 ± 0.152**	**0.245 ± 0.120**	**0.185 ± 0.139**

Table 7
Kappa score achieved on the Clinic test set for the 10 models trained for each CNN architecture with the PANDA dataset and then fine-tuned with Clinic. The mean and standard deviation are also reported for each architecture.

	VGG16	InceptionV3	DenseNet121	Custom 1	Custom 2	Custom 3
Model 1	0.755	0.737	0.725	0.584	0.516	0.523
Model 2	0.763	0.739	0.729	0.539	0.632	0.586
Model 3	0.770	0.742	0.727	0.579	0.558	0.492
Model 4	0.789	0.706	0.734	0.547	0.603	0.626
Model 5	0.713	0.723	0.739	0.542	0.632	0.603
Model 6	0.727	0.712	0.735	0.580	0.641	0.558
Model 7	0.690	0.752	0.732	0.564	0.564	0.591
Model 8	0.747	0.691	0.730	0.583	0.532	0.373
Model 9	0.724	0.727	0.713	0.606	0.570	0.584
Model 10	0.783	0.703	0.740	0.640	0.560	0.354
Mean	**0.746 ± 0.030**	**0.723 ± 0.019**	**0.731 ± 0.007**	**0.579 ± 0.028**	**0.581 ± 0.042**	**0.529 ± 0.091**

Table 8
Kappa score achieved on the Clinic test set for the 10 models trained for each CNN architecture with the PANDA and Clinic datasets combined. The mean and standard deviation are also reported for each architecture.

	VGG16	InceptionV3	DenseNet121	Custom 1	Custom 2	Custom 3
Model 1	0.685	0.666	0.739	0.783	0.763	0.761
Model 2	0.562	0.729	0.807	0.802	0.728	0.603
Model 3	0.837	0.767	0.722	0.601	0.794	0.694
Model 4	0.610	0.742	0.751	0.754	0.798	0.738
Model 5	0.775	0.708	0.751	0.760	0.668	0.607
Model 6	0.851	0.758	0.770	0.822	0.724	0.631
Model 7	0.814	0.719	0.743	0.762	0.756	0.627
Model 8	0.798	0.674	0.729	0.606	0.805	0.753
Model 9	0.810	0.754	0.686	0.780	0.736	0.724
Model 10	0.773	0.764	0.762	0.781	0.610	0.721
Mean	**0.751 ± 0.094**	**0.728 ± 0.034**	**0.746 ± 0.030**	**0.745 ± 0.074**	**0.738 ± 0.058**	**0.686 ± 0.059**

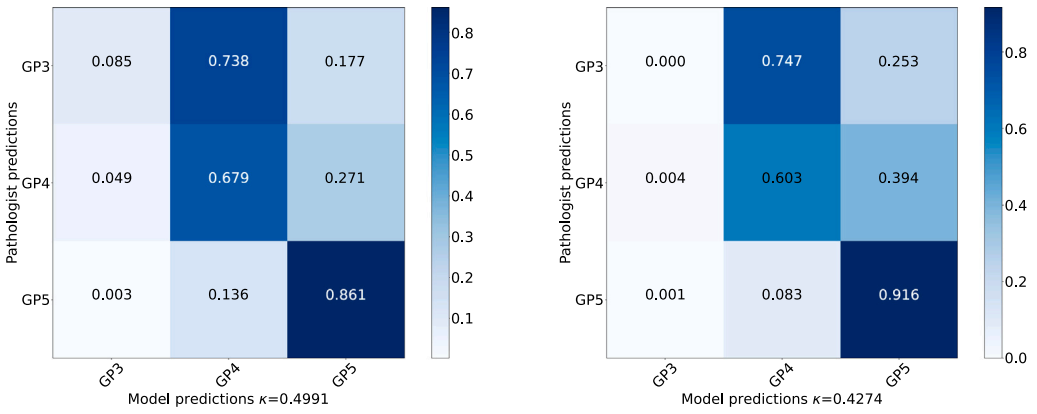
discrepancy between them, although all the stroma and benign cells included in the larger annotation would be considered, representing a decent increase of that discrepancy.

On the other hand, with respect to the inter-observer variability analysis regarding the labels of the annotations performed in Section 3.1.2, the result obtained (0.6946 κ between pathologists) is consistent with that presented in [16], which reports a similar κ value analyzed on TMAs instead of WSIs.

Regarding the CNN experiments presented in Section 3.2, four different training approaches were considered in order to compare the performance obtained with the inter-observer variability that was previously analyzed. These experiments were aimed at comparing the different training methods in terms of patch-level results on the test set.

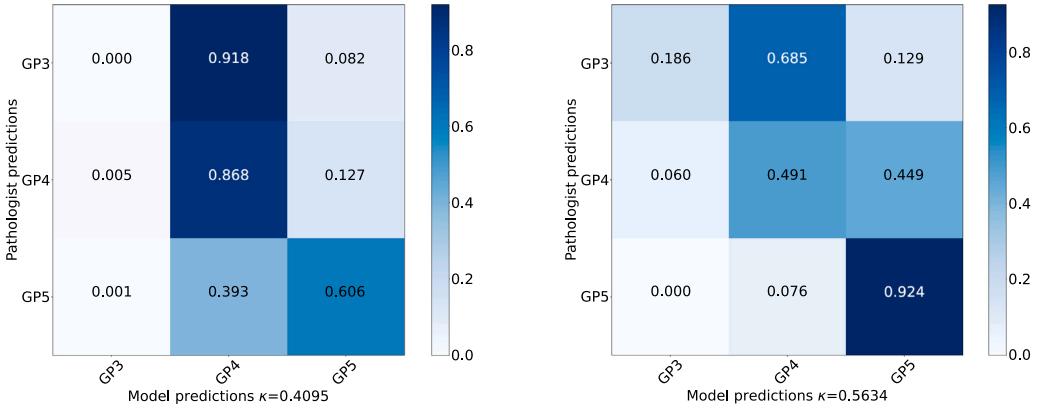
Firstly, in Section 3.2.1, six different neural network models were trained with strongly-annotated patches extracted from the Clinic dataset. Among them, three widely-known CNN architectures (VGG16, DenseNet121 and InceptionV3) were used, together with the three best custom models obtained by means of a Grid Search algorithm, which automatically explored different architectures from one convolution stage up to ten with different hyperparameters and filter sizes. In Table 4, it can be observed that the models from iteration 8 onward, which contain 9 and 10 convolution stages, are not functional. Each convolution stage reduces the size of the feature maps, making them very small and lose relevant information after 8 consecutive convolution stages, which explains why the fully-connected layers are no longer able to classify the extracted features correctly. On the other hand, iterations 6, 7 and 8 report the best results. These three architectures together with VGG16, DenseNet121 and InceptionV3 were trained ten times each, reporting the mean and the standard deviation of the results. This approach is commonly followed to reduce undesired effects introduced by the stochastic gradient descent optimizer adopted during the model optimization. From the results reported in Table 5, it can be seen that the best overall results were obtained by the DenseNet121 models with $\kappa = 0.826 \pm 0.014$, while the best result was obtained by one of the InceptionV3 models reporting $\kappa = 0.845$. Although these results were already higher than the inter-observer variability analyzed on the Clinic dataset ($\kappa = 0.6946$), other training methods were explored in order to avoid biased results obtained from models that were trained and tested on the same dataset. To this end, the largest publicly-available prostate cancer dataset was used to train the models in different ways, which allows for a higher generalization of the model due to the heterogeneity of the dataset.

The results of the experiment in which the networks were trained with PANDA and tested on Clinic (see Section 3.2.2) show a decrease



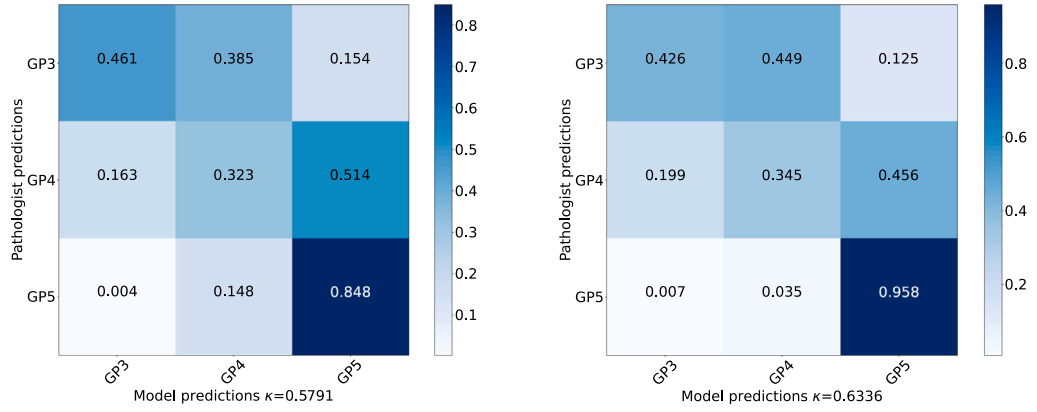
(a) Custom 1 model.

(b) Custom 2 model.



(c) Custom 3 model.

(d) DenseNet121 model.



(e) InceptionV3 model.

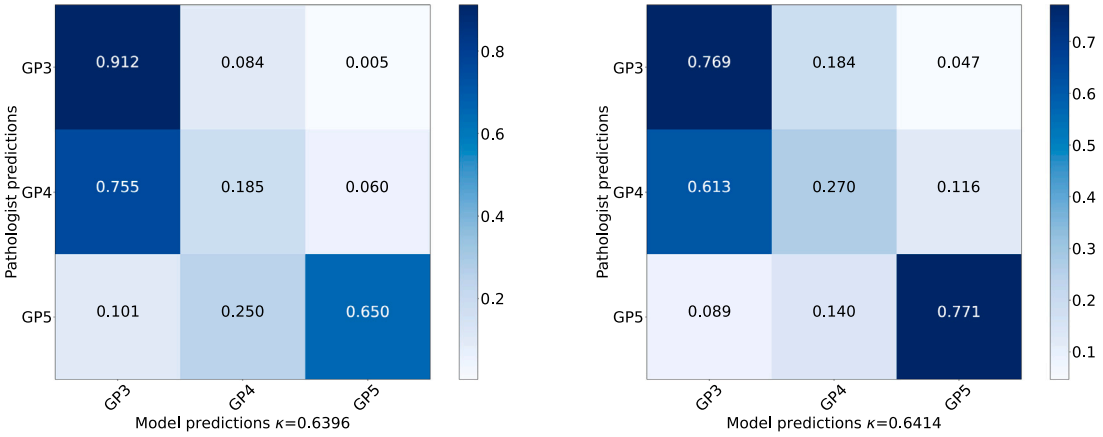
(f) VGG16 model.

Fig. 6. Confusion matrices of the best models trained with PANDA dataset.

on the performance of the models in terms of κ compared to the first experiment, where Clinic was used for both training and testing. These results were expected due to the high heterogeneity of the data and the stain variability. In these cases, models tended to classify GP3 as GP4 or vice versa in most cases. Nevertheless, complex models like VGG16 managed to overcome this aspect and achieve decent average results.

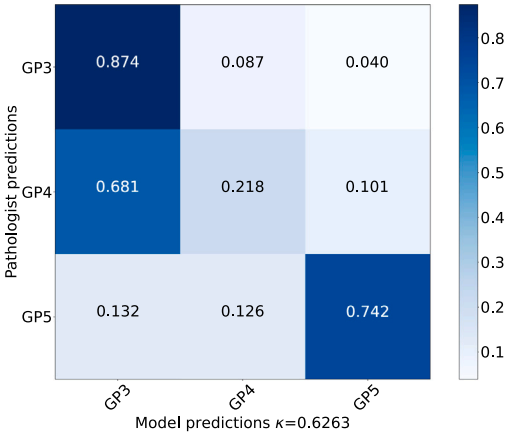
Additionally, in Section 3.2.3, the models trained with PANDA were used in a transfer learning experiment in which they were fine-tuned using the Clinic dataset. This approach allows having a more versatile and generalized model that could be used for a specific medical center after a simple and fast retraining of the last fully-connected layers. The results obtained using this method (see Table 7) show that the custom models did not perform as well as in the first experiment (see Section 3.2.1). One of the reasons that may explain this situation is the simplicity of the models, which are much smaller than VGG16, DenseNet121 and InceptionV3 and, thus, the feature extraction from input images is not as robust as in these models. On the other hand, the PANDA dataset has a much larger number of samples and is more heterogeneous than Clinic, which also explains the need for more complex architectures. The best results were obtained with VGG16, reporting an average result of $\kappa = 0.746 \pm 0.030$, with one of the models achieving $\kappa = 0.789$.

As an additional experiment, instead of training the models with either Clinic or PANDA (also considering fine-tuning), both datasets were combined in order to increase the variability of the training data. The results show a clear overall improvement when compared to

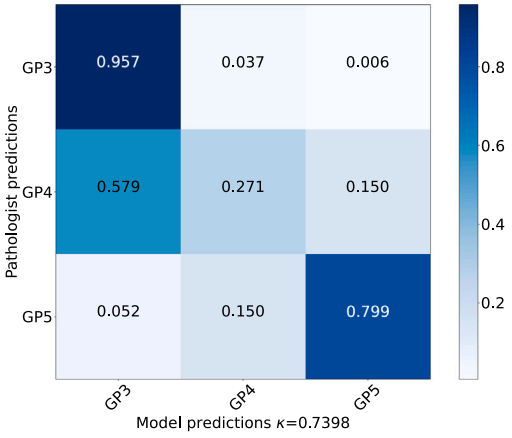


(a) Custom 1 model.

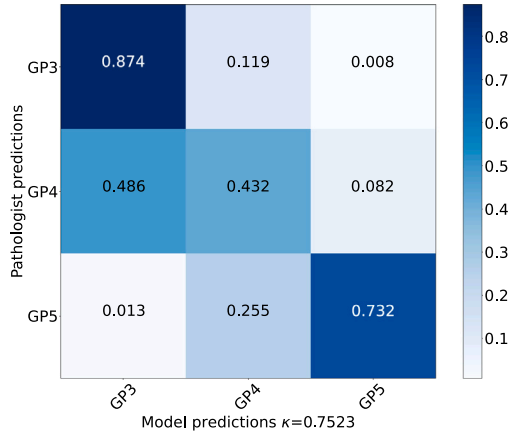
(b) Custom 2 model.



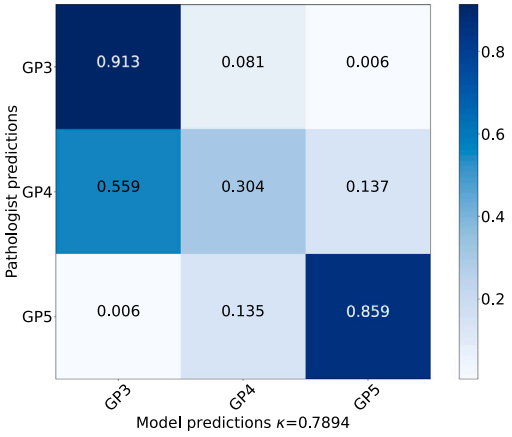
(c) Custom 3 model.



(d) DenseNet121 model.



(e) InceptionV3 model.

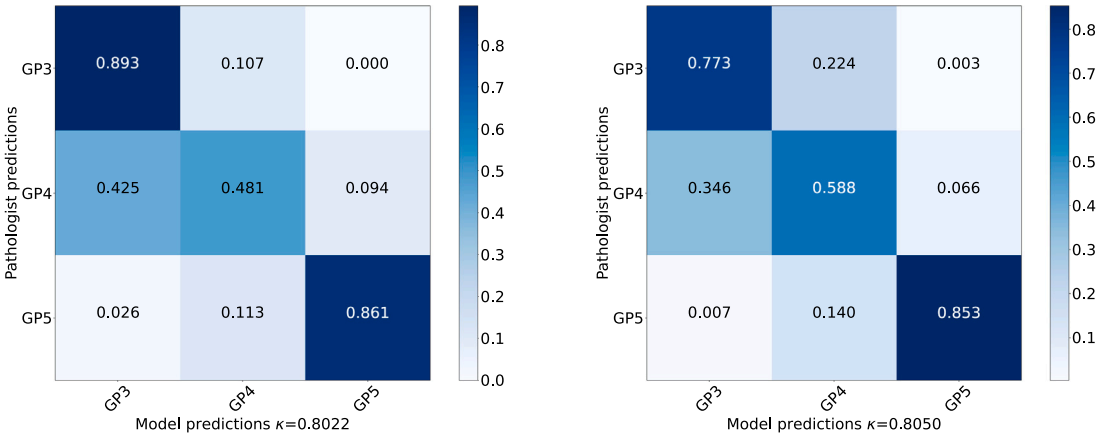


(f) VGG16 model.

Fig. 7. Confusion matrices of the best models trained with PANDA dataset and fine-tuned with Clinic dataset.

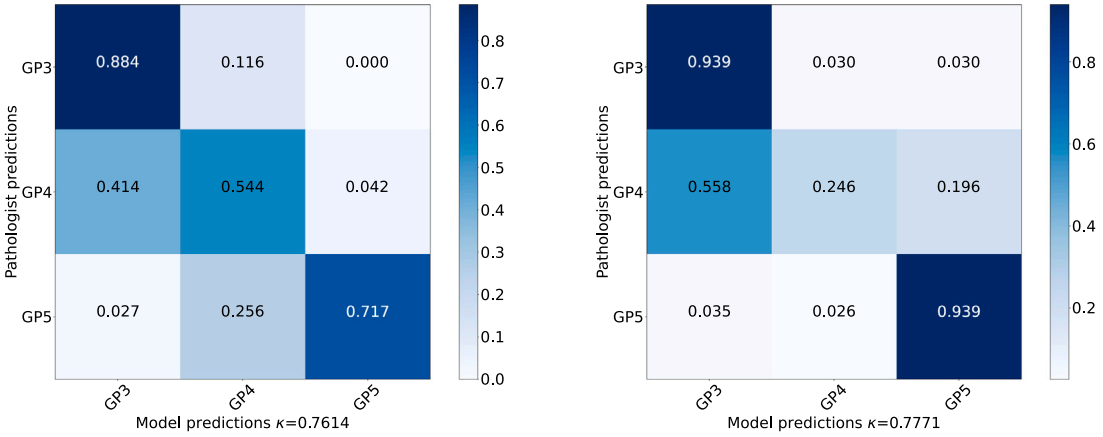
the previous experiment, particularly for the custom models. The best results are obtained with VGG16, which achieve $\kappa = 0.751 \pm 0.094$ on average, closely followed by Custom 1 models ($\kappa = 0.745 \pm 0.074$). More publicly-available datasets could have been added on this experiment in order to create models able to generalize better on unseen data. Nevertheless, achieving the best results or the highest generalization was not the main goal of this work, but to compare a few approaches with a test set with which the results could be compared to the inter-observer variability of a group of 5 pathologists from the same medical center. To this end, we considered using PANDA, the largest publicly-available prostate cancer dataset, for the proposed experiments.

Table 9 shows a general overview of the results obtained for each of the experiments performed, together with the mean κ and its standard



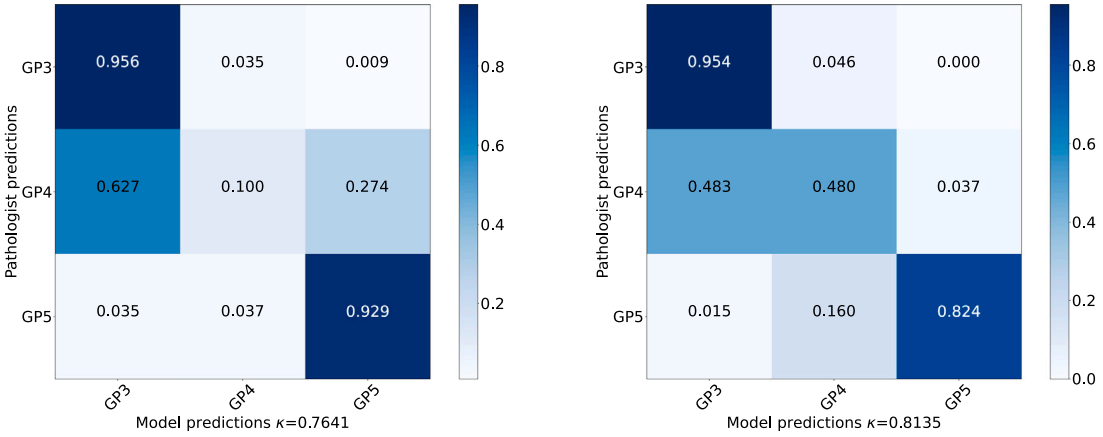
(a) Custom 1 model.

(b) Custom 2 model.



(c) Custom 3 model.

(d) DenseNet121 model.



(e) InceptionV3 model.

(f) VGG16 model.

Fig. 8. Confusion matrices of the best models trained with PANDA and Clinic datasets combined.

deviation. As can be seen, widely-known architectures achieve better results than the custom models, which sacrifice on performance while benefiting from faster computation time.

An interesting aspect to note is the fact that custom models obtained by means of the Grid Search method performed in a similar way as deeper pre-trained architectures in the last experiment (and also in the first one for some cases), while the former perform significantly worse in the PANDA training experiment and the fine-tuning experiment (experiments 2 and 3, respectively). This could be caused by two different factors. On the one hand, the Grid Search models were specifically tailored for being trained on the Clinic dataset, which could clearly lead to better performance in those cases where the dataset is used for training the models from scratch. On the other hand, in the fine-tuning experiment, the models were pre-trained with PANDA, which is a vast dataset compared to Clinic. The large amount of samples and the heterogeneity of the dataset, which consists of images from different sources, may require deeper architectures to extract more complex features and generalize better. This aspect would explain why

Table 9
Summary of the average results obtained for each of the trained models on the different experiments performed. The mean κ across the different experiments is also reported for comparison purposes.

	VGG16	InceptionV3	DenseNet121	Custom 1	Custom 2	Custom 3
Exp1: Clinic training	0.700 ± 0.056	0.797 ± 0.017	0.826 ± 0.014	0.709 ± 0.040	0.695 ± 0.044	0.725 ± 0.027
Exp2: PANDA training	0.516 ± 0.055	0.475 ± 0.109	0.476 ± 0.055	0.252 ± 0.152	0.245 ± 0.120	0.185 ± 0.139
Exp3: Transfer learning	0.746 ± 0.030	0.723 ± 0.019	0.731 ± 0.007	0.579 ± 0.028	0.581 ± 0.042	0.529 ± 0.091
Exp4: Combined datasets	0.751 ± 0.094	0.728 ± 0.034	0.746 ± 0.030	0.745 ± 0.074	0.738 ± 0.058	0.686 ± 0.059
Mean	**0.678 ± 0.059**	**0.681 ± 0.045**	**0.698 ± 0.027**	**0.571 ± 0.074**	**0.565 ± 0.066**	**0.531 ± 0.079**

pre-trained networks achieved better results than custom models when PANDA was introduced.

Gleason grading is still an open problem due to many factors including the heterogeneity of the samples and the stain variability across different sources. The recent advances in the deep learning field and computational pathology pave the road for the future integration of CAD systems into medical centers for supporting pathologists. The inter-observer variability in Gleason grading tasks is a well-known fact, and has been analyzed particularly in this work for a local dataset of WSIs obtained from Clinic Hospital (Barcelona, Spain), which was annotated by a group of 5 different pathologists from the same center. The results obtained using different training approaches and CNN architectures prove that this kind of algorithm can achieve similar results compared to those of the inter-observer variability, which was used as the baseline, and even improve them when more heterogeneous datasets are used for training deeper neural networks.

As a future work, we would like to include the trained models that achieved the highest performance in a fully automatic Gleason grading and Gleason scoring diagnostic system. To this end, a two-stage model, which performs a binary classification of benign and malignant images and then performs Gleason grade classification on the images previously classified as malignant would be considered. On the other hand, the Clinic dataset will be released for public use in the near future. This includes releasing the WSIs with the ground truth diagnostic together with the pixel-wise annotations performed by pathologists. It is intended to be an incremental dataset, meaning that it will be updated with new data upon reception and anonymization.

5. Conclusions

In this work, a comparative study of the inter-observer variability in pixel-wise annotations of prostate cancer WSIs sourced from Clinic Hospital (Barcelona, Spain) was carried out, considering Gleason patterns 3-5. This analysis was first performed at the annotation level, studying the discrepancy among five pathologists from the same source at both annotation size and label assigned. A mean pairwise area discrepancy of 46% was obtained on overlapping annotations, while a quadratic Cohen's kappa of 0.6946 was achieved when comparing the labels of the annotations. These results were compared to the performance of six CNN architectures that were trained using four different approaches. In these experiments, both the Clinic dataset, which will be public in the near future, and the largest publicly-available prostate cancer dataset (PANDA) were used. The best results were obtained by DenseNet121, which was able to obtain a mean $0.826 \pm 0.014\kappa$. Most of the trained models achieved a quadratic Cohen's κ that is similar to or even higher than the inter-observer variability of the Clinic dataset, except for experiments where this dataset was not introduced in the training loop.

The different training approaches explored in this work could help other researchers focus on the best methods for benefiting from datasets obtained from different sources in order to improve current CAD systems.

The inter-observer kappa score of 0.6946 obtained from the team of pathologists was outperformed by several CNN models presented in this work. These results prove that the application of Deep Learning in the field of computational pathology could serve as a support for pathologists in the diagnosis process, providing a second opinion or even serving as a triage tool for experts to focus on more aggressive cases first.

CRediT authorship contribution statement

José M. Marrón-Esquivel: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **L. Duran-Lopez:** Conceptualization, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **A. Linares-Barranco:** Funding acquisition, Project administration, Resources, Writing – review & editing. **Juan P. Dominguez-Morales:** Conceptualization, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, 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.

Acknowledgments

This work was partially supported by the Andalusian Regional Project, Spain (with FEDER support) DAFNE (US-1381619) and by the Spanish grant MINDROB (PID2019-105556GB-C33/AEI/10.13039/501100011033). The authors would like to thank Clinic Hospital for providing the WSIs used in this work, together with the pathologists that annotated them.

References

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

[2] Andres Matoso, Jonathan I. Epstein, Grading of prostate cancer: past, present, and future, Curr. Urol. Rep. 17 (3) (2016) 1–6.

[3] Ni Chen, Qiao Zhou, The evolving Gleason grading system, Chin. J. Cancer Res. 28 (1) (2016) 58.

[4] Kasper Drimer Berg, Birgitte Grønkaer Toft, Martin Andreas Røder, Klaus Brasso, Ben Vainer, Peter Iversen, Prostate needle biopsies: interobserver variation and clinical consequences of histopathological re-evaluation, Apmis 119 (4–5) (2011) 239–246.

[5] Alastair M Lessells, Rodney A Burnett, S Rosalind Howatson, Stephen Lang, Frederick D Lee, Kathryn M McLaren, E Robert Nairn, Simon A Ogston, Alistair J Robertson, John G Simpson, et al., Observer variability in the histopathological reporting of needle biopsy specimens of the prostate, Hum. Pathol. 28 (6) (1997) 646–649.

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

[7] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, Kaori Togashi, Convolutional neural networks: an overview and application in radiology, Insights Into Imag. 9 (4) (2018) 611–629.

[8] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, Sebastian Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (7639) (2017) 115–118.

[9] Mitko Veta, Josien PW Pluim, Paul J Van Diest, Max A Viergever, Breast cancer histopathology image analysis: A review, IEEE Trans. Biomed. Eng. 61 (5) (2014) 1400–1411.

[10] Farzaneh Shafiei, Shervan Fekri Ershad, Detection of Lung cancer tumor in CT scan images using novel combination of super pixel and active contour algorithms, Trait. Du Signal 37 (6) (2020) 1029–1035.

[11] Shervan Fekri-Ershad, Mustafa Jawad Al-Imari, Mohammed Hayder Hamad, Marwa Fadhil Alsaffar, Fuad Ghazi Hassan, Mazin Eidan Hadi, Karrar Salih Mahdi, et al., Cell phenotype classification based on joint of texture information and multilayer feature extraction in DenseNet, Comput. Intell. Neurosci. 2022 (2022).

[12] Lei Cai, Jingyang Gao, Di Zhao, A review of the application of deep learning in medical image classification and segmentation, Ann. Transl. Med. 8 (11) (2020).

[13] Lourdes Duran-Lopez, Juan Pedro Dominguez-Morales, Jesús Corral-Jaime, Saturnino Vicente-Diaz, Alejandro Linares-Barranco, COVID-XNet: A custom deep learning system to diagnose and locate COVID-19 in chest X-ray images, Appl. Sci. 10 (16) (2020) 5683.

[14] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, Clara I Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.

[15] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, Thomas J Fuchs, Clinical-grade computational pathology using weakly supervised deep learning on whole slide images, Nat. Med. 25 (8) (2019) 1301–1309.

[16] Eirini Arvaniti, Kim S Fricker, Michael Moret, Niels Rupp, Thomas Hermanns, Christian Fankhauser, Norbert Wey, Peter J Wild, Jan H Rueschoff, Manfred Claassen, Automated Gleason grading of prostate cancer tissue microarrays via deep learning, Sci. Rep. 8 (1) (2018) 1–11.

[17] Peter Ström, Kimmo Kartasalo, Henrik Olsson, Leslie Solorzano, Brett Delahunt, Daniel M Berney, David G Bostwick, Andrew J Evans, David J Grignon, Peter A Humphrey, et al., Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study, Lancet Oncol. 21 (2) (2020) 222–232.

[18] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, Zbigniew Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Li Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.

[20] Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F Steiner, Hester van Boven, Robert Vink, et al., Artificial intelligence for diagnosis and gleason grading of prostate cancer: the PANDA challenge, Nat. Med. 28 (1) (2022) 154–163.

[21] Peter Bankhead, Maurice B Loughrey, José A Fernández, Yvonne Dombrowski, Darragh G McArt, Philip D Dunne, Stephen McQuaid, Ronan T Gray, Liam J Murray, Helen G Coleman, et al., QuPath: Open source software for digital pathology image analysis, Sci. Rep. 7 (1) (2017) 1–7.

[22] Niccolo Marini, Sebastian Otálora, Henning Müller, Manfredo Atzori, 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 (2021) 102165.

[23] Lourdes Duran-Lopez, Juan P Dominguez-Morales, Antonio Felix Conde-Martin, Saturnino Vicente-Diaz, Alejandro Linares-Barranco, PROMETEO: A CNN-based computer-aided diagnosis system for WSI prostate cancer detection, IEEE Access 8 (2020) 128613–128628.

[24] Lourdes Duran-Lopez, Juan P Dominguez-Morales, Daniel Gutierrez-Galan, Antonio Rios-Navarro, Angel Jimenez-Fernandez, Saturnino Vicente-Diaz, Alejandro Linares-Barranco, Wide & Deep neural network model for patch aggregation in CNN-based prostate cancer detection systems, Comput. Biol. Med. 136 (2021) 104743.

[25] Thaína A Azevedo Tosta, Paulo Rogério de Faria, Leandro Alves Neves, Marcelo Zanchetta do Nascimento, Computational normalization of H&E-stained histological images: Progress, challenges and future potential, Artif. Intell. Med. 95 (2019) 118–132.

[26] Niccolo Marini, Manfredo Atzori, Sebastian Otálora, Stephane Marchand-Maillet, Henning Müller, 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, 2021, pp. 601–610.

[27] Mohammad Abdullah-Al-Wadud, Md Hasanul Kabir, M Ali Akber Dewan, Oksam Chae, A dynamic histogram equalization for image contrast enhancement, IEEE Trans. Consum. Electron. 53 (2) (2007) 593–600.

[28] Muhammad Suzuri Hitam, Ezmahamrul Afreen Awalludin, Wan Nural Jawahir Hj Wan Yussof, Zainuddin Bachok, Mixture contrast limited adaptive histogram equalization for underwater image enhancement, in: 2013 International Conference on Computer Applications Technology, ICCAT, IEEE, 2013, pp. 1–5.

[29] Kamlesh Lakhwani, P.D. Murarka, N.S. Chauhan, Color space transformation for visual enhancement of noisy color image, Int. J. ICT Manage. 3 (2) (2015) 9–13.

[30] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, Peter Shirley, Color transfer between images, IEEE Comput. Graph. Appl. 21 (5) (2001) 34–41.

[31] Massimo Salvi, Nicola Michielli, Filippo Molinari, Stain Color Adaptive Normalization (SCAN) algorithm: Separation and standardization of histological stains in digital pathology, Comput. Methods Programs Biomed. 193 (2020) 105506.

[32] Yushan Zheng, Zhiguo Jiang, Haopeng Zhang, Fengying Xie, Jun Shi, Chenghai Xue, Adaptive color deconvolution for histological WSI normalization, Comput. Methods Programs Biomed. 170 (2019) 107–120.

[33] Md Ziaul Hoque, Anja Keskinarkaus, Pia Nyberg, Tapio Seppänen, Retinex model based stain normalization technique for whole slide image analysis, Comput. Med. Imaging Graph. 90 (2021) 101901.

[34] Connor Shorten, Taghi M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (1) (2019) 1–48.

[35] David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, Jeroen Van Der Laak, Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology, Med. Image Anal. 58 (2019) 101544.

[36] Niccolò Marini, Sebastian Otalora, Marek Wodzinski, Selene Tomassini, Aldo Franco Dragoni, Stephane Marchand-Maillet, Juan Pedro Dominguez Morales, Lourdes Duran-Lopez, Simona Vatrano, Henning Müller, et al., Data-driven color augmentation for H&E stained images in computational pathology, J. Pathol. Inform. (2023) 100183.

[37] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, Alexandr A. Kalinin, Albumentations: Fast and flexible image augmentations, Information 11 (2) (2020).

[38] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.

[39] Karen Simonyan, Andrew Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[40] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, Kilian Q Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[41] Fabrício José Pontes, GF Amorim, Pedro Paulo Balestrassi, AP Paiva, João Roberto Ferreira, Design of experiments and focused grid search for neural network parameter optimization, Neurocomputing 186 (2016) 22–34.

[42] Lourdes Duran-Lopez, Juan P Dominguez-Morales, Antonio Rios-Navarro, Daniel Gutierrez-Galan, Angel Jimenez-Fernandez, Saturnino Vicente-Diaz, Alejandro Linares-Barranco, Performance evaluation of deep learning-based prostate cancer screening methods in histopathological images: measuring the impact of the model's complexity on its processing speed, Sensors 21 (4) (2021) 1122.

[43] Diederik P. Kingma, Jimmy Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.

[44] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, Xiaoqiang Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems, 2015, Software available from tensorflow.org.

[45] Padmavathi Kora, Chui Ping Ooi, Oliver Faust, U Raghavendra, Anjan Gudigar, Wai Yee Chan, K Meenakshi, K Swaraja, Pawel Plawiak, U Rajendra Acharya, Transfer learning techniques for medical image analysis: A review, Biocybern. Biomed. Eng. (2021).

[46] Nima Tajbakhsh, Jae Y. Shin, Suryakanth R. Gurudu, R. Todd Hurst, Christopher B. Kendall, Michael B. Gotway, Jianming Liang, Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Trans. Med. Imaging 35 (5) (2016) 1299–1312.

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

[48] Yuri Tolkach, Tilmann Dohmgörgen, Marieta Toma, Glen Kristiansen, High-accuracy prostate cancer pathology using deep learning, Nat. Mach. Intell. 2 (7) (2020) 411–418.

[49] Sebastian Otálora, Manfredo Atzori, Amjad Khan, Oscar Jimenez-del Toro, Vincent Andrearczyk, Henning Müller, Systematic comparison of deep learning strategies for weakly supervised Gleason grading, in: Medical Imaging 2020: Digital Pathology, Vol. 11320, SPIE, 2020, pp. 142–149.