IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

PROMETEO: A CNN-based computer-aided diagnosis system for WSI prostate cancer detection

L. DURAN-LOPEZ[1], JUAN P. DOMINGUEZ-MORALES[1],A. F. CONDE-MARTIN[2], S. VICENTE-DIAZ[1], and A. LINARES-BARRANCO[1] (Senior Member, IEEE)
[1]Robotics and Technology of Computers Lab. University of Seville, Sevilla, Spain
[2]Pathological Anatomy Unit, Virgen de Valme Hospital. Seville, Spain

Corresponding author: L. Duran-Lopez (e-mail: lduran@atc.us.es).

This work was partially supported by the Spanish grant (with support from the European Regional Development Fund) COFNET (TEC2016-77785-P), and by the Andalusian Regional Project PAIDI2020 (with FEDER support) PROMETEO (AT17_5410_USE).

ABSTRACT

Prostate cancer is currently one of the most commonly-diagnosed types of cancer among males. Although its death rate has dropped in the last decades, it is still a major concern and one of the leading causes of cancer death. Prostate biopsy is a test that confirms or excludes the presence of cancer in the tissue. Samples extracted from biopsies are processed and digitized, obtaining gigapixel-resolution images called whole-slide images, which are analyzed by pathologists. Automated intelligent systems could be useful for helping pathologists in this analysis, reducing fatigue and making the routine process faster. In this work, a novel Deep Learning based computer-aided diagnosis system is presented. This system is able to analyze whole-slide histology images that are first patch-sampled and preprocessed using different filters, including a novel patch-scoring algorithm that removes worthless areas from the tissue. Then, patches are used as input to a custom Convolutional Neural Network, which gives a report showing malignant regions on a heatmap. The impact of applying a stain-normalization process to the patches is also analyzed in order to reduce color variability between different scanners. After training the network with a 3-fold cross-validation method, 99.98% accuracy, 99.98% F1 score and 0.999 AUC are achieved on a separate test set. The computation time needed to obtain the heatmap of a whole-slide image is, on average, around 15 s. Our custom network outperforms other state-of-the-art works in terms of computational complexity for a binary classification task between normal and malignant prostate whole-slide images at patch level.

INDEX TERMS Convolutional Neural Networks, computer-aided diagnosis, deep learning, medical image analysis, prostate cancer, whole-slide images

I. INTRODUCTION

PROSTATE cancer is the third most commonly-diagnosed non-skin cancer and one of the leading causes of cancer death among males [1], with more than 1.25 million new cases in 2018 (7.5% of total cancer cases) and around 359k deaths worldwide (3.8% of the number of deaths of that year). With these high numbers, finding a way to improve current diagnosis systems is crucial, along with reducing the diagnosis time.

The prostate is a gland in the male reproductive system in most mammals that lies below the bladder, and whose main function is to secrete prostate fluid. Prostate cancer begins when cells of this gland grow uncontrollably. Then, these cells could invade surrounding tissues and organs (a process called infiltration) or spreading to other parts of the body (metastasis) [2].

Generally, the diagnostic steps for this cancer consists, firstly, in the realization of a digital rectal examination, and a determination of PSA (Prostate-Specific Antigen) levels in blood. In the case that the doctor finds anomalies in digital rectal examination or PSA results, a prostate biopsy is performed. This test would confirm or exclude the presence of cancer in the prostate tissue [3]. Prostate biopsy consists in obtaining samples of the prostate tissue using a needle that performs the puncture from a region that is determined through a transrectal ultrasound process. Then, these tissue

TABLE 1. Comparative study between state-of-the-art research about prostate cancer diagnosis.

Ref.	Dataset	Preprocessing step	Classifier	Classes	Performance measure
[10]	4 TMAs[1]: - Train: 73 (cancer) + 89 (normal) cores - Test: 217 (cancer) + 274 (normal) cores	Otsu's thresholding, Euclidean distance and Watershed algorithm to perform nuclear seed detection. Nuclear seed maps are used as input to the classifier.	CNN[2] (custom)	2: Cancer and normal	AUC[3] at core level: 0.974
[11]	235 WSIs[4]: - Train: 282k patches - Val: 94k patches - Test: 92k patches	Binary tissue mask obtained by Blue Ratio image to remove background.	CNN (GoogLeNet)	2: High and low GGS[5]	ACC[6] at patch level: 78.2% on test 73.52% on validation
[12]	225 WSIs: - Train: 48 (cancer) + 52 (normal) WSIs - Val: 31 (cancer) + 19 (normal) WSIs - Test: 45 (cancer) + 30 (normal) WSIs	Binary tissue mask by applying thresholding procedure based on optical density of RGB channels to remove background.	CNN (custom)	2: Cancer and normal	AUC at slide level: 0.99
[13]	513 WSIs: - Train: not specified - Test: not specified	Normalize procedure to eliminate stain variability.	R-CNN[7] (custom)	4: Stroma, benign glands, low GGS, high GGS	IOU[8] *: 79.56% OPA[9] *: 89.40% SMA[10] *: 88.78% *at tile level (set of patches)
[14]	54 patches: - Train: not specified - Test: not specified	Extraction of architectural features. Extraction of 1st-order statistical features with the average, median, standard deviation and range of the pixel values. Extraction of 2nd-order statistical features (Haralick features) from a co-occurrence matrix.	SVM[11]	2, but with different classes: Epithelium vs stroma GGS 3 vs GGS 4 GGS 3 vs epithelium GGS 3 vs stroma GGS 4 vs epithelium GGS 4 vs stroma	ACC at patch level: Epithelium vs stroma: 76.9% GGS 3 vs GGS 4: 76.9% GGS 3 vs epithelium: 85.4% GGS 3 vs stroma: 92.8% GGS 4 vs epithelium: 88.9% GGS 4 vs stroma: 89.7%
[15]	22 WSIs - Train: 17 WSIs - Test: 5 WSIs	Segmentation procedure with CNN and superpixel segmentation. Feature extraction with Bag-of-Word to remove background.	RFC[12]	2: GGS 3 and GGS 4	F1-score*: 0.8460 Sensitivity*: 0.70±0.15 Specificity*: 0.89±0.04 ACC*: 0.83±0.03 *at patch level
[16]	24859 WSIs: - Train: 70% - Val: 15% - Test: 15%	Otsu's thresholding to remove background.	CNN (ResNet34) + RNN[13]	2: Tumor and normal	AUC at slide level: 0.986
[17]	8914 WSIs: - Train: 6953 WSIs - Val: 1631 WSIs - Test: 330 WSIs	Segmentation algorithm based on Laplacian filtering.	CNN (60 Inception V3)	2: Normal and malignant 3: GGS 3, GGS 4 and GGS 5	AUC for normal and malignant*: 0.997 on validation 0.986 on test Mean pairwise kappa for GGS*: 0.62 *at slide level
[18]	1243 WSIs: - Train: 933 WSIs - Val: 100 WSIs - Test: 210 WSIs	Tissue segmentation network for extracting tissue from background. Tumor detection system to define the tumor and epithelial tissue detection system to label the images.	CNN (U-Net)	6: Benign, GGG[14] 1-5.	AUC at slide level: Benign vs malignant: 0.990 Benign and GGG 1 vs GGG≥2: 0.978 Benign and GGG 1-2 vs GGG≥3: 0.974

[1]: Tissue Microarray. [2]: Convolutional Neural Network. [3]: Area Under Curve. [4]: Whole Slide Tissue Image. [5]: Gleason Grade Score. [6]: Accuracy. [7]: Region-based Convolutional Neural Network. [8]: Intersection Over Union. [9]: Overall Pixel Accuracy. [10]: Standard Mean Accuracy. [11]: Support Vector Machine. [12]: Random Forest Classifier. [13]: Recurrent Neural Network. [14]: Gleason Grade Group.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3008868, IEEE Access

samples are processed in a laboratory and scanned, resulting on gigapixel-resolution images called Whole-Slide Images (WSIs), which are then analyzed and inspected by pathologists.

The aggressiveness of prostate cancer could be determined through a scoring system called the Gleason Grading System (GGS) [4], which ranges from 1 to 5 and describes how much the cancer from a biopsy resembles normal tissue when analyzed. Pathologists observe the structure of the cells in WSIs and assign a lower or higher score depending on whether the appearance is that of healthy or abnormal tissue, respectively. This grade is later used by the doctor to assign the most suitable treatment for the patient. However, many studies have reported interobserver variability among pathologists in the process of labeling the cancerous sections of the tissue (more than 30% degree of discrepancy in the score) [3] [5] [6].

To avoid this problem, and also to reduce pathologists' fatigue when analyzing WSIs, artificial intelligence could play an important role in this field. This emerging topic has proved its potential in image diagnostic tasks such as radiology [7], dermatology [8] and histopathology [9], among others. To this respect, Computer-Aided Diagnosis (CAD) systems have gained popularity in recent years. CAD systems are automatic or semi-automatic algorithms whose main goal is to assist doctors when making an interpretation of medical images. Recently, many researchers have investigated the application of this kind of systems to the diagnosis of prostate cancer based on different methodologies. Some of these studies use machine learning techniques, such as neural networks, Support Vector Machines (SVMs), or some complex algorithms to carry out the classification [10]–[19], while others are based on algebraic tools, such as Homology Profile algorithms, which extracts features from a structure of a topological space [20]. Many of them have performed a binary classification [10]–[12], [14]–[16], distinguishing between cancerous and normal tissue or between different GGS scores, whereas others have performed a multi-class detection [13], [17]–[19].

For this kind of systems, preprocessing the information could be a key factor to make it easier for the classifier to extract the most relevant features from the input images. Background and noise removal are key processes to consider when working with histopathological images. Otsu's thresholding [21] is one of the most well-known and used methods for extracting background and tissue from WSIs [10], [19]. In [11], the Blue Ratio method, which detects nuclei from cells in stained images, is used to obtain tissue regions. Other simpler mechanisms to remove background are based on thresholding procedures on the optical density of the RGB channels [12].

Stain normalization has also proved to be useful for histopathological images, since they reduce color variations that could have been produced in the staining process of the tissue sample [22]. This has been used in different cancer studies based on histopathological images [23]. In [24], the authors compared the effect of applying different stain normalization methods in histopathological images for liver, breast, kidney and colorectal cancer.

Table 1 presents a comparison of some of these studies, summarizing the characteristics of the dataset, the preprocessing step applied to the data, the main classification method procedure of the CAD system, the number of classes taken into account and the results obtained with their corresponding performance measure. These works have used many different techniques for the preprocessing step, although apart from [14], which uses SVMs, the rest have performed either the classification or part of the preprocessing by using Convolutional Neural Networks (CNNs). These complex architectures have increased in popularity in the recent years thanks to the rise in the computation capabilities of current general purpose computers, reducing the gap of achieving a robust and accurate CAD system.

In this work, a novel deep-learning-based CAD system for prostate cancer detection in WSI images is presented to support pathologists in this task. A CNN was trained and tested over a new dataset that was built and labeled with the supervision of expert pathologists after processing the images with novel algorithms to improve cancer detection and robustness across WSIs from different hospitals and scanners.

The main contributions of this work include the following:

- A novel filter to discriminate areas without tissue including noise and external agents.
- A comparative study based on the application of stain normalization in prostate WSI images.
- A 9-layer custom CNN model, trained and validated from scratch, to reduce the computation time needed to process WSIs with competitive accuracy.

The rest of the paper is structured as follows: in section II, the materials and methods are presented, focusing on the dataset that was used for this work (including preprocessing and data augmentation), along with the neural network model that was trained and tested. Then, section III presents the results obtained in the system for different evaluation metrics, which are described. After this, discussions, some limitations of the proposed system and conclusions are presented in sections IV, V and VI, respectively.

II. METHODOLOGY

A. DATASET

Training a CNN requires a large amount of data to make the classifier learn and converge to the wanted solution. The lack of free and open datasets with the sufficient amount of samples, and with reliable labels associating the pixels in every image with a specific class, is always a restriction when trying to develop a CAD system for medical image analysis.

For this work, a novel dataset that was analyzed and labeled by expert pathologists was created. In this dataset, malignant regions of the WSIs considered by the pathologist for such diagnosis were specified. This kind of labels could

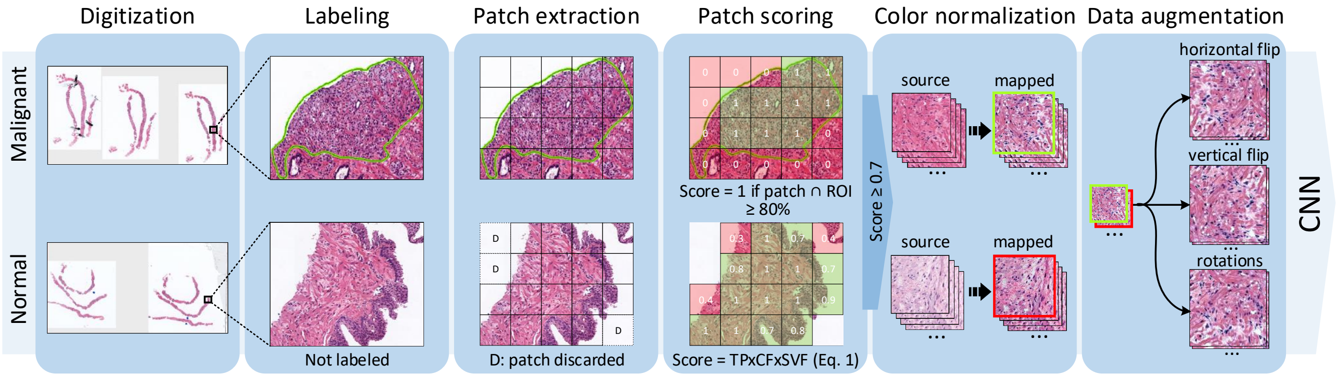


FIGURE 1. Flow chart of the whole dataset acquisition and the different preprocessing steps applied.

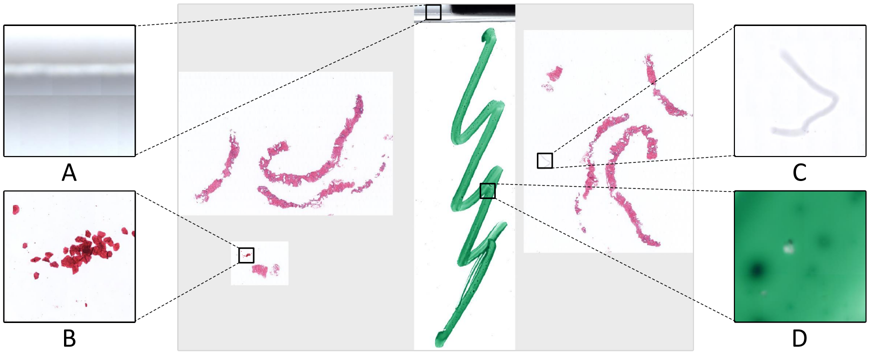


FIGURE 2. WSI with unwanted areas: regions which correspond to the edge of the slide cover (A), cells from external tissue not related to the prostate (B), external agents such as dirt (C) and zones highlighted with pen (D).

provide the necessary information to train a learning system in order to extract relevant features from the cell structures contained in them and, thus, detect specific patterns. This approach was not considered for other prostate cancer datasets (such as TCGA) that were used in some of the works mentioned in section I, in which a general diagnosis was given to WSIs instead of specifying malignant regions. These regions would provide much more information for the training of a CNN.

Figure 1 depicts the whole process applied for obtaining our dataset.

1) Data acquisition and labeling

To obtain a reliable dataset, a collaboration with the Pathological Anatomy Unit of Virgen de Valme Hospital in Seville (Spain) was established. They provided a large set of prostate cancer cases obtained from different patients. These cases consisted in different Hematoxylin and Eosin (H&E) stained slides (diagnosed as normal or malignant) obtained from needle core biopsy. Then, they were digitized with a VENTANA iScan HT[1] scanner from Roche Diagnostics.

Once the biopsies were scanned and digitized, the following step consisted in labeling the WSIs. To this end, a desktop software application was designed and developed in C# and Windows Presentation Foundation (WPF) with Microsoft ® .NET Framework with the purpose of categorizing specific regions of the tissue that are malignant. Using this application, experienced pathologists examined WSIs in order to find malignant areas, considered as Regions of Interest (ROIs), indicating the GGS score that they belong to, and thus, labeling each of the WSI images. For a more precise and comfortable labeling process, pathologists used computer drawing pads from Wacom® to mark the ROIs inside WSIs.

The essential attributes details for the dataset creation are summarized in Table 2.

2) Patch sampling

Due to the large size of the WSIs obtained from the process presented in section II-A1 (100k×100k pixels approximately), using them as a direct input for the CNN is not doable. To this end, these images were divided into small patches (100×100 pixels at 10× optical magnification) in order to obtain a dataset that the neural network could work with for the training, validation and testing steps, ensuring that all patches of a patient are only in one of these subsets. This division would also have some other effects. First of all, it would speed up the computation time for processing a com-

[1] https://diagnostics.roche.com/global/en/products/instruments/ventana-iscan-ht.html

TABLE 2. Dataset summary.

Attributes	Details
Staining method	Hematoxylin and Eosin stain
Scanner	VENTANA iScan HT from Roche Diagnostics
Scanner resolution	0.25 μm per pixel
Total number of WSIs	97
Optical magnification	10×

plete WSI, since unwanted areas such as noisy regions of the image or background would not be taken into account. Then, this would also increase the overall accuracy, robustness and reliability of the system, since more images would be considered for training the network. Finally, the CAD system would also be more precise in locating malignant areas of the tissue, which could better help pathologists, rather than just predicting if a whole WSI is malignant or not, as an unique and global diagnosis.

The quality of the dataset is crucial in the training step as well as when testing the network. The lesser number of noisy patches the dataset contains, the more robust and better fitted the training step of the network would be, leading to achieving better results. For these reasons, it is important to discard all unwanted regions (background and noisy regions) from the dataset and only consider areas which contain prostate tissue. Figure 2 shows some common noisy agents that could be present in WSIs.

To obtain the dataset, different patch-extraction algorithms were applied for WSIs labeled as normal and malignant. It is important to mention that patches labeled as normal were only obtained from WSIs diagnosed as normal, and patches labeled as malignant were obtained from ROIs of WSIs diagnosed with cancer, avoiding possible malignant tissue regions that the pathologist could have missed when labeling a malignant WSI. For malignant WSIs, the ROIs selected by the pathologists were framed with a polygon, which was then scanned by overlapping patches (with 50% overlap between them) as in [13] and [16], due to the smaller amount of malignant patches in comparison with the normal ones. Overlapping was only applied to malignant WSIs in the cross-validation set, and not in the test set (see section II-B2). Those patches which had at least 80% of its area within the ROI were considered, and the rest of them were discarded. For normal WSIs, all patches which contained tissue were extracted, following two consecutive processes: first, background patches were discarded based on an RGB value threshold, where patches with a mean color value close to either white or black were removed (below 30 and above 230, using a 8-bit color depth); then, patches corresponding to unwanted areas (noise) were discarded by applying a novel filter process based on Deron Eriksson's patch scoring formula[2].

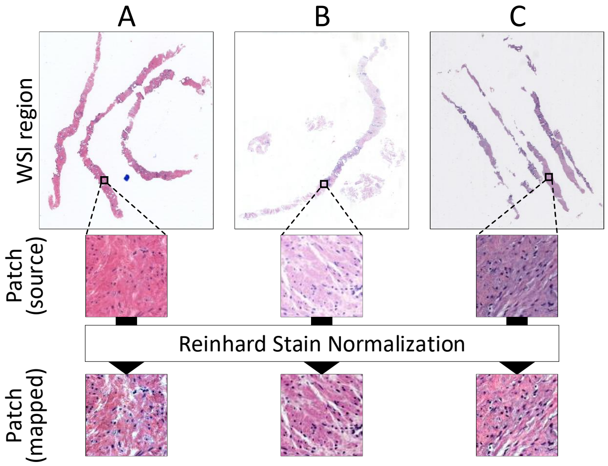


FIGURE 3. Examples of the application of Reinhard stain-normalization on three patches (source) from three WSIs (A, B and C) from different scanners, obtaining normalized patches (mapped).

This filter applies a score (in a scale that ranges from 0 to 1) to each extracted patch depending on three subfilters (see (1)). Following this score, if the patch exceeded a threshold (established at 0.7), then the patch was considered for the dataset, if not, it was discarded.

$$\text{Score} = \text{TP} \times \text{CF} \times \text{SVF} \tag{1}$$

Where TP stands for Tissue Percentage; CF, Color Factor; and SVF, Saturation and Value Factor.

TP measures the amount of tissue that the patch contains, scoring it from 0 to 1, by counting the number of pixels that do not correspond to background. The more tissue the patch contains, the higher the score it will be given.

CF measures (from 0 to 1) the area of the patch that is inside H&E's color range (which is between pink and blue, including purple, depending on whether the region is acidic or basic). For this, each of the patches were first converted from RGB to HSV scale, which consists of three channels: hue (H), saturation (S) and brightness/value (V). Then, the score assigned to CF depends on the percentage of pixels whose hue lie within H&E's color range.

SVF measures the dispersion (standard deviation) of the saturation and brightness channels of the patch after being converted to HSV scale. As patches which contain tissue have a medium-high dispersion due to their low uniformity, those that do not have tissue or that have a small amount of tissue score lower SVF.

The number of patches obtained from normal and malignant WSIs after applying the mentioned steps is shown in Table 3, where the GGS distribution is also reported. Around

[2] https://github.com/deroneriksson/python-wsi-preprocessing

TABLE 3. Dataset classes distribution.

Categories	No. of WSI	No. of patches
Malignant	70	19905, where: 6404 (32.17%) GGS 3 9791 (49.19%) GGS 4 3710 (18.64%) GGS 5
Normal	27	19772
Total	**97**	**39677**

50% of the total amount of patches correspond to normal, and the rest to malignant.

3) Preprocessing step

Histology images could present unwanted color variations caused by different factors such as the staining procedure that was performed, the equipment that was used for doing it and the color responses of digital scanners in the digitization process, among others. When comparing WSIs, their color could be very different even if the images are obtained from the same scanner. Therefore, color normalization methods, which reduce the variability of H&E stain appearance, could be useful to improve the classifier. This could also make the system more robust and stable when predicting or inferring over new unseen samples from different hospitals and scanners with which the network has not been trained with.

To this end, a color normalization processing, called Reinhard stain-normalization [25], [26], was applied. With this color normalization method, the mean and standard deviation of each channel of a source image are matched to that of a target image by applying a linear transformation in a perceptual colourspace (the $l\alpha\beta$ colourspace of [27]), obtaining the resulting mapped image. This process is defined by equations (2), (3) and (4).

$$l_{\text{mapped}} = \frac{l_{\text{source}} - \bar{l}_{\text{source}}}{\hat{l}_{\text{source}}} \hat{l}_{\text{target}} + \bar{l}_{\text{target}} \tag{2}$$

$$\alpha_{\text{mapped}} = \frac{\alpha_{\text{source}} - \bar{\alpha}_{\text{source}}}{\hat{\alpha}_{\text{source}}} \hat{\alpha}_{\text{target}} + \bar{\alpha}_{\text{target}} \tag{3}$$

$$\beta_{\text{mapped}} = \frac{\beta_{\text{source}} - \bar{\beta}_{\text{source}}}{\hat{\beta}_{\text{source}}} \hat{\beta}_{\text{target}} + \bar{\beta}_{\text{target}} \tag{4}$$

Where $\bar{l}$, $\bar{\alpha}$, and $\bar{\beta}$ are the channel means; $\hat{l}$, $\hat{\alpha}$, and $\hat{\beta}$ are the channel standard deviations (calculated over all the pixels in the image). This process was applied to every patch (source) in the dataset, considering target as the mean over all the patches in the training set (dashed purple in Figure 5). An example of the application of this process can be seen in Figure 3.

4) Data augmentation

In Deep Learning algorithms, the more images the dataset has, the more robust and stable the system will be. Also, having a larger dataset helps to avoid overfitting, since the network has more different data to train with. However, this is not always the case (e.g., adding more noisy samples will not help), and this is why having a clean dataset with region-specific labels is so important.

For this reason, data augmentation techniques were applied to our dataset in order to increase the number of images for the training step, and thus, to contemplate many other cases. Different transformations were performed to the original patches, thus, for each training patch, a horizontal flip and a vertical flip were applied, along with rotations in the whole 360° range with steps of 1°, where the missing information in the corners after rotating the patch was filled by mirroring. Therefore, we obtained 2×2×360 new patches from each original patch.

B. DEEP LEARNING FRAMEWORK

1) Convolutional Neural Network architecture

CNNs are a particular class of deep, feed-forward neural networks. They have become the most popular network architectures in the Deep Learning field due to their success on image processing and classification tasks [28]. The main difference between CNNs and other feed-forward neural networks is the application of convolution operations to extract features from the input image. In addition to the convolutions, CNNs have other kinds of layers that improve and accelerate the learning and the inference by downsampling the amount of data that is generated, as well as for the classification step.

In this work, PROMETEO, a custom CNN was developed to perform the prostate cancer detection task. It is a supervised neural network whose architecture is shown in Figure 4. This network consists of five convolution stages. A convolution stage consists of the following layers: convolution, batch normalization, rectified linear unit and 2×2 pooling. These 5 layers have 64 5×5, 64 3×3, 128 3×3, 128 3×3 and 256 3×3 filters, respectively, connected to three consecutive fully-connected layers with 256, 128 and 128 units, respectively. Finally, a Softmax decision layer with two units gets the output from the last fully-connected layer and generates the result of the classification, identifying between normal and malignant patches. Different architectures were tested in this work, including the VGG16 [29], VGG19 [29], MobileNet [30] and DenseNet121 [31] architectures, although this custom CNN was selected based on the fact that achieved the best results.

2) Training, validating and testing the system

As mentioned in previous sections, CNNs and other deep learning algorithms need a large amount of samples for the training phase. When using these architectures, the dataset is commonly divided into three different sets for training, validating and testing the model, respectively, where the training set is by far the one with more samples.

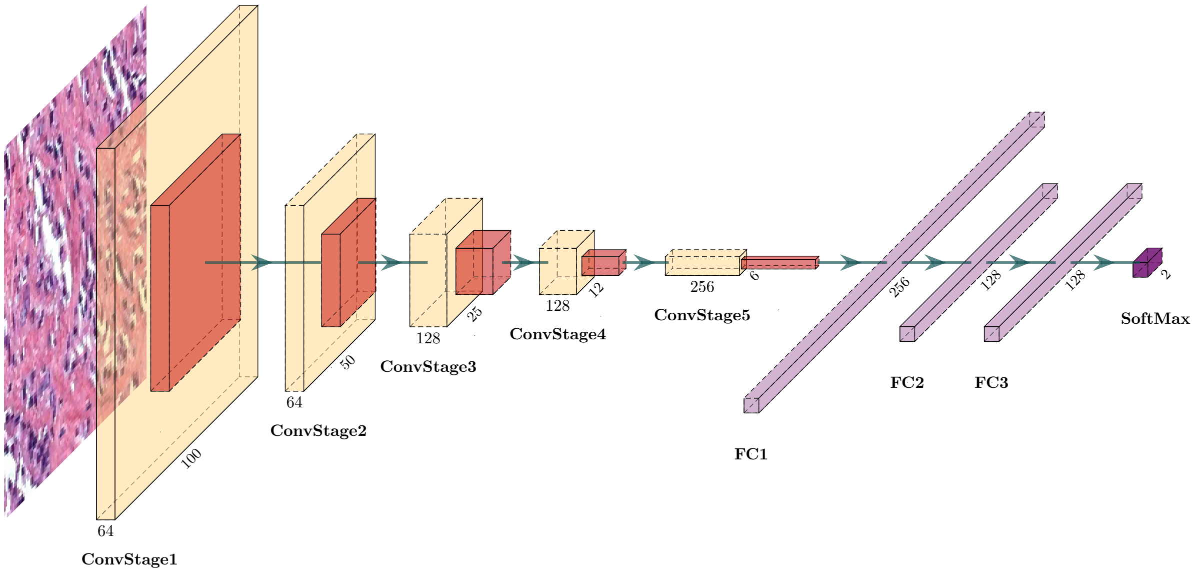
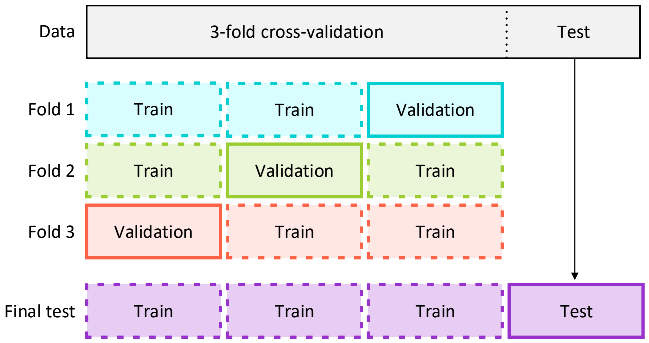


FIGURE 4. Diagram of the architecture of the CNN. Each convolution stage (ConvStageX) consists of convolution, batch normalization, ReLU and 2×2 max pooling layers. Each fully connected stage (FCX) consists of dense, batch normalization, ReLU and dropout (0.5) layers. Convolution kernels are: 5×5, 3×3, 3×3, 3×3, 3×3, respectively.



Data	3-fold cross-validation			Test
Fold 1	Train	Train	Validation	
Fold 2	Train	Validation	Train	
Fold 3	Validation	Train	Train	
Final test	Train	Train	Train	Test

FIGURE 5. 3-fold cross-validation and final test diagram. The dataset was divided into four subsets. Two of them were used for training each fold and one for validation. After that evaluation, those three subsets were used to train a final model and the remaining one was used to test the performance of the system.

At the same time, to measure the generalization ability of the model, cross-validation is usually performed. There are different types of cross-validation; the one we used in this work was the K-fold stratified cross-validation (where $K = 3$). First, the dataset was split in two sets: 75% was used to perform the 3-fold cross-validation and the remaining 25% for performing a final test of the system.

For performing cross-validation, the 3-fold cross-validation set was divided again into three different subsets, where patches in each of these subsets were also divided following a patient-level split. Each subset consisted of, approximately, 50% cancer and 50% normal cases. Then, the network was trained for 200 epochs with a batch size of 32 using Adadelta optimizer [32] and validated three times (once per fold), using two of the subsets for training and the remaining one for validating the system. The results for cancer detection were evaluated as an average of the 3-fold cross-validation results.

After obtaining these results, a final test was performed, using the whole 3-fold cross-validation set (75% of the dataset) for training and then testing with the 25% set that was left apart. Figure 5 shows a diagram about the dataset division for the 3-fold-cross-validation and the final test.

In this work, TensorFlow[3] and Keras[4], which are two widely known Deep Learning frameworks/libraries, were used to design, train and test the network.

3) Evaluation metrics

In order to present the capabilities of this implemented CAD system, different evaluation metrics were used. These are accuracy (Equation 5), precision (Equation 6), sensitivity (Equation 7), specificity (Equation 8), F1-score (Equation 9), and Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) curve. All of them were measured at patch level.

$$\text{Accuracy} = 100 \times \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{5}$$

$$\text{Precision} = 100 \times \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

$$\text{Sensitivity} = 100 \times \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

[3] https://www.tensorflow.org
[4] https://keras.io

$$\text{Specificity} = 100 \times \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{8}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{9}$$

Where TP and FP denote true positive cases (when the system diagnoses a malignant patch correctly) and false positive cases (the system detects a malignant patch in a region where the tissue does not correspond to a tumor), respectively. TN and FN denote true negative cases (the system classifies a normal patch as normal) and false negative cases (the system classifies a malignant patch as normal), respectively.

The ROC curve shows the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The AUC is a commonly used metric that measures the area that is under the ROC curve, where an area of 1 represents a perfect test.

III. RESULTS

A. QUANTITATIVE EVALUATION

The evolution of the loss and accuracy over 200 epochs for each fold, both for the stain-normalized dataset and for the original one that was not normalized, is shown in Figure 6 and Figure 7, respectively.

The ROC curve was calculated for the same cases that were taken into account in the loss and accuracy plots (Figures 6 and 7), along with their corresponding AUC value, which are shown in Figure 8 and Figure 9.

Table 4 presents the results obtained from each of the cross-validation sets that were trained and validated, considering the stain-normalized dataset and the one that was not normalized. These results consists of the evaluation metrics that were introduced in section II-B3, comparing both approaches by calculating the average over the validation sets.

After the cross-validation was performed, and as was explained in section II-B2, the three subsets were used for training and the remaining 25% of the dataset was used to test the network (see Figure 5). With this, the stain-normalized approach achieved 99.14% accuracy, while the not-preprocessed achieved 99.98% (see Table 4). Figures 8 and 9 present the ROC curves for these two tests.

As can be seen from these results, both approaches achieved very high scores in all the metrics that were studied for this classification task, with the dataset that was not normalized performing slightly better (less than 1.5% increase in accuracy). However, as was mentioned in previous sections, these results were obtained with WSIs from the same hospital (Virgen de Valme), and to measure the performance of both approaches with WSIs obtained from different hospitals and scanners, a new test was carried out, which is presented in section III-D.

B. COMPARISON WITH OTHER METHODS

The results obtained in the previous section were compared with different state-of-the-art architectures and classifiers using the same dataset. The following well-known CNN models were used to extract features from the dataset: MobileNet, DenseNet121, VGG16 and VGG19. Instead of training these networks from scratch, whose architectures are more complex than the one that was developed for this work, their weights were obtained by using the transfer learning technique from the ImageNet dataset [33]. Along with these four models, two different classifiers were tested: Support Vector Machine (SVM) and SoftMax. Moreover, each of the architectures was also fine-tuned, meaning that the weights from ImageNet were adjusted using backpropagation to increase the recognition rate over our dataset. The accuracy results for each of the possible combinations are presented in Table 5.

C. EXPERT PATHOLOGISTS' VERIFICATION

In addition to the numerical results that were obtained in the previous section, a validation was also performed by expert pathologists.

To this end, the network trained for the final test was used. With that model, a prediction was performed over the WSIs from the test subset. To perform a prediction, all patches from WSIs were read and only those which passed the patch filters mentioned in section II-A2 were stain normalized and predicted by the CNN. These predictions were represented in a heatmap graph over the original WSI image. An example can be seen in Figure 10, where the ground truth annotations from the pathologist are also shown. These heatmaps were given to different pathologists together with their corresponding WSIs in order to validate the predictions obtained from the network. The results of the CNN presented by the heatmap mark the same regions that pathologists labeled in the original WSI, with the exception of some isolated false positives, which are indicated.

D. TESTING WITH WSIs FROM DIFFERENT HOSPITALS

The results presented in section III-A are promising, although it is important to highlight that, as was mentioned in previous sections, for both training and testing the network, only images from a single hospital were taken into account, which also means images from one laboratory and a specific scanner. A new experiment was carried out in order to measure the performance of the network when using new images obtained from other hospitals. This also allows determining whether the stain-normalization step was better or not compared to the same images without applying any kind of color normalization.

To perform this experiment, new WSIs were obtained from two different hospitals: Puerta del Mar Hospital (Cádiz, Spain) and Clínic Barcelona Hospital (Barcelona, Spain). It is important to mention that this new images were not labeled the same way as the ones that were used to perform the previous experiments. These WSIs were only diagnosed as normal or malignant, without indicating which specific areas of the tissue were relevant for the pathologists to make that decision. Therefore, this new experiment consisted in measuring the number of false positives against true negatives

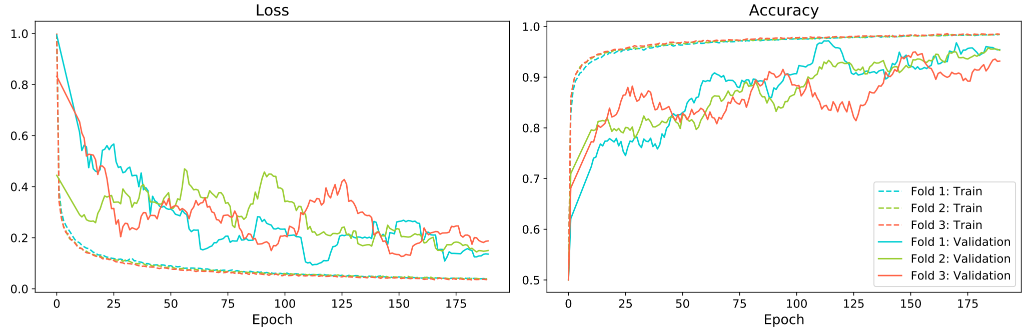


FIGURE 6. Loss and accuracy evolution when training with the three cross-validation sets using the stain-normalized dataset.

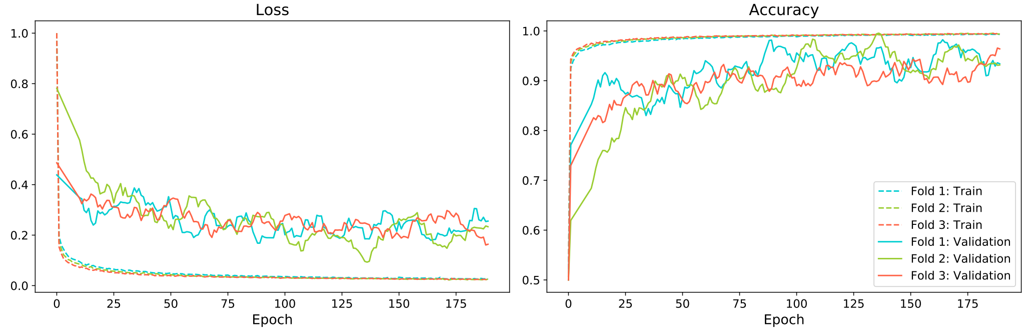


FIGURE 7. Loss and accuracy evolution when training with the three cross-validation sets using the dataset that was not stain-normalized.

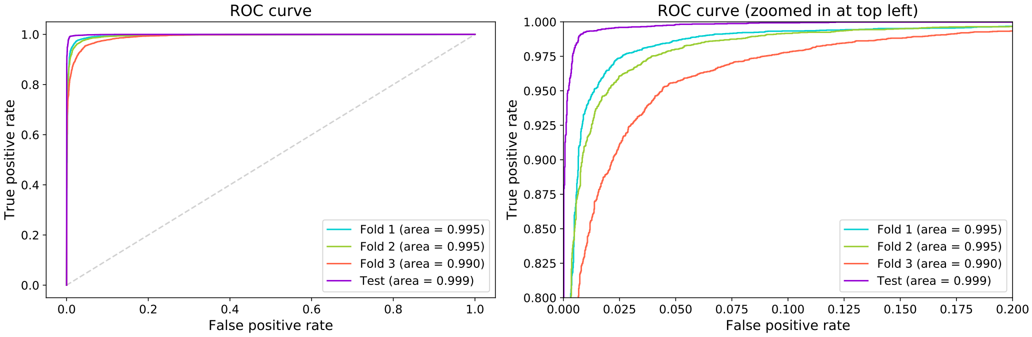


FIGURE 8. Left: ROC curve for each cross-validation set and the test set when using the stain-normalized dataset. Right: zoomed in at top left.

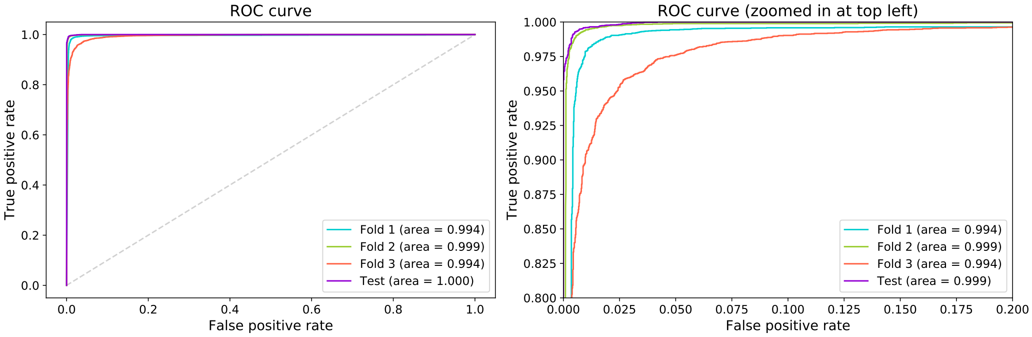


FIGURE 9. Left: ROC curve for each cross-validation set and the test set when using the dataset that was not stain-normalized. Right: zoomed in at top left.

TABLE 4. Results obtained from each cross-validation fold and the final test.

	Set	Dataset	Accuracy (%)	Specificity (%)	Sensitivity (%)	Precision (%)	F1 score (%)	AUC
Cross-validation	1st fold	Stain-normalized	97.43	97.42	97.44	97.47	97.45	0.995
		Not normalized	98.54	98.99	98.11	99.00	98.55	0.994
	2nd fold	Stain-normalized	96.7	97.85	95.63	97.85	96.73	0.995
		Not normalized	99.24	100.00	98.50	100.00	99.24	0.999
	3rd fold	Stain-normalized	95.35	96.25	94.49	96.28	95.37	0.990
		Not normalized	96.43	95.10	97.72	95.34	96.52	0.994
	Average	**Stain-normalized**	**96.49**	**97.17**	**95.83**	**97.2**	**96.51**	**0.993**
		Not normalized	**98.07**	**98.03**	**98.11**	**98.11**	**98.10**	**0.996**
Final test	**Test**	**Stain-normalized**	**99.14**	**99.18**	**99.10**	**99.27**	**99.19**	**0.999**
		Not normalized	**99.98**	**100**	**99.97**	**100**	**99.98**	**0.999**

TABLE 5. Results comparison for different state-of-the-art methods. Best accuracies for each architecture model are highlighted in bold.

Model	Classifier	Fine-tuning	Accuracy (%)
VGG16	SoftMax	No	86.36
		Yes	**94.76**
	SVM	No	85.37
		Yes	93.85
VGG19	SoftMax	No	83.87
		Yes	**93.54**
	SVM	No	85.11
		Yes	91.22
MobileNet	SoftMax	No	81.48
		Yes	98.96
	SVM	No	80.58
		Yes	**99.08**
DenseNet121	SoftMax	No	78.47
		Yes	96.82
	SVM	No	78.00
		Yes	**97.77**

(specificity) detected by the network in total for all WSIs diagnosed as normal for each hospital. WSIs diagnosed as malignant were not taken into consideration for a sensitivity study due to the fact that there was no ground truth that could be used to evaluate the network when testing it with the patches obtained from them. Instead, a statistical study based on Student's t-test is later presented to compare the predicted patches' distribution between normal and malignant WSIs.

From Clínic Barcelona Hospital, 100 new WSIs diagnosed as normal were used, whereas a total of 79 were considered from Puerta del Mar Hospital: 33 of them were obtained from needle core biopsy (the same procedure as Virgen de Valme Hospital and Clínic Hospital) and the remaining 46 WSIs were obtained from incisional biopsy.

Figure 11 shows the mean specificity and standard deviation for each of the three sets from different hospitals, comparing the stain-normalization algorithm (96.08 ± 2.85, 94.82 ± 3.52 and 96.26 ± 2.20, respectively) to the original images (93.31 ± 6.43, 95.87 ± 8.57 and 95.94 ± 3.42, respectively).

Since malignant WSIs only provided a global diagnosis, we could not calculate the sensitivity at patch level. Then, an evaluation relying on the slide-level label was performed, comparing the probability distributions estimated by the CNN for normal and malignant WSIs for each external hospital. To carry out this evaluation, 129 new WSIs diagnosed as malignant from Clínic Barcelona Hospital and 65 new malignant WSIs from Puerta del Mar Hospital (26 obtained from needle core biopsy and 39 from incisional biopsy) were considered, along with the ones diagnosed as normal that were used in the previous experiment. Patches from both normal and malignant WSIs were predicted following the same procedure explained in section III-C with the model that was trained with stain-normalized patches and also with the one that was not (in this case, patches extracted from the WSIs from external hospitals were not stain-normalized in the preprocessing step). The average and standard deviation of the percentage of malignant patches in relation to the total amount of tissue patches (those that passed the patch filters mentioned in section II-A2) that the models predicted, were calculated for each hospital. Statistical Student's t-test was performed to measure how significant the difference between the results obtained for normal and malignant WSIs were. For the t-test, two values were generated: the t-statistic and the critical t-value. If the first is greater than the second, the test concludes that there is a statistically significant difference between the results obtained for normal and malignant WSIs. The results of this evaluation are presented in Table 6, where the impact of using stain-normalization is also shown.

As can be seen from the results obtained when performing the predictions, there is a statistically significant difference

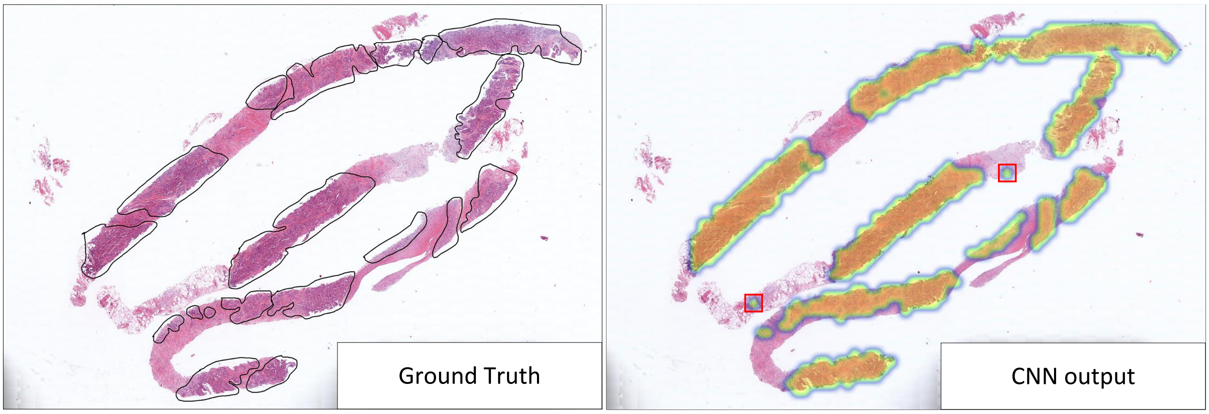


FIGURE 10. Left: WSI taken from the test subset with ground truth labels from pathologists. Right: output of the CNN represented with a heatmap. Isolated false positives marked with red squares.

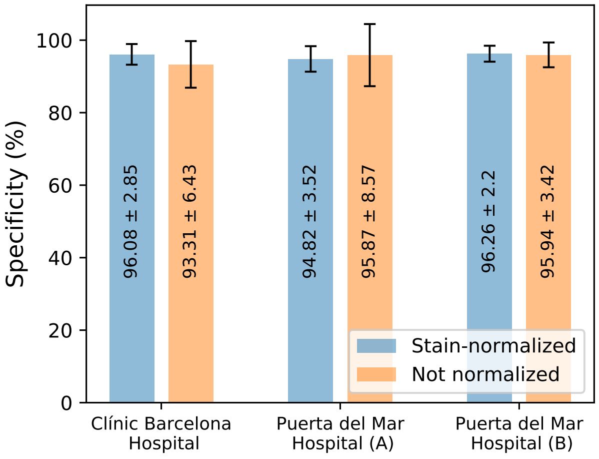


FIGURE 11. Mean specificity and standard deviation achieved by the CNN with WSIs obtained from Clínic Barcelona Hospital (Barcelona. Spain), and Puerta del Mar Hospital (Cádiz, Spain). A and B were extracted with incisional biopsy and needle core biopsy, respectively.

between the results obtained for normal and malignant WSIs when using stain-normalization as part of the preprocessing step. On the other hand, significant differences cannot be achieved when predicting without having applied the normalization process to the input patches before, except for the WSIs obtained from Clínic Barcelona Hospital.

Figure 12 presents three extreme cases from Puerta del Mar Hospital obtained with needle core biopsy. The first case (A) shows a malignant WSI in which the system detected a high quantity of malignant patches (∼45% of the tissue). On the other hand, the second one (B), corresponds to a malignant WSI with around a 6% of the tissue predicted as malignant. Finally, in the third case (C), a normal WSI is shown, in which the system mistakenly detected 5% of the patches that correspond to tissue as malignant. These malignant WSIs present pen marks drawn by the pathologist that globally diagnosed the slide before being scanned, which roughly delimit malignant areas of the tissue. As can be seen in Figure 12, C has a relatively high quantity of patches detected as malignant. However, these patches are scattered across the tissue and, hence, not focusing on a specific region, which clearly represents the error of the system. On the other hand, in B, the small quantity of patches detected as malignant are mostly focused inside the area delimited

TABLE 6. Results of the statistical evaluation performed with malignant and normal WSIs from external hospitals, where Avg%ppm stands for the average of the percentage of patches predicted as malignant, and Std for its standard deviation. Cases where t-static > critical t-value (statistically significant difference found between normal and malignant distributions) are highlighted in bold. A and B were extracted with incisional biopsy and needle core biopsy, respectively.

			Malignant	Normal
Clínic Barcelona	Stain-normalized	Avg%ppm	12.22%	3.92%
		Std	9.29%	2.85%
		t-statistic (critical t-value)	**8.24** **(1.98)**	
	Not normalized	Avg%ppm	15.46%	6.69%
		Std	10.17%	6.43%
		t-statistic (critical t-value)	**7.29** **(1.98)**	
Puerta del Mar (A)	Stain-normalized	Avg%ppm	11.47%	5.18%
		Std	9.43%	3.52%
		t-statistic (critical t-value)	**3.93** **(2.01)**	
	Not normalized	Avg%ppm	7.35%	4.13%
		Std	9.97%	8.57%
		t-statistic (critical t-value)	1.58 (1.99)	
Puerta del Mar (B)	Stain-normalized	Avg%ppm	14.00%	3.74%
		Std	11.87%	2.20%
		t-statistic (critical t-value)	**4.35** **(2.05)**	
	Not normalized	Avg%ppm	3.35%	4.06%
		Std	4.08%	3.42%
		t-statistic (critical t-value)	-0.71 (2.01)	

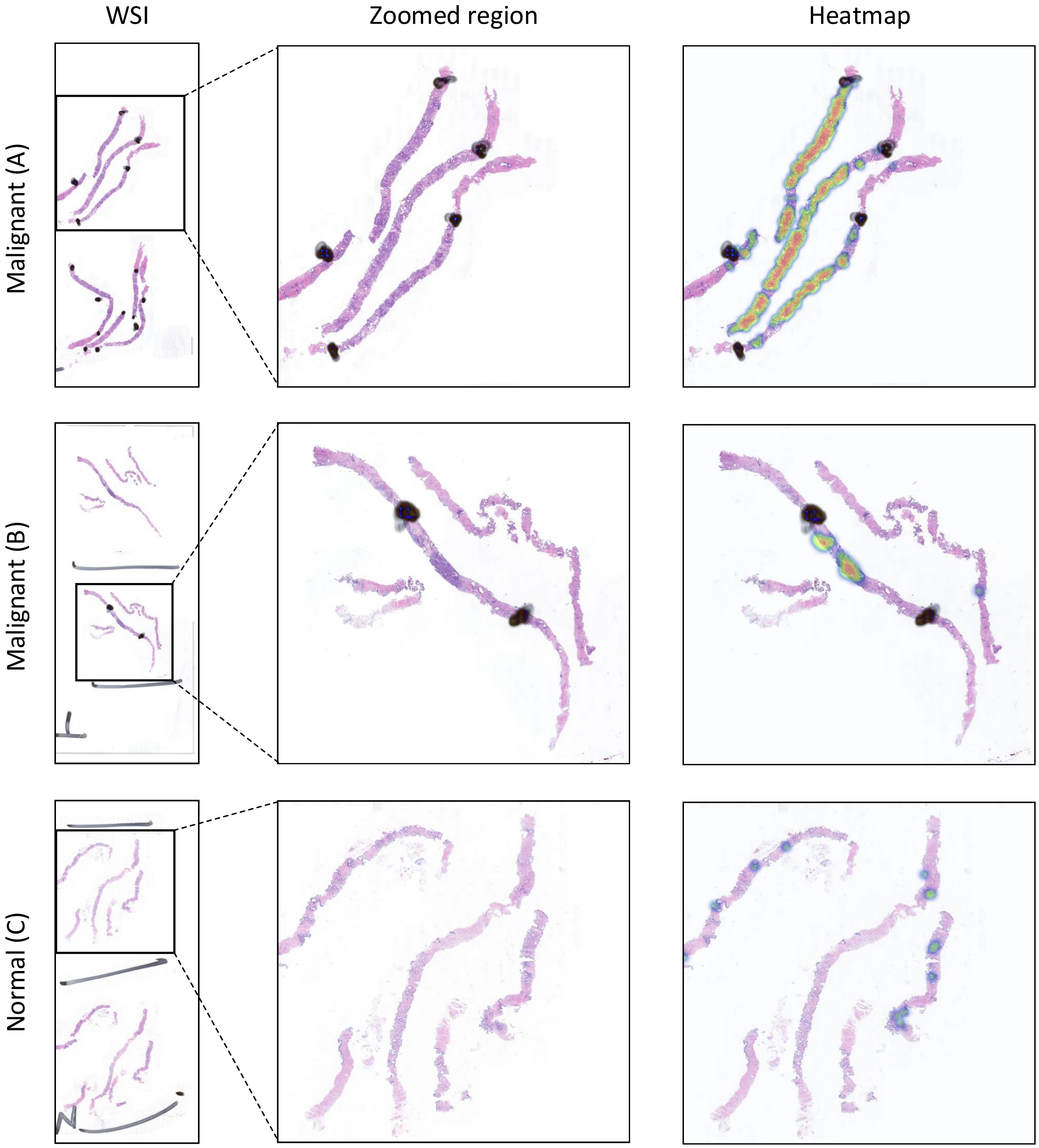


FIGURE 12. Heatmaps generated by the system for three different WSIs from Puerta del Mar Hospital. A and B correspond to WSIs globally diagnosed as malignant with high and low quantity of malignant patches detected by the system, respectively, while C represents a normal WSI with a high error rate in the prediction. Zoomed regions are presented for better visualization.

by the pen marks. After being revised by a pathologist, it was confirmed that the malignant area matches the heatmap, while the rest corresponds to normal tissue, except for a small area that is partially overlapped by the bottom pen mark. Finally, the heatmap presented for A shows that the system detects most of the malignant tissue correctly based on the pen marks.

Further details on the limitations of the study presented in this section are discussed in section V.

E. PERFORMANCE EVALUATION

As mentioned in section II, the fatigue that the pathologist undergoes when examining many consecutive WSIs was the main motivation and inspiration of this work. With this CAD system, pathologists could be aided in the process of looking for malignant areas in such large images, speeding up the process and having a second opinion. For this reason, the time that the system takes to evaluate a WSI should be as short as possible. To this end, a performance evaluation was carried out to determine the advantages of this CAD system.

This evaluation was performed over the 97 WSIs from Virgen de Valme Hospital, measuring the time that each of the images takes to go through the whole process: from the extraction of all patches, to the prediction of each patch, including filtering and stain-normalization. These times were obtained with an Intel® Core™ i7-8700K CPU at 3.70 GHz.

Figure IV top shows the mean time (in milliseconds) that the system takes to process one patch (8.9 ms in total) for each of the processing steps. Figure IV bottom shows the mean time (in seconds) that is needed for the CAD system to process a complete WSI on average (calculated over 97 WSIs) (18.78±6.55 s in total for a stain-normalized WSI and 16.41±5.39 s for a WSI that has not been normalized), detailing how much time it requires for each of the processing steps. As an example, the WSI shown in Figure 10 took 12.9 s to complete the whole process.

As can be seen, for a single patch, the extraction is the process that takes less time (around 0.6 ms). However, when processing a WSI, the patch extraction is the step that involves the longest amount of time. This is due to the fact that, in that step, all patches from a WSI have to be read and analyzed, but not all of them are processed in the following steps. Most of the patches are discarded before being scored to remove unwanted areas. Then, only those which are not background and pass the scoring step are normalized and predicted by the CNN.

To compare the network that was developed with the different models that were tested in section III-B, the average time that each of these took to process a single patch was measured over the same WSIs. Compared with the 2.8 ms that our model took to predict a patch, MobileNet took 4.88 ms; VGG16, 5.85 ms; VGG19, 6.14 ms; and DenseNet121, 13.8 ms, in the best cases, due to their greater number of layers and higher complexity.

IV. DISCUSSION

As can be clearly seen from the results obtained in section II-B3, both the stain-normalized version of the dataset and the original one achieved very high recognition rates for the prostate cancer detection task. However, the second one performed slightly better than the first in the 3-fold cross-validation step and in the final test, which can be due to different reasons. First of all, color differences could be one of the factors that the network learns for distinguishing between malignant and normal patches, since the H&E stain makes malignant regions tend to a more purple-like color due to the stronger hematoxylin stain balance because of the higher nucleic acid content in those areas. Then, normalizing all patches to a target color could imply losing relevant information for the classification.

However, when testing with different hospitals whose WSIs were not used to train the system and which present different color variations, the results changed when comparing both approaches. In that case, the mean specificity of both approaches is still around 95%. However, when looking at the standard deviation of the specificity, the difference is clearer: the stain-normalized was more stable while achieving almost the same result. This could be caused by the fact that, thanks to the normalization, the patches were more homogeneous in terms of color, and the network was able to extract more relevant features based on the cell structures (which are more complex to detect than color differences) during the training phase. Hence, the stain-normalization could make the system more robust and stable to images from new hospitals and scanners where color variations exist. This idea was confirmed when performing the Student's t-test over the percentage of malignant area of the tissue predicted by the CNN for normal and malignant WSIs without applying stain-normalization, which showed that there was no statistically significant difference between the two classes for two out of the three external sources. These results were also studied in [34], where the authors conclude that training a deep CNN with stain-normalized images did not improve the results and, in some cases, they were worse than the baseline. However, the authors state that this technique improved the generalization of the CNN for classification tasks using digital pathology images. Our results also confirm this idea regarding the application of stain normalization to prostate cancer histopathological images.

Apart from the preprocessing step, another option for improving the results obtained from the CAD system could be adding a postprocessing layer to remove either isolated false negatives and/or isolated false positives, as the ones that were present in the heatmaps in Figure 10 (marked with red squares) and Figure 12 (C).

State-of-the-art works, such as [10], [12], [16] and [17], also performed a classification between normal and malignant tissue. However, since these works performed the classification and obtained the metrics at a different level (core-level and slide-level), results cannot be directly compared. Moreover, those works use different datasets, which also does not allow a strict and fair comparison with the results obtained in this work. Therefore, our network was compared with different well-known pre-trained models, which were tested on the same dataset that we used. Some of them obtained similar results after a fine-tuning process, as presented in Table 5. However, in terms of performance, due to the higher complexity of these models compared to our proposal, the average time that they take to predict a single patch is higher than that of our network, as presented in section III-E. This means that, when using our network, the CAD system would be able to process more WSIs in the same amount of time than any of the other models that were tested, as well as to achieve a slightly higher accuracy. These pre-trained models were much faster to train than our network, since it was trained from scratch, and their accuracies are close to ours. However, since the training process is a step that only has to be done once, it is worth to have a longer training process in order to obtain a lighter computational algorithm with better accuracy for its production phase (predicting every new WSI

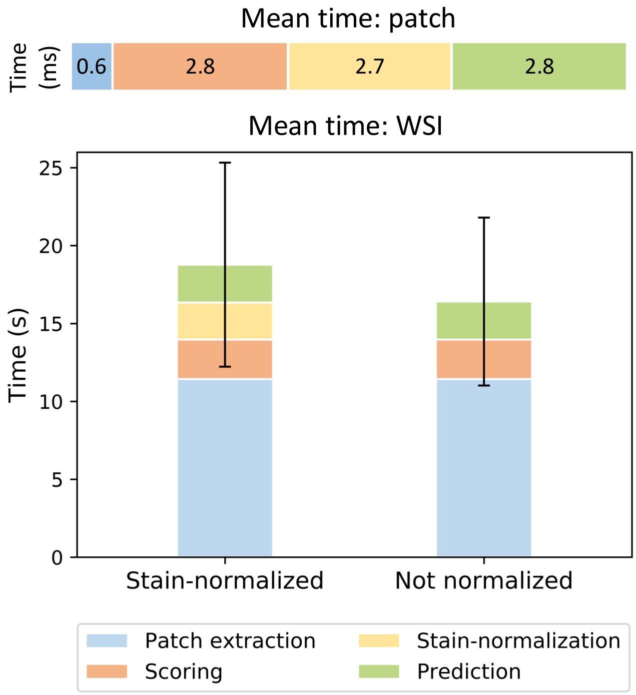


FIGURE 13. Mean execution time for patches (top) and WSIs (bottom), calculated over 97 WSIs, detailing between patch extraction, scoring, stain-normalization and prediction.

that is processed in a hospital). We have also carried out a comprehensive comparison with other state-of-the-art works in terms of the number of operations (OPS) performed by the network. In our case, based on the number and size of the layers, our network performs around 350 MOPS (10^6 operations) per patch. Other works that present high accuracies for a binary classification task in WSIs, such as [11], [16], [17], use well-known models that, based on [35], perform more than 1 GOPS (10^9 operations) per input patch. The custom model presented in [12] needs to perform more than 660 MOPS per patch. Our custom CNN outperforms other state-of-the-art works in terms of computational complexity for a binary classification task in prostate histopathological images between normal and malignant WSIs.

As a future work, we would like to develop a more complex CAD system based on the one proposed on this work. This system would consists of two different CNNs: the first one would discriminate between normal and malignant patches, and the second one would receive the patches classified as malignant and identify the GGS score that they belong to. We are already working with pathologists in order to get new WSIs labeled by them in order to train the first CNN with more images and also to start developing the second one for building the new CAD system.

V. LIMITATIONS OF THE STUDY

In this study, we present and evaluate a tool, called PROMETEO, that aims to aid pathologists in their routine work. PROMETEO is a CNN-based system able to analyze a WSI for patch-level classification of normal and malignant tissue. It offers a heatmap over the original WSI to highlight the malignant tissue detected together with statistical values about the distribution of the patches. The current state of the system presents some limitations which are discussed in the following subsections.

1) Dataset size

The size of the dataset (obtained from ∼100 WSIs of one particular hospital) has demonstrated that robust results for cancer detection at patch level can be achieved after training and testing our presented model. Nevertheless, the invaluable work from pathologists in labeling these WSIs to extract patches from the presented dataset must continue in order to obtain a competitive dataset valid for deeper classification with Gleason scores, and to combine information from different hospitals to improve results on classifying WSIs from different sources not used in the training process.

2) Sensitivity on external hospitals

In order to calculate the sensitivity at patch level when predicting WSIs from external hospitals and scanners, region-specific labels from pathologists are required. However, these were not available for external hospitals. This limitation made us perform a different evaluation relying on their global available label by comparing the probability distributions of malignant patches predicted by the system for normal and malignant WSIs. To this end, the statistical t-test was performed in order to discriminate if predictions for normal and malignant WSIs were significantly different. This test has demonstrated that, even though there is a limitation on reporting the results from the external hospitals, the distributions of predicted patches on malignant and normal cases are clearly different when using the stain-normalization in the process, which validates the application of this method.

VI. CONCLUSION

In this work, the authors have presented a novel CAD system based on deep learning algorithms (CNNs) for discriminating between malignant and normal regions in WSI images obtained from Virgen de Valme Hospital (Seville, Spain) and labeled by expert pathologists. A custom CNN, called PROMETEO, consisting of 9 layers (5 convolution stages, 3 fully connected layers and a Softmax layer) was trained, validated and tested by means of a 3-fold stratified cross-validation technique with 100×100 patches extracted from the WSIs. These patches were filtered with a novel patch extraction and scoring algorithm which removed unwanted areas such as pen marks and external agents.

The authors have also studied the impact of using a stain-normalization algorithm on the patches for improving the classification of the system. The results show that the application of this kind of normalization is not relevant when working with images obtained from the same hospital/scanner, although it could potentially be useful for developing a stable and "universal" CAD system which could achieve better results (than the one trained with images that are not stain-normalized) due to the color variations that WSIs from different scanners present because of the H&E stain process.

The network presented in this work achieves, in the best case of the cross-validation process, a mean accuracy of 98.07%, a mean specificity of 98.03%, a mean sensitivity of 98.11%, a mean precision of 98.11%, a mean F1-score of 98.10% and a mean AUC of 0.996 over the three cross-validation folds. The system was also evaluated over a test set, obtaining 99.98% accuracy, 100% specificity, 99.97% sensitivity, 100% precision, 99.98% F1-score and 0.999 AUC in the best case. Different state-of-the-art methods were tested with the same dataset to compare the performance of the system. In the case of MobileNet with SVM as classifier, the accuracy achieved is similar to that of our model, with a difference of 0.9%. However, if the execution time needed to process a WSI is considered, our model is 75% faster, approximately, which is very relevant when working with real-time CAD systems.

The system is able to generate a heatmap of the input WSI, indicating the regions that the network has detected as malignant. This could help pathologists in their task by reducing fatigue and the time they take to analyze a WSI.

REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a Cancer Journal for Clinicians,* vol. 68, no. 6, pp. 394–424, 2018.

[2] K. Garber, "A tale of two cells: discovering the origin of prostate cancer," *JNCI: Journal of the National Cancer Institute,* vol. 102, no. 20, pp. 1528–1535, 2010. DOI: 10.1093/jnci/djq425.

[3] K. D. Berg, B. G. Toft, M. A. Røder, K. Brasso, B. Vainer, and P. Iversen, "Prostate needle biopsies: interobserver variation and clinical consequences of histopathological re-evaluation," *Apmis,* vol. 119, no. 4-5, pp. 239–246, 2011.

[4] A. Matoso and J. I. Epstein, "Grading of prostate cancer: past, present, and future," *Current Urology Reports,* vol. 17, no. 3, pp. 25, 2016.

[5] A. M. Lessells, R. A. Burnett, S. R. Howatson, S. Lang, F. D. Lee, K. M. McLaren, E. R. Nairn, S. A. Ogston, A. J. Robertson, J. G. Simpson, et al., "Observer variability in the histopathological reporting of needle biopsy specimens of the prostate," *Human Pathology,* vol. 28, no. 6, pp. 646–649, 1997.

[6] M. McLean, J. Srigley, D. Banerjee, P. Warde, and Y. Hao, "Interobserver variation in prostate cancer Gleason scoring: are there implications for the design of clinical trials and treatment strategies?," *Clinical Oncology,* vol. 9, no. 4, pp. 222–225, 1997.

[7] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights Into Imaging,* vol. 9, no. 4, pp. 611–629, 2018.

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

[9] M. Veta, J. P. Pluim, P. J. Van Diest, and M. A. Viergever, "Breast cancer histopathology image analysis: A review," *IEEE Transactions on Biomedical Engineering,* vol. 61, no. 5, pp. 1400–1411, 2014.

[10] J. T. Kwak and S. M. Hewitt, "Nuclear architecture analysis of prostate cancer via convolutional neural networks," *IEEE Access,* vol. 5, pp. 18526–18533, 2017.

[11] O. J. del Toro, M. Atzori, S. Otálora, M. Andersson, K. Eurén, M. Hedlund, P. Rönnquist, and H. Müller, "Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade Gleason score," in *Proc. SPIE Medical Imaging 2017: Digital Pathology,* vol. 10140, 2017, Art. no. 101400O.

[12] G. Litjens, C. I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen-Van De Kaa, P. Bult, B. Van Ginneken, and J. Van Der Laak, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Scientific Reports,* vol. 6, Art. no. 26286, 2016.

[13] W. Li, J. Li, K. V. Sarma, K. C. Ho, S. Shen, B. S. Knudsen, A. Gertych, and C. W. Arnold, "Path R-CNN for prostate cancer diagnosis and Gleason grading of histological images," *IEEE Transactions on Medical Imaging,* vol. 38, no. 4, pp. 945–954, 2018.

[14] S. Doyle, M. Hwang, K. Shah, A. Madabhushi, M. Feldman, and J. Tomaszeweski, "Automated grading of prostate cancer using architectural and textural image features," in *Proc. IEEE ISBI 2017: From Nano to Macro,* 2017, pp. 1284–1287.

[15] J. Ren, E. Sadimin, D. J. Foran, and X. Qi, "Computer aided analysis of prostate histopathology images to support a refined Gleason grading system," in *Proc. SPIE Medical Imaging 2017: Image Processing,* vol. 10133, 2017, Art. no. 101331V.

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

[17] P. Ström, K. Kartasalo, H. Olsson, L. Solorzano, B. Delahunt, D. M. Berney, D. G. Bostwick, A. J. Evans, D. J. Grignon, P. A. Humphrey, et al., "Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study," *The Lancet Oncology,* vol. 21, no. 2, pp. 222–232, 2020.

[18] W. Bulten, H. Pinckaers, H. van Boven, R. Vink, T. de Bel, B. van Ginneken, J. van der Laak, C. Hulsbergen-van de Kaa, and G. Litjens, "Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study," *The Lancet Oncology,* vol. 21, no. 2, pp. 233–241, 2020.

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

[20] C. Yan, K. Nakane, X. Wang, Y. Fu, H. Lu, X. Fan, M. D. Feldman, A. Madabhushi, and J. Xu, "Automated Gleason grading on prostate biopsy slides by statistical representations of homology profile," *Computer Methods and Programs in Biomedicine,* vol. 194, Art. no. 105528, 2020.

[21] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 9, no. 1, pp. 62–66, 1979.

[22] F. Ciompi, O. Geessink, B. E. Bejnordi, G. S. De Souza, A. Baidoshvili, G. Litjens, B. Van Ginneken, I. Nagtegaal, and J. Van Der Laak, "The importance of stain normalization in colorectal tissue classification with convolutional networks," in *Proc. IEEE ISBI 2017,* pp. 160–163, 2017.

[23] S. Vesal, N. Ravikumar, A. Davari, S. Ellmann, and A. Maier, "Classification of breast cancer histology images using transfer learning," in *Proc. ICIAR 2018,* 2018, pp. 812–819.

[24] S. Roy, A. kumar Jain, S. Lal, and J. Kini, "A study about color normalization methods for histopathology images," *Micron,* vol. 114, pp. 42–61, 2018.

[25] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer Graphics and Applications,* vol. 21, no. 5, pp. 34–41, 2001.

[26] D. Magee, D. Treanor, D. Crellin, M. Shires, K. Smith, K. Mohee, and P. Quirke, "Colour normalisation in digital histopathology images," in *Proc. Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy (MICCAI Workshop),* vol. 100, 2009, pp. 100–111.

[27] D. L. Ruderman, T. W. Cronin, and C.-C. Chiao, "Statistics of cone responses to natural images: implications for visual coding," *Journal of the Optical Society of America A,* vol. 15, no. 8, pp. 2036–2045, 1998.

[28] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proc. IEEE,* vol. 86, no. 11, pp. 2278–2324, 1998.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556,* 2014.

[30] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861,* 2017.

[31] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. CVPR,* 2017, pp. 4700–4708.

[32] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701,* 2012.

[33] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR,* 2009, pp. 248-255.

[34] S. Otálora, M. Atzori, V. Andrearczyk, A. Khan, and H. Müller, "Staining invariant features for improving generalization of deep convolutional neural networks in computational pathology," *Frontiers in Bioengineering and Biotechnology,* vol. 7, pp. 198, 2019.

[35] A. Canziani, A. Paszke, and E. Culurciello, "An analysis of deep neural network models for practical applications," *arXiv preprint arXiv:1605.07678,* 2016.

L. DURAN-LOPEZ received the B.S. degree in Biomedical Engineering in 2016 and the M.S. degree in Biomedical Research in 2017, both from the University of Seville (Sevilla, Spain). On September 2017, she started her Ph.D. in the department of Computer Architecture and Technology, at the University of Seville.

Since 2017, she has worked as a Research Fellow in the Robotics and Technology of Computers Lab. from the University of Seville. Her research interests include image processing, medical image analysis, computer-aided diagnosis systems and deep learning, particularly, convolutional neural networks.

J. P. DOMINGUEZ-MORALES received the B.S. degree in computer engineering, the M.S. degree in computer engineering and networks, and the Ph.D. degree in computer engineering (specializing in neuromorphic audio processing and spiking neural networks) from the University of Seville (Sevilla, Spain), in 2014, 2015 and 2018, respectively.

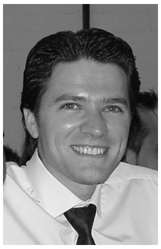
From October 2015 to December 2018, he was a PhD student in the Architecture and Technology of Computers Department of the University of Seville with a research grant from the Spanish Ministry of Education and Science. Since January 2019, he has been working as Assistant Professor in the same department. His research interests include medical image analysis, convolutional neural networks, computer-aided diagnosis systems, neuromorphic engineering, spiking neural networks, neuromorphic sensors and audio processing. In 2016 he became a member of the European Neural Network Society, and he has been a member of IEEE for four years.

A. F. CONDE-MARTIN received the B.S. degree in Medicine (specializing in anatomic pathology) from the University of Seville (Sevilla, Spain), in 1992.

From 1993 to 1997, he was a residence in the Pathology unit at the Department of Pathology in Virgen del Rocío Hospital (Sevilla, Spain). Since 1998, he works as general pathologist. He is part of the Pathology unit at Virgen de Valme Hospital (Sevilla, Spain). Since 2008, he is member of the editorial committee of Revista Española de Patología. His research interests include digital pathology, uropathology, gastrointestinal pathology and hematopathology.

S. VICENTE-DIAZ received his B.S. degree in Computer Science and his Ph.D. from the University of Seville (Sevilla, Spain), in 1996 and 2001, respectively.

Since 2010 is Associate Professor at the same University. Currently, he is ViceDean of the Computer Engineering School (2010-2013). He has been a researcher for the Robotics Technology of Computers Lab. since 1996. He is author/co-author of more than 40 papers in refereed international journals and conferences in the fields of robotics, accessibility, e-health, embedded systems and bioinspired systems. He has participated in more than 20 research projects and contracts. He has participated in EU projects FLEX, CAVIAR and CARDIAC. He is co-funder of the Spin-Off COBER, mainly devoted to biomedical robotics.



A. LINARES-BARRANCO (M'06) received the B.S. degree in computer engineering, the M.S. degree in industrial computer science, and the Ph.D. degree in computer science (specializing in computer interfaces for neuromorphic systems) from the University of Sevilla, Sevilla, Spain, in 1998, 2002, and 2003, respectively.

After working in some companies (ABENGOA, IMSE, Air-force), he started as an Assistant Professor at the Architecture and Technology of Computers Department of the University of Sevilla in 2001. In 2009, he was promoted to Associate Professor (civil-servant). He worked as the secretary of the department from 2013-2017. Since 2017 he is head of department. In 2014 was visiting professor with the UZH-ETHZ at the Institute of Neuroinformatics. He has visited Bielefeld University (CITEC) in 2018 and Ulster University in 2020 with Salvador de Madariaga funds. His research interests include VLSI for FPGA digital design, neuromorphic engineering for interfaces, sensor's processing, motor control and deep-learning. He is co-funder of the Spin-Off COBER, mainly devoted to biomedical robotics.