Going Deeper through the Gleason Scoring Scale: An Automatic end-to-end System for Histology Prostate Grading and Cribriform Pattern Detection

Julio Silva-Rodríguez[a], Adrián Colomer[b], María A. Sales[c], Rafael Molina[d], Valery Naranjo[b]

[a]Institute of Transport and Territory, Universitat Politècnica de València, Valencia, Spain
[b]Institute of Research and Innovation in Bioengineering, Universitat Politècnica de València, Valencia, Spain
[c]Anatomical Pathology Service, University Clinical Hospital of Valencia, Valencia, Spain
[d]Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain

Abstract

Background and Objective:

Prostate cancer is one of the most common diseases affecting men worldwide. The Gleason scoring system is the primary diagnostic and prognostic tool for prostate cancer. Furthermore, recent reports indicate that the presence of patterns of the Gleason scale such as the cribriform pattern may also correlate with a worse prognosis compared to other patterns belonging to the Gleason grade 4. Current clinical guidelines have indicated the convenience of highlight its presence during the analysis of biopsies. All these requirements suppose a great workload for the pathologist during the analysis of each sample, which is based on the pathologist's visual analysis of the morphology and organisation of the glands in the tissue, a time-consuming and subjective task.

In recent years, with the development of digitisation devices, the use of computer vision techniques for the analysis of biopsies has increased. However, to the best of the authors' knowledge, the development of algorithms to automati-

cally detect individual cribriform patterns belonging to Gleason grade 4 has not yet been studied in the literature. The objective of the work presented in this paper is to develop a deep-learning-based system able to support pathologists in the daily analysis of prostate biopsies. This analysis must include the Gleason grading of local structures, the detection of cribriform patterns, and the Gleason scoring of the whole biopsy.

Methods:

The methodological core of this work is a patch-wise predictive model based on convolutional neural networks able to determine the presence of cancerous patterns based on the Gleason grading system. In particular, we train from scratch a simple self-design architecture with three filters and a top model with global-max pooling. The cribriform pattern is detected by retraining the set of filters of the last convolutional layer in the network. Subsequently, a biopsy-level prediction map is reconstructed by bi-linear interpolation of the patch-level prediction of the Gleason grades. In addition, from the reconstructed prediction map, we compute the percentage of each Gleason grade in the tissue to feed a multi-layer perceptron which provides a biopsy-level score.

Results:

In our SICAPv2 database, composed of 182 annotated whole slide images, we obtained a Cohen's quadratic kappa of 0.77 in the test set for the patch-level Gleason grading with the proposed architecture trained from scratch. Our results outperform previous ones reported in the literature. Furthermore, this model reaches the level of fine-tuned state-of-the-art architectures in a patient-based four groups cross validation. In the cribriform pattern detection task, we obtained an area under ROC curve of 0.82. Regarding the biopsy Gleason scoring, we achieved a quadratic Cohen's Kappa of 0.81 in the test subset.

Shallow CNN architectures trained from scratch outperform current state-of-the-art methods for Gleason grades classification. Our proposed model is capable of characterising the different Gleason grades in prostate tissue by extracting low-level features through three basic blocks (i.e. convolutional layer + max pooling). The use of global-max pooling to reduce each activation map

2

has shown to be a key factor for reducing complexity in the model and avoiding overfitting. Regarding the Gleason scoring of biopsies, a multi-layer perceptron has shown to better model the decision-making of pathologists than previous simpler models used in the literature.

1. Introduction

Worldwide, prostate cancer (PCa) is the second most common cancer in men, with 1.3 million new patients in 2018 [1]. According to the World Health Organisation, the yearly number of new cases will increase by more than 40% in this decade [2]. The main tool to diagnose PCa, once clinical explorations or blood test suggest its presence, is the prostate biopsy. Small portions of the tissue are extracted with a needle, laminated, stained with Hematoxylin and Eosin (H&E) and finally stored in crystal. Then, the sample is analysed under the microscope by the pathologist, determining the presence and grade of cancerous patterns depending on the morphology and organisation of the glands, nuclei and lumen using the Gleason grading system [3]. In this system, different cancer patterns in the tissue are grouped in different grades according to the prognosis of the cancer. In particular, for two-dimensional tissue slides, the Gleason grades (GG) range from 3 to 5, correlating inversely with the degree of gland differentiation of the tissue. The Gleason grade 3 (GG3) includes atrophic well differentiated and dense glandular regions. The GG4 contains cribriform, ill-formed, large-fused and papillary glandular patterns. Finally, GG5 includes isolated cells or file of cells, nests of cells without lumina formation and pseudo-roseting patterns. Examples of patterns belonging to different grades are presented in Figure 1.

Pathologists classify by visual inspection the tissue regions, detecting the presence of one or more Gleason patterns and, finally, diagnose the combined Gleason score according to the most prominent grades (e.g. the combined grade $5 + 4 = 9$ would be assigned to a sample in which the main cancerous Gleason
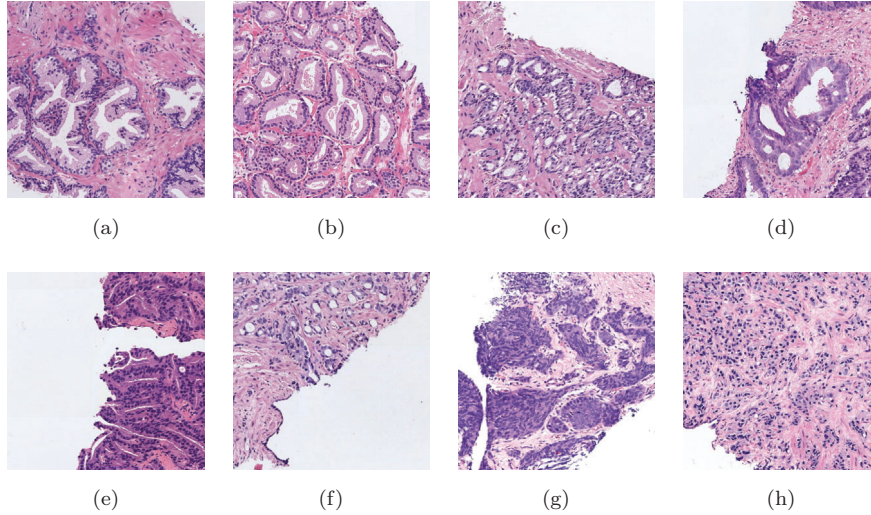
3

Figure 1: Patches of H&E histology samples presenting different Gleason patterns. (a): Non-cancerous well-differentiated glands; (b): Gleason grade 3 containing atrophic dense patterns; (c): Gleason grade 4 containing large fused glandular patterns; (d): Gleason grade 4 containing cribriform patterns; (e): Gleason grade 4 containing papillary structures; (f): Gleason grade 4 containing individual poorly-formed glands; (g): Gleason grade 5 including nests of cells without lumen formation; (g): Gleason grade 5 containing files of isolated cells.

grade is 5 followed by the grade 4). Therefore, the combined Gleason score ranges from 6 to 10, and it is assigned to the whole biopsy. This score is currently the best marker of prostate cancer prognosis and it defines the treatment to apply [4]. However, the Gleason scoring of histological prostate biopsies is a high time-consuming and repetitive task, which has intra and inter pathologist variability. Moreover, after the last International Society of Urological Pathology (ISUP) Consensus Conference in 2014 [5], new guidelines have been included that increase the pathologists' workload. In particular, it is recommended to also report the percentage of Gleason grade 4 in the sample, mainly for regions scored as $3 + 4 = 7$, where a higher percentage of Gleason grade 4 indicates the convenience of an earlier treatment [6], and the presence of cribriform glandular patterns, which indicate worse prognosis than the presence of other Gleason

grade 4 patterns [7, 8]. Computer-Aided Diagnosis systems (CAD) support the work of pathologists and increase the objectivity in the this process. These are based on the digitisation of the histological crystals, obtaining whole slide images (WSIs) and developing computer vision algorithms to detect the cancerous regions inside the biopsy (or WSI).

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

Regarding the images, mainly three types of histological images have been used: WSIs, prostactetomies and Tissue Micro Arrays (TMAs). TMAs are clusters of representative tumor areas extracted manually by pathologists [9]. TMAs are used for testing new techniques in a large number of different tumour samples. One of the main limitations of TMAs lies in the small amount of tissue that can be included in each samples, which may not be representative of the whole tumor region in epithelial tumors with heterogeneous patterns [10]. This is the case of prostate cancer, which has different patterns for each Gleason grade, as previously mentioned. Non-cancerous patterns that could confuse CAD systems, as the inflamed tissue or benign multi-nucleation, could be lost using TMAs. Thus, the strategy based on TMA analysis is not used in clinical practice [11] and it is more convenient to develop CAD systems based on raw WSI analysis. A model trained using large databases of WSIs could be used for both WSIs and prostactetomies. The works in [12, 13, 14, 15, 16, 17, 18] follow the strategy of WSI analysis, while in [19, 20, 21] the authors use TMAs to develop the CAD models.

With regard to the objectives to be addressed, some works focus just on the detection of prostate cancer against non-cancerous tissue [13, 17] or on the first-stage prostate cancer detection [22]. A full analysis of Gleason grades

5

from 3 to 5 is usually limited by the size of the collected database, and the low prevalence of Gleason grade 5. Due to that, numerous researchers classify differentiating among non-cancerous samples, low grade (Gleason grade 3), and high Grade (Gleason grade $\geq$ 4) [18, 23, 24] or among non-cancerous, Gleason grade 3, and Gleason grade 4 [12, 15]. The most recent works tried to predict the full Gleason grading (Benign - Grade 3 - Grade 4 - Grade 5) in [19, 20, 21] but only using TMAs cores. To the best of the authors's knowledge, works analysing deeper the Gleason grades, this is, focusing on the automatic detection of individual patterns of a Gleason grade (i.e. cribriform pattern, which belongs to the Gleason grade 4 group) do not exist. This work represents an attempt in this direction.

Finally, concerning the techniques used to deal with the different mentioned objectives, the most common approach to analysed both is to perform a patch-based strategy (see Figure 2). The motivation for using this strategy is the large size of both TMAs and, especially WSIs, together with hardware limitations.
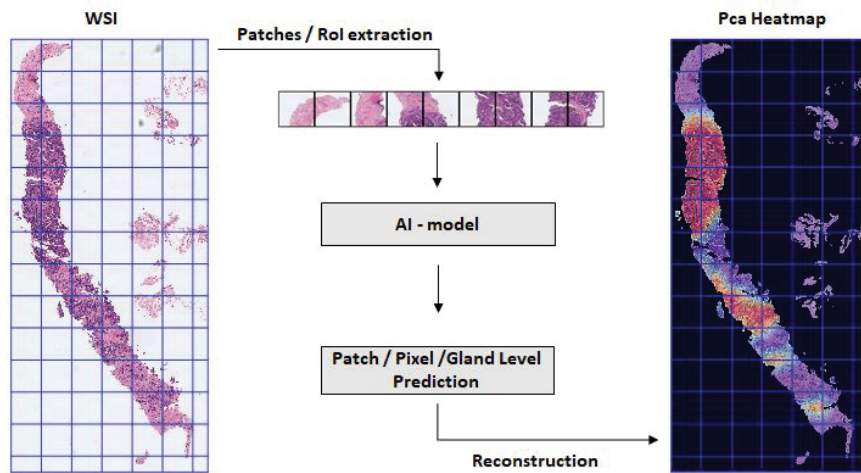

Figure 2: General workflow for high resolution histology slides processing.

Below, we will focus only on the description of the different techniques used, until now, for the patch-level Gleason grading. In the literature we can find approaches based on classic machine learning techniques with a hand-crafted

feature extraction and deep learning algorithms (automatic feature extraction) by means of convolutional neural networks (CNN). In Nir et al. (2018) [20] a comparison between both approaches is carried out with a database of 333 cores of TMAs. Glands and nuclei are segmented to obtain features related to their size, intensity distributions and number of elements in each patch at different resolutions. Those are combined with full patch-level features related to the colour distribution and SURF descriptors to fit different machine learning models as linear discriminant analysis, linear regression, support vector machines, and random forests. Those models are compared with a U-Net CNN. The best result reported is a Cohen's quadratic kappa (κ) overall agreement measure of 0.51 obtained by the linear regression model. Nevertheless, in a later publication by Nir et al. (2019) [21] a κ of 0.60 was obtained by fine-tuning the CNN architecture MobileNet. In Arvaniti et al. (2018) [19] a larger database is used, with 886 cores. The patch-level grading is addressed through fine-tuning different CNN architectures such as VGG16, InceptionV3, ResNet50, DenseNet121, and MobileNet. The best results are reported with the last one, achieving a κ of 0.67 in the training set and 0.55 in the test one.

Regarding the classification of the Gleason score for the whole biopsy (whole slide image), only a few works have addressed it, and only using TMAs. The common strategy used is to obtain the percentage of each grade in the analysed image and to assign the first and second components above a threshold as primary and secondary grades respectively. In Arvaniti et al. (2018) [19] the full Gleason scoring, using TMAs, is addressed, archiving κ of 0.75. Unfortunately, this simple model did not perform for extreme cases, for example $5 + 5 = 10$. In this case, a precision of 0.10 is reported in this work. In addition, the primary and secondary grades are not just related to the proportion of the different grades in the tissue, but also to the severity of each grade (e.g. GG5 could be diagnosed as secondary grade even having less proportion than GG4 or GG3 in the tissue).

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

- Detection of the cancerous regions in the tissue according to the Gleason grading system.

- Detection of cribriform patterns.

- Calculation of the percentage of each Gleason grade in the biopsy.

- Gleason scoring of the whole biopsy, taking into account not only the grade proportion but also its severity.

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

8

to the whole biopsy. The obtained results show the good performance of this model which outperforms the previous state-of-the-art methods.
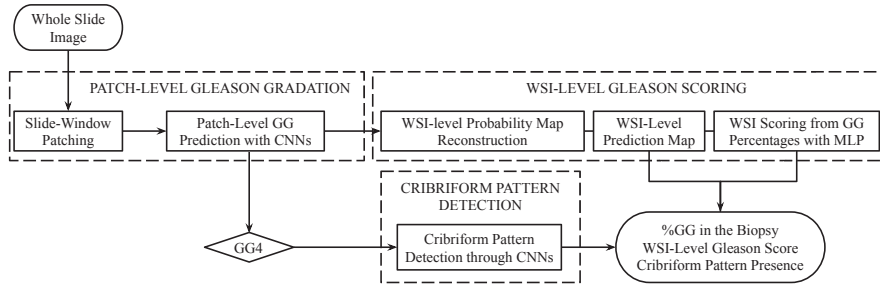

Figure 3: Flowchart in which the different blocks of our system are presented. Taking as input a prostate whole slide image (WSI), the system performs a patch-level Gleason grade prediction through convolutional neural networks. If one patch is classified as Gleason grade 4 (GG4), a cribriform pattern detection is carried out by fine-tuning the model of the previous stage. Finally, the regions in the WSI are reconstructed and a pixel-level Gleason grade assignement is carried out. The WSI-level Gleason scoring is performed with a multi-layer perceptron taking as input the percentage of the Gleason grades in that region.

The paper is organised as follows, in Section 2 we introduce the database used in this work, SICAPv2, a large set of prostate whole slide images with pixel-level annotations of the Gleason grades and WSI-level annotations of the Gleason scores assigned by expert pathologists. In Section 3 we describe the methodological details of our proposed CAD system, based on CNNs able to predict the Gleason grade and presence of cribriform pattern in local patches of the WSIs. From those local predictions, in this section we also detail the process of predicting the WSI-level Gleason score. In Section 4 we describe the performed experiments in order to validate our models. In particular, Section 4.2 describes the experiments related to the patch-level Gleason grading, Section 4.3 the detection of cribriform patterns and in Section 4.4 we present our results related to the biopsy-level Gleason scoring. Finally, Section 5 summarises the conclusions extracted with the carried out experiments.

9

2. Materials: SICAP database

The database presented in this paper, SICAPv2, is, to the best of the authors's knowledge, the largest public collection of prostate H&E biopsies with local-level annotations of Gleason grades. SICAPv2 is an extension the database introduced in [17] and will be publicy available after the publication of this paper.

After analysing the literature, four main prostate cancer tissue image databases were found. The largest database with prostate biopsies was released by The Cancer Genome Atlas project[1] [25] with up to 720 prostate biopsy slides. Nevertheless, the lack of annotations at both the local and biopsy levels of the Gleason grades restricts the use of these data. The database shared by Arvaniti et al. [19] includes pixel-level annotations of Gleason patterns from 886 small regions of slides (cores of TMAs). Unfortunately, as discussed earlier, those cores do not represent the heterogeneous patterns of local structures of prostate cancer and benign lesions, so they lack clinical relevance for the slide-level Gleason score diagnosis. Similar limitations are found in the recent database from the challenge Gleason19 in the MICCAI 2019 conference[2], with 331 cores annotated by different pathologists, and the dataset used in [16], composed by 625 isolated patches. Although those databases contribute to the validation of different algorithms, the lack of large databases with clinical reference of heterogeneous patterns has been a limiting factor for the scientific community to develop deep-learning-based methods which demand a large amount of data. One of the contributions of this work is the publication of a large database of WSIs containing biopsy-level labels (i.e. Gleason scores for each biopsy) and pixel-level Gleason grades annotations, in which for the first time, the presence of cribriform glandular regions is indicated.

SICAPv2 database includes 155 biopsies from 95 different patients who signed the pertinent informed consent. The tissue samples where sliced, stained

---

[1]https://portal.gdc.cancer.gov/
[2]https://gleason2019.grand-challenge.org/Home/

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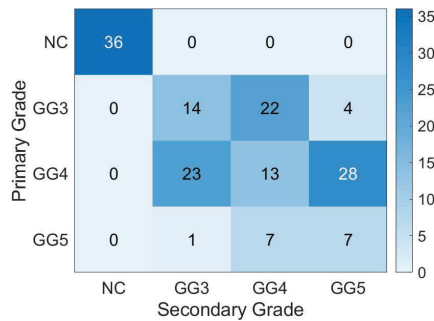

Figure 4: Description of the Gleason scores in the SICAPv2 database. Co-occurrence matrix of primary and secondary Gleason grades in each biopsy. NC: non cancerous, GG3: Gleason grade 3, GG4: Gleason grade 4 and GG5: Gleason grade 5.

The local cancerous patterns were annotated using an in-house software based on the OpenSeadragon libraries [26], following the Gleason scale and indicating the presence of cribriform glandular structures. In order to process the large WSIs, these were down-sampled to $10x$ resolution and divided into patches of size 512^2 and overlap of 50% between them. Those values were previously optimised for the detection of cancerous patterns in [17]. A mask of the presence of tissue in the patches was obtained by applying the Otsu threshold method. To develop the model able to predict the main Gleason grade, patches with less than 20% of tissue were excluded. In addition, patches without cancerous

patterns annotated by the pathologists belonging to cancerous biopsies where also discarded. After this procedure, the database contains 4417 non-cancerous patches, 1635 labelled as GG3, 3622 as GG4, and 665 as GG5. Note that if one patched contained more than one annotated grade, the majority grade was assigned as label. 763 GG4 patches also contain annotated cribriform glandular regions. A summary of the database description is presented in Table 1.

Table 1: SICAPv2 database description. Amount of whole slide images and their respective biopsy-level primary label (first row) and number of patches of each one of the Gleason categories (second row).

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

The data collected by Arvaniti et al. [19] was also utilised to validate the models produced in our study. The cores were resized to match the resolution used in our models and patched to the dimensions used in our database. By this approach, each one of these cores is approximately equivalent to one of our patches. Thus, 115 non-cancerous images, 274 patches labelled as GG3, 210 GG4, and 104 GG5 were used to validate our work in an external database. Also, the patches shared by Gerytch et al. [13] were used in our work for the validation of our proposed model. After normalisation of the images to match our methodology, 32 non-cancerous images, 95 patches labelled as GG3, 216 GG4, and 70 GG5 were obtained.

3. Methods

3.1. Patch-Level Gleason Grading

The patch-level classification in the different Gleason grades is carried out by means of convolutional neural networks. We propose a self-designed base-model architecture (from now on called $FSConv$) which consists of a simple convo-

lutional architecture with three convolutional layers and dimensional reduction operation employing max-pooling layers (Table 2).

Table 2: *FSConv* architecture description. It consists of three blocks with convolutional filters, ReLU activation and max-pooling operation.

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

After the automatic feature extraction blocks (base model), we introduce as top model a global-max-pooling layer. To show the superior performance of this architecture, different configurations already applied in the literature to the same problem, have been also tested as top models and are described next.

One of the main approaches is the flattening of the activation volume resulting from the final convolutional block and the class prediction through consecutive fully-connected layers. In this case, overfitting is addressed by means of a random dropout of a percentage of the neurons in each training iteration. Nevertheless, these top-model architectures include a large number of parameters to optimise, increasing the complexity of the model, and they are sensitive to the location of the structures in the image. This problem is usually dealt with data augmentation techniques, applying, for example, random rotations and translations to the images. Other approaches propose the convenience of using global-average pooling on the last feature maps as regulariser to make the model translation-invariant and decrease its complexity [27]. This technique is used in [19] for the prediction of prostate cancer Gleason degree with fine-tuned models. Due to the use of a patch-based strategy with sliding window, the location and amount of the cancerous structures in the image is not controlled.

Thus, as shown in Figure 5, some patches could have small portions of cancerous tissue. The global-average pooling layer takes into account the information in the whole activation map, and in those cases, the output of the filter that detects this pattern could be diminished. To make the models robust to the amount and location of cancerous tissue, we propose in this work the use of the global-max-pooling layer to play the role of the global-average pooling. All different configurations, fully-connected layer with ReLU activation and dropout regularisation (FC), global-average-pooling (GAP) and global-max-pooling (GMP) layers and their combinations are implemented and their performance is discussed in this work.
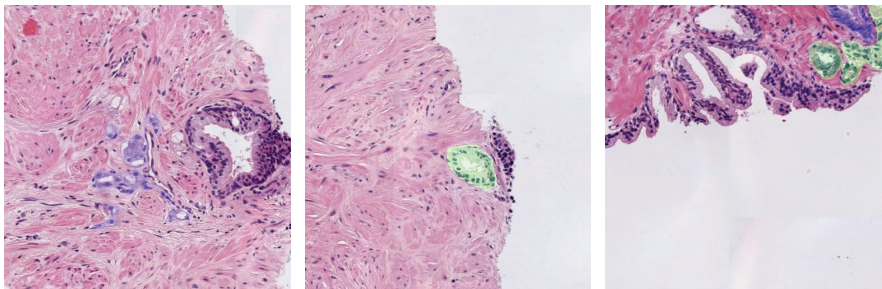

Figure 5: Patches with small amount of cancerous tissue. Green: GG3, Blue: GG4.

For comparison, together with the proposed architecture trained from scratch, we fine-tuned several well-known architectures: VGG19 [28], ResNet-50 [29], InceptionV3 [30] and MobileNetV2 [31]. All of them were pre-trained in the Imagenet data set [32]. For the feature extraction stage, the base model from those pre-trained models is extracted and partially retrained. This strategy is usually used to transfer the knowledge obtained in extracting features from a large database to specific domains where the amount of data is limited. Nevertheless, the patterns of the images used during the training are very different from the histology ones. To keep just the low-level features (contours, combination of basic colours, general shapes, etc.) from the pre-trained models, the weights of just the first convolutional blocks are frozen, while the rest are re-trained to

adapt the model to the specific application. The layer from which the freezing strategy is applied is empirically optimised for each model, and is specified in the experimental part of the paper, in Section 4.2.

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

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

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

Stochastic Gradient Descend is applied as optimiser and the training procedure is performed using mini-batches. The values of learning rate and batch size are fixed empirically for each configuration and experiment, and they are specified in Section 4.2. Data augmentation techniques are used on the training set applying random rotations and translations to the images.

3.2. Cribriform Pattern Detection

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

15

Stochastic Gradient Descent is used as optimiser applied on mini-batches and including data augmentation with random rotations, translations and brightness variations.

3.3. Whole Slide Image Gleason Scoring

To predict the Gleason score of the WSI, it is necessary to compute the tissue percentage of each Gleason grade present in the WSI. For that purpose, the first step is to apply the patch-level classification (section 3.1). Then, for each pixel, the predicted probabilities for each class is estimated by bilinearly interpolating the predicted probabilities of the closest patches in terms of euclidean distance to the center of the patches. Thus, a probability map per class (i.e NC, GG3, GG4, and GG5) is obtained per each WSI. Finally, the percentage of each Gleason grade is calculated after assigning each pixel the class, c, with the highest probability.

The pathologist's decision making while assigning a Gleason score to a WSI takes into account both the percentage of each Gleason grade and the severity of each grade. To model this process, we propose to train a Multi-Layer Perceptron (MLP) to automatically predict the combined Gleason scoring of a biopsy, by means of a multi-class classification task. This task requires the prediction of both primary and secondary Gleason grades. To address it, MLP is selected as a suitable classifier, due to its flexibility to adapt the architecture to perform a multi-output classification. The proposed MLP architecture consists of a branch with two outputs (see Figure 6). The branch is composed of two fully-connected layers with 16 and 8 neurons respectively, and ReLU as activation function. The branch is then divided into two output layers: one for the primary Gleason grade and one for the secondary grade. These output layers are composed of four neurons each, one neuron per target class (i.e. NC, GG3, GG4 or GG5) and soft-max as activation function. The loss function used is the categorical cross-entropy.
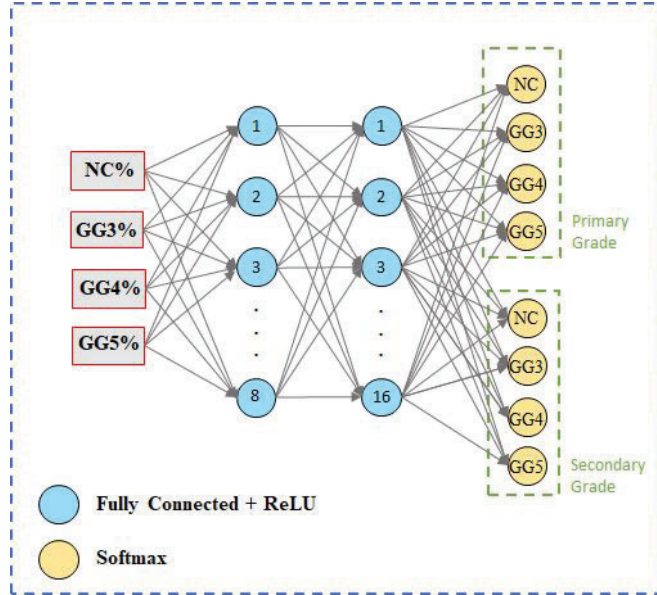
Figure 6: Proposed Multi-Layer Perceptron (MLP) for the whole slide image Gleason scoring. The model takes as input the percentage of each Gleason grade in the whole slide image, and is composed by a main branch with two fully-connected layers and two outputs. The intermediate layers consist of 8 and 16 neurons respectively and ReLU as activation function. The output layers present one neuron per target class and soft-max activation. NC: non cancerous, GG3: Gleason grade 3, GG4: Gleason grade 4, GG5: Gleason grade 5.

4. Experiments

In this section we present the results of the different experiments carried out to show the performance of the proposed approach for the different classification tasks: patch-level classification, cribriform pattern detection and WSI scoring. In all cases, when possible, we also present a comparison with current state of the art methods and discuss the obtained results.

4.1. Database Partitioning and Metrics

In order to train the models and optimise the hyperparameters involved in this process, the database was divided following a cross-validation strategy. In particular, each patient was exclusively assigned to one fold with the aim of

17

avoiding overestimation of the performance of the system [21] and ensuring its ability of generalisation. Thus, the database was divided into 5 groups containing approximately 20% of the patches each one. Notice that this process was carried out trying to guarantee the class balance character between sets. A summary of the resulting partition is presented in Table 3.

Table 3: Database partition description: number of patients-patches for each grade in each validation fold (4-fold cross-validation) and test subset.

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

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

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

$$F1S_c = 2 \times \frac{precision_c \times sensitivity_c}{precision_c + sensitivity_c} \qquad (2)$$

Cohewhere c indicates the predicted classes.

However, an automatic method should be less penalised when classifying a GG5 tissue as GG4 than as NC, even more so when taking into account the inter and intra-observer variability. In the literature, this fact is addressed using the Cohen's quadratic kappa (κ) metric [33]. The metric κ ranges from -1 to 1, being directly proportional to the level of agreement between observers (-1 no agreement, 1 total agreement). Although there is not objective interpretation of which are the reasonable values for κ in medical applications, recent proposals [34] define a moderate agreement if κ is higher than 0.6, while a strong agreement is stated when κ is higher than 0.8.

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

In order to evaluate the system for the detection of cribriform patterns, the area under the Receiver Operating Characteristic (ROC) curve (AUC) was used. In medical applications, a system is considered reliable if the AUC value exceeds 0.80 [35]. The predicted labels are obtained by thresholding the scores (cribriform if the probability is above 0.5), and then evaluated by means of ACC, sensitivity and specificity.

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

4.2. Patch-Level Gleason Grading

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

4.2.1. FSConv Architecture Benchmarking

After optimising the hyperparameters (learning rate, batch size, number of epochs, etc.), table 4 shows the obtained results in the validation sets for the

19

proposed network $FSConv$ with different top models: fully-connected layers (FC), global-max pooling (GMP), global-average pooling (GAP), or a combination of them (GAP+FC or GMP+FC). Table 4 also presents the results for the best tested fine-tuned architectures, VGG19 and RestNet, using the same top models as $FSConv$. The optimum hyperparameters were: learning rate of 0.01 for $FSConv$ and 0.0001 for the finned-tunned networks, batch size of 32 images and 200 epochs in all cases. The base model of the fine-tuned networks were also optimised, being selected to freeze the first convolutional block for VGG19 and setting all layers as trainable for RestNet. Futhermore, Table 5 presents a comparison in terms of storage space (in kilobytes, KB) and number of trainable parameters of each architecture.

Table 4: Results for patch-level Gleason grades prediction in the validation set. The performance of the different models ResNet, VGG19 and $FSConv$ are presented with the different configurations of top models. The metrics presented are the accuracy (ACC), the F1-Score (FS1), computed per class and its average, and the Cohen's quadratic kappa (κ). GMP: global-max pooling, GAP: global-average pooling and FC: fully-connected layers.

Experiment	ACC	F1S				Avg-F1S	κ
		NC	GG3	GG4	GG5		
VGG19+FC	0.7218 ± 0.0411	$\mathbf{0.8871 \pm 0.0178}$	0.6639 ± 0.0509	0.6041 ± 0.1694	0.5206 ± 0.0996	0.6689 ± 0.0650	0.7346 ± 0.0324
VGG19+GMP	0.7213 ± 0.0542	0.8729 ± 0.0207	0.6480 ± 0.0609	0.6032 ± 0.1673	$\mathbf{0.5450 \pm 0.0943}$	0.6673 ± 0.0766	0.7174 ± 0.0641
VGG19+GMP+FC	0.7273 ± 0.0424	0.8860 ± 0.0194	0.6821 ± 0.0633	0.6093 ± 0.1508	0.5313 ± 0.0820	0.6772 ± 0.0651	$\mathbf{0.7474 \pm 0.0648}$
VGG19+GAP	0.7306 ± 0.0460	0.8814 ± 0.0267	0.6434 ± 0.0961	0.6530 ± 0.1164	0.5138 ± 0.0847	0.6729 ± 0.0472	0.7175 ± 0.0730
VGG19+GAP+FC	0.7246 ± 0.0485	0.8795 ± 0.0130	0.6905 ± 0.0601	0.6099 ± 0.1542	0.5216 ± 0.1185	0.6754 ± 0.0724	0.7179 ± 0.0623
ResNet+FC	0.6952 ± 0.0316	0.8383 ± 0.0151	0.6670 ± 0.0753	0.5726 ± 0.1271	0.4845 ± 0.0534	0.6406 ± 0.0550	0.6811 ± 0.0463
ResNet+GMP	0.6879 ± 0.0380	0.8368 ± 0.0181	0.6424 ± 0.0729	0.5567 ± 0.1315	0.5069 ± 0.0739	0.6357 ± 0.0609	0.6780 ± 0.0330
ResNet+GMP+FC	0.6991 ± 0.0220	0.8458 ± 0.0137	0.6748 ± 0.0811	0.5521 ± 0.1230	0.4925 ± 0.4925	0.6413 ± 0.0440	0.6890 ± 0.0534
ResNet+GAP	0.6965 ± 0.0269	0.8487 ± 0.0131	0.6777 ± 0.0834	0.5455 ± 0.1240	0.5019 ± 0.0405	0.6434 ± 0.0552	0.6927 ± 0.6927
ResNet+GAP+FC	0.7024 ± 0.0287	0.8471 ± 0.0075	0.6826 ± 0.0890	0.5556 ± 0.1268	0.5184 ± 0.0523	0.6509 ± 0.0557	0.6982 ± 0.0427
$FSConv$+FC	0.7330 ± 0.0303	0.8395 ± 0.0437	0.6503 ± 0.0229	0.6964 ± 0.0606	0.5441 ± 0.1294	0.6826 ± 0.0207	0.6809 ± 0.0273
$FSConv$+GMP	$\mathbf{0.7622 \pm 0.0075}$	0.8766 ± 0.0167	$\mathbf{0.7277 \pm 0.0228}$	$\mathbf{0.7093 \pm 0.0540}$	0.5364 ± 0.1062	$\mathbf{0.7125 \pm 0.0251}$	0.7328 ± 0.0465
$FSConv$+GMP+FC	0.7286 ± 0.0610	0.8724 ± 0.0341	0.6955 ± 0.0374	0.6317 ± 0.2017	0.4529 ± 0.0379	0.6631 ± 0.0592	0.7200 ± 0.0405
$FSConv$+GAP	0.5317 ± 0.0886	0.6830 ± 0.0805	0.3228 ± 0.2408	0.4418 ± 0.2587	0.3391 ± 0.1835	0.4467 ± 0.1501	0.4153 ± 0.2376

Regarding the results obtained in the fine-tuned models, the use of architectures with residual blocks provided slightly worse results than the sequential approach, similarly as the previous results reported in the literature where sequential models used to outperform residual ones [17, 19, 20]. In relation to the use of different top models, no differences were found in the accuracy of the fine-tuned architectures, observing similar results for all of them.

Table 5: Number of parameters and memory usage of the different CNN architectures tested for the patch-level Gleason grading task. KB: kilobytes.

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

In relation to $FSConv$ architecture, interesting results were obtained while testing the use of different top models. The best performing architecture to validate the system is the one with global-max pooling, $FSConv$+GMP. The outperforming of the global-max pooling compared to the fully-connected configuration could be explained by the reduction in the number of weights to be optimised (see Table 5), making the model simpler and more capable of generalising to new images, and by the invariance to the pattern location provided by the global-pooling operations. However, the $FSConv$ model did no converge properly using global-average poling in the top model ($FSConv$+GAP), an effect non observed in the case of fine-tuned architectures. The explanation of this behaviour could be related to the receptive field of the model. The receptive field is defined as the region of the image involved in the cross-correlation

21

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

Paying attention to Table 4 and taking into account all the figures of merit, we conclude that $FSConv$+GMP configuration is the best performing one for the patch-level Gleason grading. In the validation set used, this model outperforms the VGG19+GMP+FC architecture in terms of accuracy (0.7622 compared to 0.7273) and average F1-score (0.7125 against 0.6772). Furthermore, the $FSConv$+GMP model performs specially well when distinguishing between GG3 and GG4, the most difficult task in the pathologists' work, reaching F1-scores of 0.7277 and 0.7093 respectively (see Table 4). This is the first time in the literature that self-defined architectures trained from scratch outperform fine-tuned architectures from the state-of-the-art pre-trained in Imagenet for Gleason grading. Moreover, the reduced amount of parameters (2×10^7 in the VGG19+GMP+FC model against 6×10^5 in the $FSConv$+GMP model, see Table 5), makes more convenient the $FSConv$ architecture for deployment. Thus, the model $FSConv$+GMP was trained using all the images in the cross-validation sets in order to evaluate its performance in the external test cohort.

The results of the proposed model for the test set and a comparison of them with previous state-of-the-art works are reported in Table 6. κ value increases up to 0.77 in the test subset for $FSConv$+GMP. In comparison with previous studies, our results outperform the state of the art, obtaining almost a

22

strong agreement between our model and the pathologist, while just moderate agreement ($\kappa = 0.55$ [19]) was obtained previously in the test set. Figure 7 shows the performance evaluation of $FSConv$. In particular, the confusion matrix for validation and test subsets are presented. From this figure, it can be observed that most of the errors occur between adjacent classes.

Table 6: Results for the patch-level Gleason grading in the test set for the model $FSConv$+GMP and comparison with previous literature. The metrics presented are accuracy (ACC), F1-Score (1S), computed per class and its average, and Cohen's quadratic kappa (κ). Note that for the results reported in previous literature not all the metrics were reported. GMP: global-max pooling.

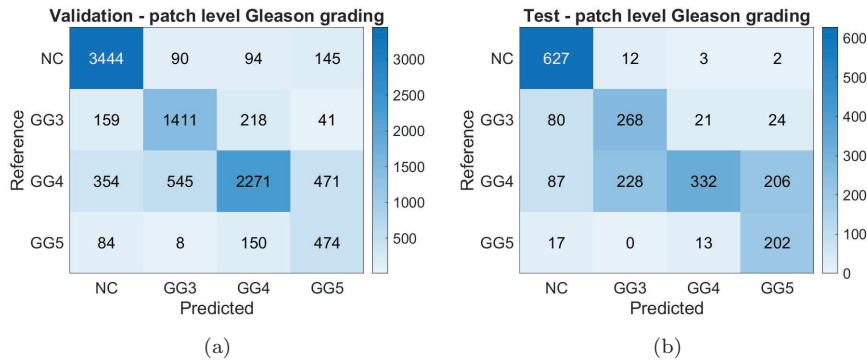| Experiment		ACC	F1S				Avg-F1S	κ
			NC	GG3	GG4	GG5		
$FSConv$+GMP	Test	0.67	0.86	0.59	0.54	0.61	0.65	**0.77**
Arvaniti et al. [19]	Validation	-	-	-	-	-	-	0.67
	Test	-	-	-	-	-	-	0.55
Nir et al. [21]	Validation	-	-	-	-	-	-	0.61

Figure 7: Confusion Matrix of the patch-level Gleason grades prediction done by $FSConv$ network in (a) validation set and (b) test set.

4.2.2. Model Interpretation

One of the main drawbacks of deep learning models in medical practice is the lack of interpretability. This fact creates distrust in the clinicians, the final

users of CAD systems. To deal with this problem, in this research we study the interpretability of the trained models by means of the Class Activation Maps technique (CAMs). Both VGG19+GMP+FC (the best fine-tuned model) and $FSConv$+GMP models are compared in this section using CAMs.

This technique was proposed in [36] as a procedure to obtain a heatmap indicating the regions of the input image to which the model is paying attention to predict certain class. CAMs for both models are obtained for images correctly classified (see Figure 8) and for images miss-classified by the VGG19 model (see Figure 9). These illustrations are organised as follows: the first row corresponds to the original patch, and the second and third rows show the CAMs for VGG19 and $FSConv$ models, respectively. In Figure 8 each column shows an example per class: NC, GG3, GG4 and GG5 accordingly. The main difference in the results obtained by VGG19 and $FSConv$ is the best differentiation between GG3 and GG4 by the second model (see Table 4), the most difficult task in the pathologists' work. In Figure 9 three of those cases are presented in each column: two cases predicted by the VGG19 as GG3 and one as GG5, respectively. Those cases were correctly classified as GG4 by $FSConv$ model.

CAMs obtained for VGG19 in NC, GG3 and GG4 show that the model is basing the decision in glandular regions detected and classified correctly. In the case of GG5, the highlighted region presents a group of single cells and infiltrating cords without lumen formation, characteristic patterns of poor differentiate tissue in GG5. In the case of $FSConv$ architecture, the CAM heatmap does not detect large regions, but small dots instead. Although the glandular regions are not detected, paying attention to the position where the dots are pointing at, we can extract interesting insights (see Figure 8). In the case of GG4, the map is activated in a small nest belonging to a fused-glands structure with irregular cribriform shape. Regarding the GG3 image, the dot indicates thick cytoplasm in different medium-sized tubular glands. In the image marked as GG5, the CAM highlights single isolated cells with hyperchromasia. Less interpretable is the CAM obtained in the NC image, where any gland is detected. We speculate that the model carries out this classification by dismissing the presence of cancer-
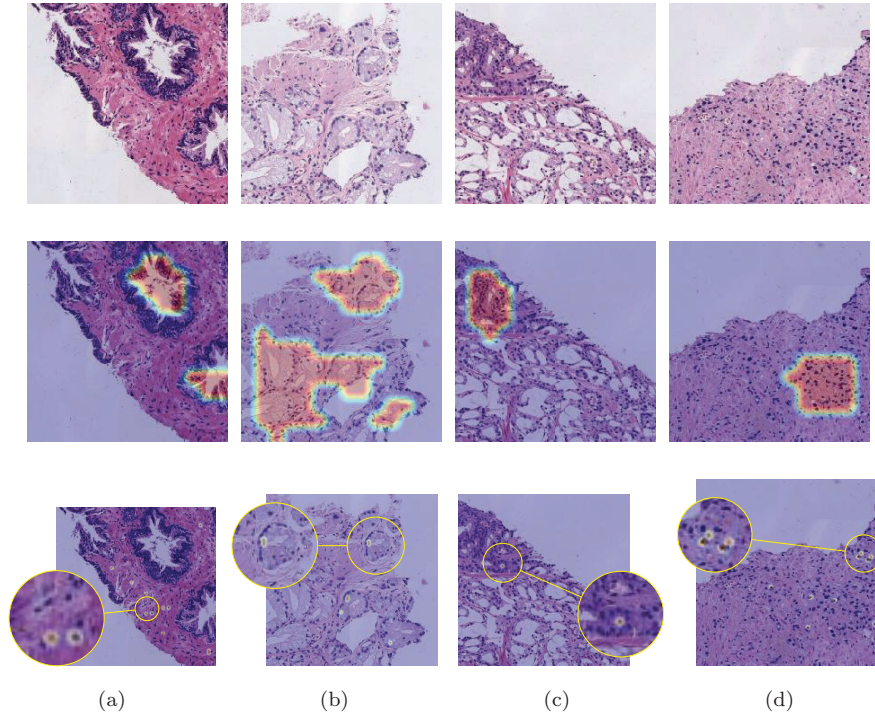
24

Figure 8: Original image (first row) and Class Activation Maps (CAMs) obtained by the VGG19 model (second row) and the $FSConv$ network (third row) in four images correctly classified. Non-Cancerous (a), Gleason grade 3 (b), Gleason grade 4 (c) and Gleason grade 5 (d).

ous patterns. Regarding the cases where VGG19 miss-classifies GG4 in Figure 9, a correct detection of the regions of interest is observed. However, these glandular regions are not correctly classified as GG4, while $FSConv$ model does it just paying attention to closed lumens in small ill-formed glands. At this stage of understanding, we believe that this fact is the cause of the different performance by both models. VGG19 focuses the prostate cancer detection on detecting epithelial and glandular regions, and these structures present a larger heterogeneity than its basic components (colour and size of individual glands, diameter and opening degree of lumens in the glandular region, etc.). This could be the reason why the VGG19 generalises slightly worse than $FSConv$.
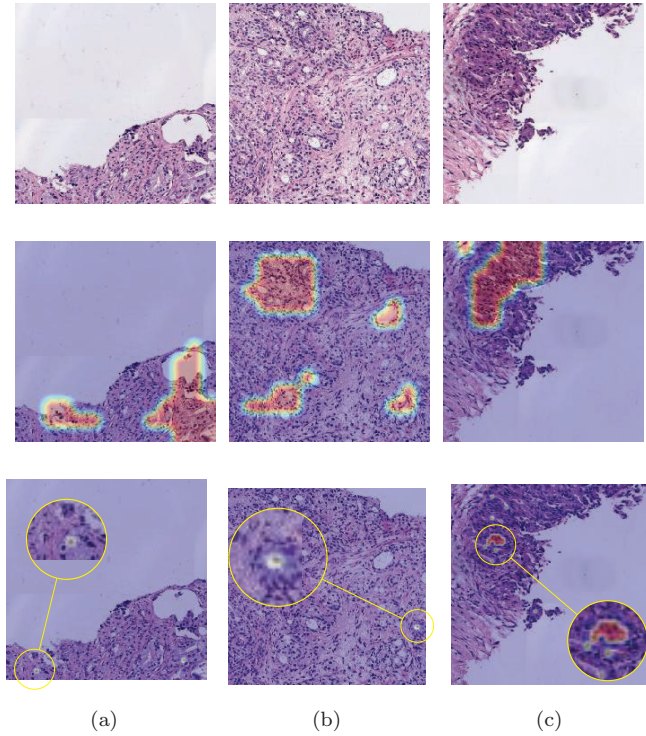
25

Figure 9: Original images (first row) and Class Activation Maps (CAMs) obtained on the VGG19 model (second row) and the $FSConv$ network (third row) in images with GG4 correctly classified by the $FSConv$. The VGG19 model classification of those cases is GG3 in (a) and (b) and GG5 in (c).

4.2.3. Validation on External Databases

With the purpose of testing the generalization capability of the trianed model, $FSConv$ net was validated on two external databases. The databases used were shared by Arvaniti et al. [19] and Gerytch et al. [13]. The first database is composed of 886 cores from Tissue-Micro Arrays digitised at 40× magnification, and the second has 625 patches of prostate histology images at 20× magnification. Each core was resized to 10× resolution and a central patch with dimensions 512^2 was extracted. For both databases, the ground truth was generated following the procedure in [19]. Non-cancerous patches were ex-

26

tracted from images with only benign structures annotated, labels GG3, GG4, and GG5 were assigned to patches with only the corresponding grade annotated. Examples of the obtained images from the Arvaniti et al. and Gerytch et al. databases are presented in the first and second rows of Figure 10, respectively. Note that the H&E stain color images are different from those appearing in the SICAPv2 database (see Figure 1 for examples of the images used to train the developed models). To normalise the colour distribution of the images in external databases, the method presented in [37] was used after applying a channel-wise histogram matching of the external images to a SICAPv2 database reference image. This image was selected by the expert pathologists involved in this work based on its structural and colour properties. Then, our best performing model, i.e. *FSConv*, was used to predict and evaluate our performance on the external databases. Table 7 and Figure 11 show the obtained figures of merit and confusion matrices, respectively.
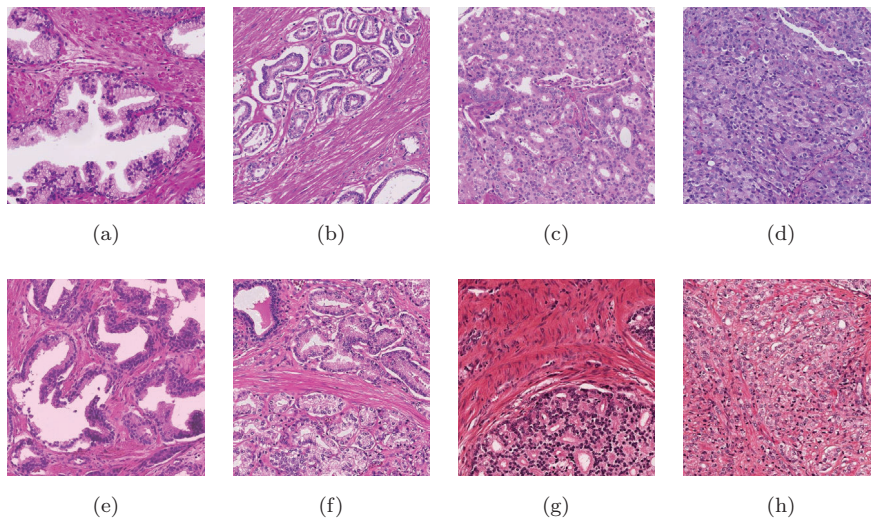

Figure 10: Examples of patches used from the external database from Arvaniti et al. (first row) and Gerytch et al. (second row). (a) and (e): Benign glands; (b) and (f): Patches containing GG3 patterns; (c) and (d): Patches containing GG4 patterns; (d) and (h): Patches containing GG5 patterns.

27

Table 7: Results of the patch-level Gleason grading in the Arvaniti and Gerytch databases by our proposed model, $FSConv$. The metrics presented are accuracy (ACC), F1-Score (F1S), computed per class and its average, and Cohen's quadratic kappa (κ).

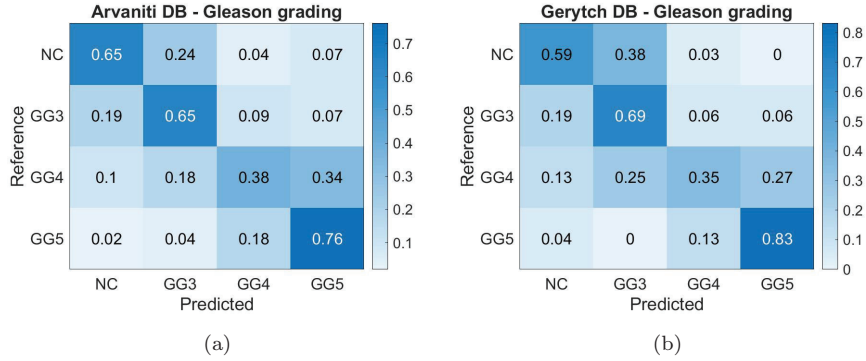| Database	ACC	F1S				Avg-F1S	κ
		NC	GG3	GG4	GG5		
Arvaniti et al. [19]	0.5861	0.5660	0.6858	0.4688	0.5603	0.5702	0.6410
Gerytch et al. [13]	0.5136	0.2901	0.6162	0.4990	0.4958	0.4753	0.5116

Figure 11: Confusion Matrix of the patch-level Gleason grades prediction in external databases using the proposed $FSConv$ model. (a): Arvaniti database and (b): Gerytch database.

The obtained results in Arvaniti et al. database were slightly worse than the ones reached in our test cohort. The macro-averaged F1 score was 0.57, while 0.65 was obtained in the test cohort (see Table 6). To the best of the authors's knowledge, this is the first time in the literature that a model trained for patch-level Gleason grading in tested on an external database. This is a challenging task, taking into account the known inter-pathologist variability of the Gleason grading task and the differences in the histology sample preparation. Thus, the difference in the results could be explained by those factors. In comparison to the results obtained in [19] on this database, the reported κ in the test subset was 0.55 (see Table 6), while the κ obtained by our model was 0.64. Our proposed model outperforms the current state of the art on this set of images, even though we used the whole database for testing, and they reported the

28

result on a specific test subset.

₅₂₀ Regarding the obtained results on the Gerytch et al. database, a macro-averaged F1 score of 0.47, and a κ of 0.51 were obtained. Note that the small amount of non cancerous patches in this database (32 patches with only benign annotation, compared to 116 in Arvaniti et al. set) could be negatively affecting the figures of merit. Unfortunately, to the best of the authors' knowledge, no ₅₂₅ work has been reported on the use of the entire set of grades on this database, which makes the comparison impossible.

4.3. Cribriform Pattern Detection

To detect cribriform patterns in GG4 patches, $FSConv$ trained in the Gleason grading stage was re-trained as specified in the subsection 3.2 with a learning ₅₃₀ rate of 0.001 and a batch size of 32 samples during 200 epochs. The results were optimised freezing the weights of the convolutional filters at different depths. Concretely, at filters $conv_1$, $conv_2$ and $conv_3$ (see Table 2 for $FSConv$ architecture details). The output probability of each model was used to compute the Receiver Operative Curve (ROC) and evaluate the Area Under Curve (AUC). ₅₃₅ Then, probabilities were thresholded to output a positive classification when they are above 50%. The results obtained for the cross-validation set are presented in Table 8, and the Receiver-Operative-Curve in Figure 12 (a).

Table 8: Results in the detection of cribriform pattern in the validation set. The accuracy (ACC), Sensitivity, specificity and area under ROC curve (AUC) are presented for the fine-tuned $FSConv$ model freezing up to the convolutional layers $conv_1$, $conv_2$ or $conv_3$.

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

The best results were obtained for the validation set by the network whose weights were frozen up to the layer $conv_2$. Thus, just the last layer, $conv_3$ ₅₄₀ and the output neuron were trained. The accuracy obtained through this configuration was 0.8225, with a sensitivity and specificity of 0.8993 and 0.5223,
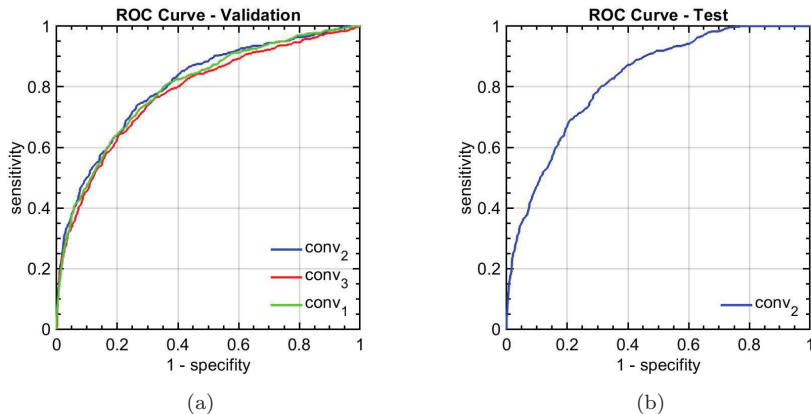
Figure 12: ROC curves obtained for cribriform pattern detection in samples with Gleason grade 4.

respectively. The reached AUC was 0.8225. Slightly better results were obtained by this model in the test subset. The ROC computed in the test subset is presented in Figure 12 (b), and it encloses an AUC of 0.8240. This value is
₅₄₅ at the permissible confidence level of systems for medical applications, above 0.80 [35]. Although the accuracy value decreases to 0.7239, the sensitivity and specificity are more balanced, with values 0.7168 and 0.7586, respectively. To the best of the authors's knowledge, this is the first time that the detection of cribriform patterns in histology prostate images is addressed and evaluated,
₅₅₀ so that it is not possible to establish comparison with previous works. Nevertheless, the studies comparing the inter-observer variability of the Gleason patterns classification show the challenging character of this task. In [38] the reproducibility in this problem was studied with 23 genitourinary pathologists. The consensus was achieved for cribriform glands in only 23% of the cases, and
₅₅₅ a consensus was not reached in how to classify the complex fused glands with cribriform shapes. We observed that the misclassified instances in our approach were mainly due to this kind of pattern. In Figure 13 few representative examples are presented, being (d), (e), and (f) images with complex fused glands that the model misclassified as cribriform pattern. Therefore, the results obtained

30

by the model are auspicious, and its main limitation is the misclassification of patterns with large inter-pathologist variability.
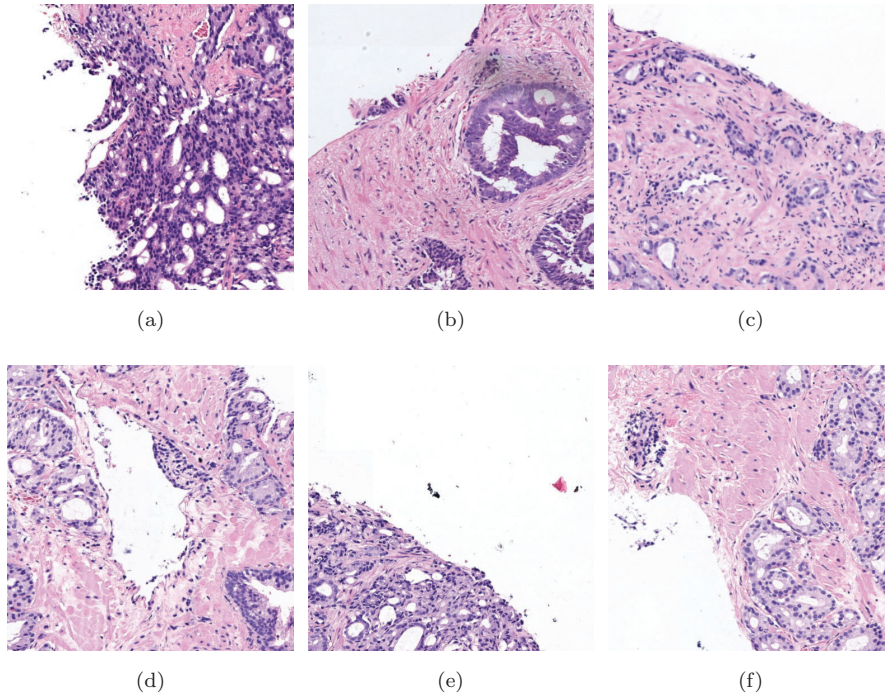

Figure 13: Examples of the system performance in the test subset for cribriform pattern detection. (a): True Positive, (b): True Positive, (c): True Negative, (d): False Positive, (e): False Positive, (f): False Positive.

4.4. WSI-Level Gleason Scoring

Once the patch-level prediction is performed with model $FSConv$, the probability maps for each Gleason grade are obtained, as specified in the subsection 3.3. The usability of these maps in the clinical practice were qualitatively validated by expert pathologists with satisfactory results.

Different examples of the test subset are presented in Figures 14, 15, and 16. These figures are organised as follows: in the first column, the WSI with pixel-level annotations (a) and pixel-level predictions (b) are presented, while in the

31

second, the heatmaps of GG3 (c), GG4 (d) and GG5 (e) are shown from top to bottom, respectively. The regions of interest in the WSIs are highlighted with a higher resolution window to facilitate visualisation. The example in Figure 14 is a biopsy with Gleason score $3 + 4 = 7$, the biopsy in Figure 15 corresponds to a $3 + 3 = 6$ sample and the case in Figure 16, $5 + 5 = 10$. Finally, a non-cancerous case is presented in Figure 17.
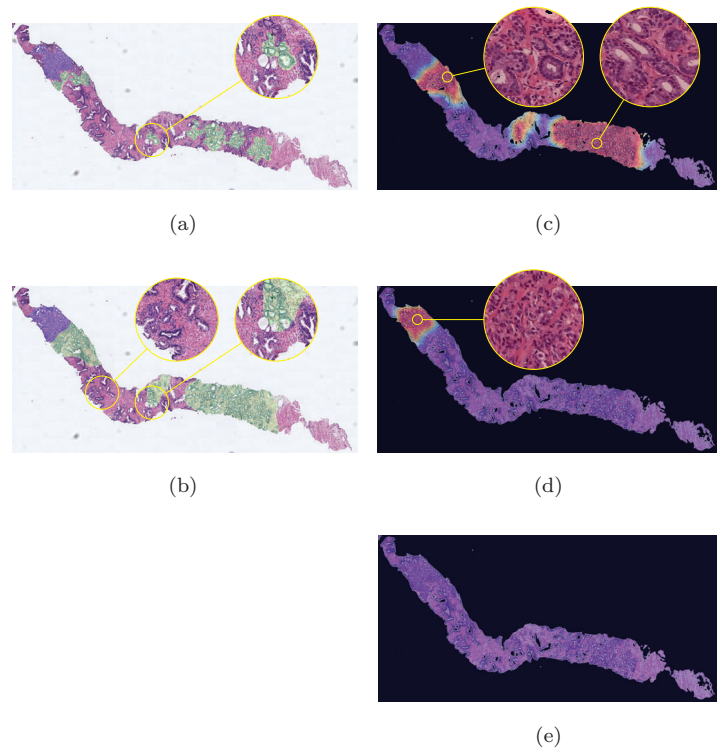

(a)

(c)

(b)

(d)

(e)

Figure 14: Whole slide image level prediction of a biopsy diagnosed as Gleason Score $3+4 = 7$. (a): manual annotations, (b): system predictions. Green: GG3, Blue: GG4, red: GG5. (c): GG3 heatmap, (d): GG4 heatmap, (e): GG5 heatmap.

In the case presented with Gleason score $3 + 4 = 7$ (see Figure 14), the GG3 and GG4 regions are correctly classified. In a subsequent review of this case, pathologists detected that some glands in the right region without pathologist's annotations in the ground truth and classified as GG3 by the model were
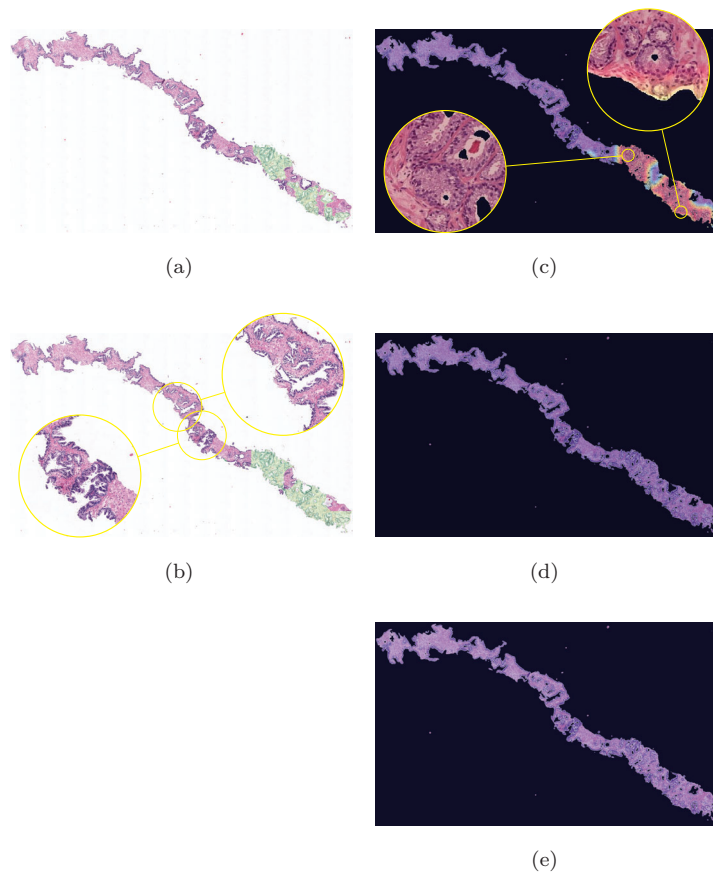
32

Figure 15: Whole slide image level prediction of a biopsy diagnosed as Gleason Score $3+3 = 6$. (a): manual annotations, (b): system predictions. Green: GG3, Blue: GG4, red: GG5. (c): GG3 heatmap, (d): GG4 heatmap, (e): GG5 heatmap.
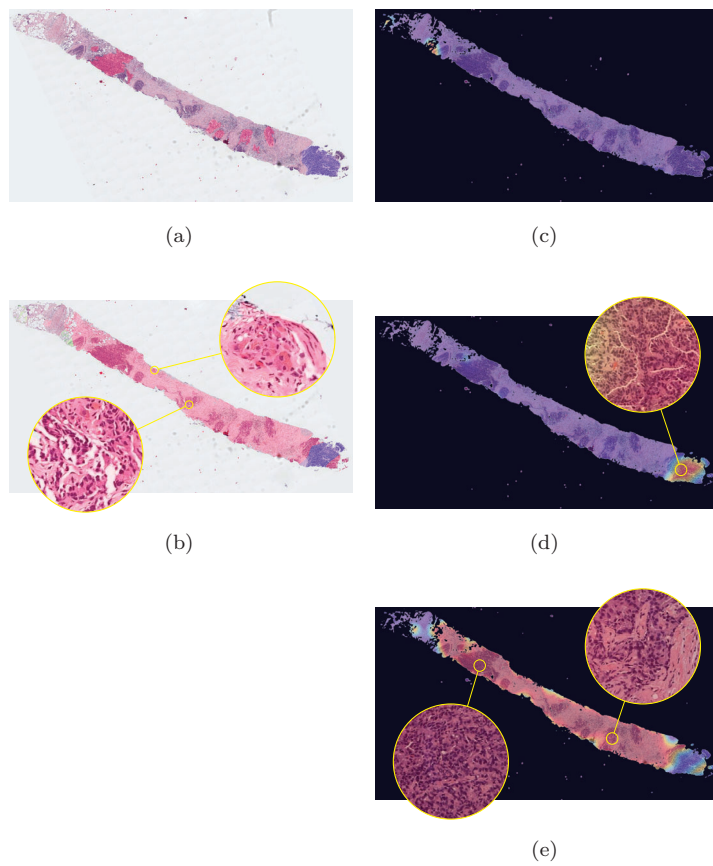
(a)

(c)

(b)

(d)

(e)

Figure 16: Whole slide image level prediction of a biopsy diagnosed as Gleason Score 5+5 = 10 (a): manual annotations, (b): system predictions. Green: GG3, Blue: GG4, red: GG5. (c): GG3 heatmap, (d): GG4 heatmap, (e): GG5 heatmap.
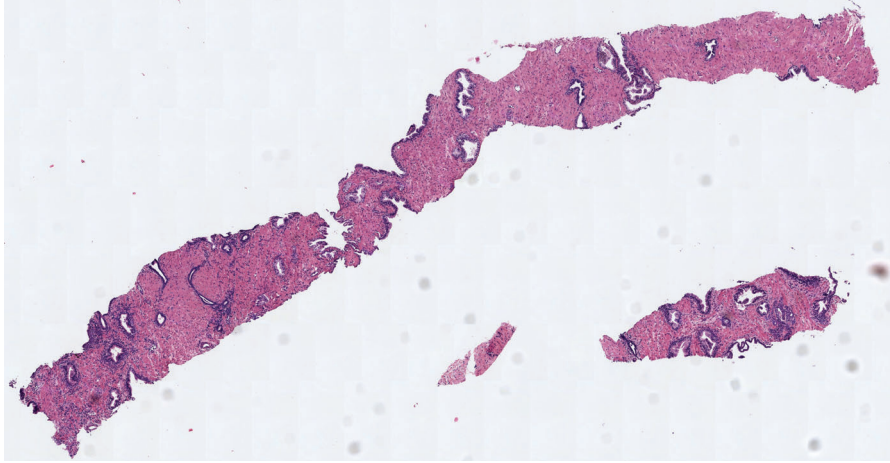
Figure 17: Non-cancerous biopsy without Gleason grades detected by the model.

actually cancerous patterns. Additionally, the few non-cancerous dilated and fusiform glands were correctly classified as non-cancerous (see Figure 14 (b), regions of interest highlighted). Regarding the biopsy with Gleason score of $3 + 4 = 7$, the model correctly detects the region with GG3 glands, but due to the patch resolution (512^2 pixels) some nearby stroma regions are highlighted as cancerous. Finally, analysing the case with a score of $5 + 5 = 10$, a papilar GG4 pattern is being correctly detected. The same occurs in the GG5 regions with isolated cells and pseudorosetting patterns. Nevertheless, in regions with a score of $GS \geq 9$ some stroma regions are frequently highlighted as GG5 by the model. This phenomenon does not occur in stroma of biopsies with $GS < 9$, as can be seen in the other cases. This fact suggests that the model could be detecting some hidden pattern of interest in the structure of the stroma in these regions.

Then, the percentages corresponding to each grade per WSI were obtained as specified in the methodology (Section 3.3). The proposed architecture MLP was then trained using as input the percentages obtained in the cross-validation subset. Adam optimiser was used, with a learning rate of 0.01, and a constant decay to zero over the 2000 epochs. The batch size was 32. The training strategy

35

was leave-one-out.

This proposed approach is compared with the method proposed by Arvaniti [19] using $T = 10\%$ as minimum number of pixels with a certain label to be consider the corresponding grade in the WSI grading. The confusion matrix at biopsy level obtained for both methods is presented in Figure 18, and Cohen's quadratic kappa (κ) was calculated as a figure of merit.
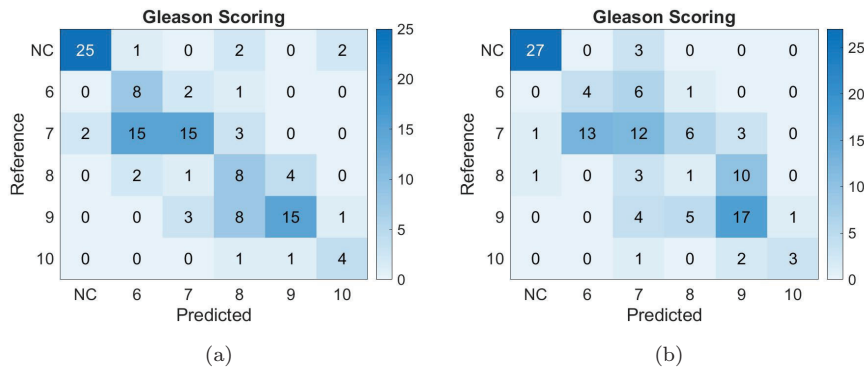

Figure 18: Confusion Matrix of the whole slide image level Gleason scoring in the validation cohorts. (a): Method proposed in [19]; (b): MLP model.

The κ value obtained for Arvaniti's approach was 0.7693, in line with the results presented in [19] using their own database (using TMAs), where the obtained κ value was 0.75. Better results were obtained with the proposed model MLP (see Figure 18 (b)), obtaining a κ value of 0.8177. The main difference between methods was observed in few samples misclassified as Gleason score 8 and Gleason score 10 by Arvaniti's proposal which were correctly classified by our model. Therefore, our proposed strategy seems to model better the pathologist's decision to assign a Gleason score to the full image of the slide than the previous scoring methodology. The results obtained in the test subset by MLP model are similar to those obtained for the validation cohorts, with a κ value 0.8168.

5. Conclusions and future work

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

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

37

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

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

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

References

[1] World Cancer Research Foundation, Prostate cancer statistics (2019).
    URL http://www.wcrf.org

[2] World Health Organization, Global cancer observatory (2019).
    URL http://gco.iarc.fr

[3] D. F. Gleason, Histologic grading of prostate cancer: A perspective, human pathology (1992).

[4] J. Gordetsky, J. Epstein, Grading of prostatic adenocarcinoma: Current state and prognostic implications, Diagnostic Pathology 11 (1) (2016) 2–9. doi:10.1186/s13000-016-0478-2.

[5] J. I. Epstein, L. Egevad, M. B. Amin, B. Delahunt, J. R. Srigley, P. A. Humphrey, The 2014 international society of urological pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma definition

of grading patterns and proposal for a new grading system, American Journal of Surgical Pathology 40 (2) (2016) 244–252. `doi:10.1097/PAS.0000000000000530`.

[6] M. Sharma, H. Miyamoto, Percent Gleason pattern 4 in stratifying the prognosis of patients with intermediate-risk prostate cancer, Translational Andrology and Urology 7 (Suppl 4) (2018) S484–S489. `doi:10.21037/tau.2018.03.20`.

[7] O. Hassan, A. Matoso, Clinical significance of subtypes of Gleason pattern 4 prostate cancer, Translational Andrology and Urology 7 (Suppl 4) (2018) S477–S483. `doi:10.21037/tau.2018.02.06`.

[8] T. van der Kwast, G. J. van Leenders, On cribriform prostate cancer, Translational Andrology and Urology 7 (2018) 145–154. `doi:10.21037/tau.2017.12.33`.

[9] H. Remotti, Tissue microarrays: Construction and use, Methods in Molecular Biology`doi:10.1007/978-1-62703-287-2`.

[10] M. H. Khouja, M. Baekelandt, A. Sarab, J. M. Nesland, R. Holm, Limitations of tissue microarrays compared with whole tissue sections in survival analysis, Oncology letters 1 (2010) 827–831. `doi:10.3892/ol{\_}00000145`.

[11] D. Voduc, C. Kenney, T. O. Nielsen, Tissue Microarrays in Clinical Oncology, Bone 18 (1) (2009) 88–97. `doi:10.1038/jid.2014.371`.

[12] S. Doyle, M. Hwang, K. Shah, A. Madabhushi, M. Feldman, J. Tomaszeweski, Automated grading of prostate cancer using architectural and textural image features, 2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro - Proceedings (2007) 1284–1287`doi:10.1109/ISBI.2007.357094`.

[13] A. Gertych, N. Ing, Z. Ma, T. J. Fuchs, S. Salman, S. Mohanty, S. Bhele, A. Velásquez-Vacca, M. B. Amin, B. S. Knudsen, Machine learning ap-

39

proaches to analyze histological images of tissues from radical prostatectomies, Computerized Medical Imaging and Graphics 46 (2015) 197–208. `doi:10.1016/j.compmedimag.2015.08.002`.

[14] O. Jiménez del Toro, M. Atzori, S. Otálora, M. Andersson, K. Eurén, M. Hedlund, P. Rönnquist, H. Müller, Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade Gleason score, Medical Imaging 2017: Digital Pathology 10140 (2017) 101400O. `doi:10.1117/12.2255710`.

[15] J. Ren, E. Sadimin, D. J. Foran, X. Qi, Computer aided analysis of prostate histopathology images to support a refined Gleason grading system, Medical Imaging 2017: Image Processing 10133 (2017) 101331V. `doi:10.1117/12.2253887`.

[16] N. Ing, Z. Ma, J. Li, H. Salemi, C. Arnold, B. S. Knudsen, A. Gertych, Semantic segmentation for prostate cancer grading by convolutional neural networks, Digital Pathology Conference 1 (June) (2018) 46. `doi:10.1117/12.2293000`.

[17] A. E. Esteban, M. López-Pérez, A. Colomer, M. A. Sales, R. Molina, V. Naranjo, A new optical density granulometry-based descriptor for the classification of prostate histological images using shallow and deep Gaussian processes, Computer Methods and Programs in Biomedicine 178 (2019) 303–317. `doi:10.1016/j.cmpb.2019.07.003`.

[18] M. Lucas, I. Jansen, C. D. Savci-Heijink, S. L. Meijer, O. J. de Boer, T. G. van Leeuwen, D. M. de Bruin, H. A. Marquering, Deep learning for automatic Gleason pattern classification for grade group determination of prostate biopsies, Virchows Archiv 475 (1) (2019) 77–83. `doi:10.1007/s00428-019-02577-x`.

[19] E. Arvaniti, K. S. Fricker, M. Moret, N. Rupp, T. Hermanns, C. Fankhauser, N. Wey, P. J. Wild, J. H. Rüschoff, M. Claassen, Automated

₇₃₀

Gleason grading of prostate cancer tissue microarrays via deep learning, Scientific Reports 8 (1) (2018) 1–11. `doi:10.1038/s41598-018-30535-1`.

[20] G. Nir, S. Hor, D. Karimi, L. Fazli, B. F. Skinnider, P. Tavassoli, D. Turbin, C. F. Villamil, G. Wang, R. S. Wilson, K. A. Iczkowski, M. S. Lucia, P. C. Black, P. Abolmaesumi, S. L. Goldenberg, S. E. Salcudean, Automatic grading of prostate cancer in digitized histopathology images: Learning ₇₃₅ from multiple experts (2018). `doi:10.1016/j.media.2018.09.005`.

[21] G. Nir, D. Karimi, S. L. Goldenberg, L. Fazli, B. F. Skinnider, P. Tavassoli, D. Turbin, C. F. Villamil, G. Wang, D. J. Thompson, P. C. Black, S. E. Salcudean, Comparison of Artificial Intelligence Techniques to Evaluate Performance of a Classifier for Automatic Grading of Prostate Cancer ₇₄₀ From Digitized Histopathologic Images, JAMA network open 2 (3) (2019) e190442. `doi:10.1001/jamanetworkopen.2019.0442`.

[22] G. García, A. Colomer, V. Naranjo, First-stage prostate cancer identification on histopathological images: Hand-driven versus automatic learning, Entropy 21 (4). `doi:10.3390/e21040356`.

₇₄₅ [23] Y. Ma, Z. Jiang, H. Zhang, F. Xie, Y. Zheng, H. Shi, Y. Zhao, J. Shi, Generating region proposals for histopathological whole slide image retrieval, Computer Methods and Programs in Biomedicine 159 (2018) 1–10. `doi:10.1016/j.cmpb.2018.02.020`.

[24] W. Li, J. Li, K. V. Sarma, K. C. Ho, S. Shen, B. S. Knudsen, A. Gertych, ₇₅₀ C. W. Arnold, Path R-CNN for Prostate Cancer Diagnosis and Gleason Grading of Histological Images, IEEE Transactions on Medical Imaging 38 (4) (2019) 945–954. `doi:10.1109/TMI.2018.2875868`.

[25] J. N. Weintein, E. A. Collisson, G. B. Mills, k. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, The Cancer Genome ₇₅₅ Atlas Pan-Cancer Analysis Project, NIH Public Access 4518 (2013) 219–223. `doi:10.1038/ng.2764`.

[26] Openseadragon, `http://openseadragon.github.io/`, accessed: 10-07-2018.

[27] M. Lin, Q. Chen, S. Yan, Network In Network, International Conference of Learning Representations (2014) 1–10.
URL `http://arxiv.org/abs/1312.4400`

[28] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, International Conference on Learning Representations 1 (2014) 1–14.
URL `http://arxiv.org/abs/1409.1556`

[29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2016-Decem (2016) 770–778. `doi:10.1109/CVPR.2016.90`.

[30] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the Inception Architecture for Computer Vision, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2016-Decem (2016) 2818–2826. `doi:10.1109/CVPR.2016.308`.

[31] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 2017-Janua (2017) 2261–2269. `doi:10.1109/CVPR.2017.243`.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database, 2009 IEEE Conference on Computer Vision and Pattern Recognition`doi:10.1109/CVPR.2009.5206848`.
URL `http://arxiv.org/abs/1409.1556`

[33] J. Cohen, Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit, Psychological Bulletin 70 (4) (1968) 213–220. `doi:10.1037/h0026256`.

[34] M. L. McHugh, Interrater reliability: the kappa statistic, Lessons in biostatistics 22 (3) (2012) 276–282.

[35] J. A. Swets, Measuring the Accuracy of Diagnostic Systems, Science 240, (4857) (1988) 1285–1293. `doi:10.1126/science.3287615`.

[36] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning Deep Features for Discriminative Localization, Proceedings of the IEEE conference on computer vision and pattern recognition (2016) 2921–2929.

[37] A. Vahadane, T. Peng, S. Albarqouni, M. Baust, K. Steiger, A. M. Schlitter, A. Sethi, I. Esposito, N. Navab, Structure-preserved color normalization for histological images, Proceedings - International Symposium on Biomedical Imaging 2015-July (2015) 1012–1015. `doi:10.1109/ISBI.2015.7164042`.

[38] C. F. Kweldam, D. Nieboer, F. Algaba, M. B. Amin, D. M. Berney, A. Billis, D. G. Bostwick, L. Bubendorf, L. Cheng, E. Compérat, B. Delahunt, L. Egevad, A. J. Evans, D. E. Hansel, P. A. Humphrey, G. Kristiansen, T. H. van der Kwast, C. Magi-Galluzzi, R. Montironi, G. J. Netto, H. Samaratunga, J. R. Srigley, P. H. Tan, M. Varma, M. Zhou, G. J. van Leenders, Gleason grade 4 prostate adenocarcinoma patterns: an interobserver agreement study among genitourinary pathologists, Histopathology 69 (3) (2016) 441–449. `doi:10.1111/his.12976`.