Computers in Biology and Medicine 136 (2021) 104743

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Wide & Deep neural network model for patch aggregation in CNN-based prostate cancer detection systems

L. Duran-Lopez [a,b,c,d,*], Juan P. Dominguez-Morales [a,b,c,d], D. Gutierrez-Galan [a,b,c,d], A. Rios-Navarro [a,b,c,d], A. Jimenez-Fernandez [a,b,c,d], S. Vicente-Diaz [a,b,c,d], A. Linares-Barranco [a,b,c,d]

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

ARTICLE INFO

Keywords:
Prostate cancer
Deep learning
Convolutional neural networks
Computer-aided diagnosis
Patch aggregation
Whole-slide images
Medical image analysis

ABSTRACT

Prostate cancer (PCa) is one of the most commonly diagnosed cancer and one of the leading causes of death among men, with almost 1.41 million new cases and around 375,000 deaths in 2020. Artificial Intelligence algorithms have had a huge impact on medical image analysis, including digital histopathology, where Convolutional Neural Networks (CNNs) are used to provide a fast and accurate diagnosis, supporting experts in this task. To perform an automatic diagnosis, prostate tissue samples are first digitized into gigapixel-resolution whole-slide images. Due to the size of these images, neural networks cannot use them as input and, therefore, small subimages called patches are extracted and predicted, obtaining a patch-level classification. In this work, a novel patch aggregation method based on a custom Wide & Deep neural network model is presented, which performs a slide-level classification using the patch-level classes obtained from a CNN. The malignant tissue ratio, a 10-bin malignant probability histogram, the least squares regression line of the histogram, and the number of malignant connected components are used by the proposed model to perform the classification. An accuracy of 94.24% and a sensitivity of 98.87% were achieved, proving that the proposed system could aid pathologists by speeding up the screening process and, thus, contribute to the fight against PCa.

1. Introduction

According to GLOBOCAN, prostate cancer (PCa) is the second most frequently diagnosed cancer and the fifth leading cause of cancer death in men, with more than 1.41 million cases in 2020 and around 375,000 deaths worldwide [1]. It is estimated that PCa cases will increase with around 1,000,000 new cases in 2040, according to the World Health Organization (WHO) [2].

Generally, the first step to diagnose PCa consists in a Digital Rectal Exam (DRE), which is the primary test for the initial clinical assessment of the prostate. If an abnormal result for DRE is found, a Prostate-Specific Antigen (PSA) analysis is performed as a screening method for the investigation of a tumor. Then, in case of a positive PSA, a *trans*-rectal ultrasound-guided biopsy is considered, which is the most certain test to confirm or exclude the presence of cancer [3]. With this technique, prostate samples are obtained, which are processed in a laboratory and scanned, producing gigapixel-resolution images called Whole-Slide Images (WSIs). These images are analyzed by pathologists to provide a final diagnosis with the corresponding cancer treatment.

The use of Artificial Intelligence (AI) in image analysis has had a huge impact in recent years [4,5], mainly due to the computational advances and the accessibility of its algorithms for researchers. Its application in the biomedical field has expanded considerably, particularly the use of Deep Learning (DL), which has become one of the most popular AI techniques for image recognition in the last years [6]. These algorithms could play an important role as screening methods to report a second opinion and assist doctors in specific image analysis tasks [7,8]. Particularly, this approach has recently been widely used in digital histopathology, where Convolutional Neural Networks (CNNs) and other different DL mechanisms are trained to analyze and detect

* Corresponding author. Robotics and Tech. of Computers Lab., Universidad de Sevilla, 41012, Seville, Spain.
E-mail address: lduran@atc.us.es (L. Duran-Lopez).

https://doi.org/10.1016/j.compbiomed.2021.104743
Received 4 April 2021; Received in revised form 3 August 2021; Accepted 3 August 2021
Available online 14 August 2021

malignant tissue in WSIs. Since CNNs cannot use an entire WSI as input due to their high resolution, which would require a huge memory and processing capacity, a common approach to this problem is to extract smaller subimages from them, called patches. Therefore, the CNN is able to analyze the WSIs at patch level and then report the classification results obtained.

Previous works, such as [9–13], have followed this patch-level classification strategy in order to develop DL-based Computer-Aided Diagnosis (CAD) systems for PCa detection in digitized histopathological images, reporting accurate results with different metrics and datasets. Among them, to the best of the authors' knowledge, PROMETEO [14] achieved the fastest and least complex model [15] while also obtaining state-of-the-art results, leading to the most-plausible edge-computing solution for PCa detection. This was achieved by means of a 9-layer custom CNN trained and validated with a set of patches after applying different processing steps, including patch filtering, stain normalization and data augmentation. This allowed achieving 99.98% accuracy, 99.98% F1 score and 0.999 Area Under Curve (AUC) on a separate test set.

Since the results obtained when using CNNs are reported at patch level, different techniques have been proposed in the literature in order to combine them and generate a slide-level classification result, which could be of great importance for developing a fast PCa screening system. This technique is known as patch aggregation. Among the different studies that can be found in the literature, some performed different patch aggregation techniques based on Recurrent Neural Networks (RNNs) [10], Random Forests (RFs) [10] and other Machine Learning (ML) or statistical alternatives [9,13], achieving accurate solutions and leading to precise screening methods that could help pathologists in their task.

This work presents a custom novel Wide & Deep (W&D) model for aggregating the patch-level classification results obtained from the PROMETEO CAD system into a global slide-level class. This approach allows providing a fast screening method for PCa detection at WSI level, while also benefiting from the spatial resolution obtained at patch level. The promising results obtained, which have also been compared to other state-of-the-art ML-based approaches, show that the proposed solution could aid pathologists when analyzing histopathological images, discriminating between positive and negative PCa samples while speeding up the whole process.

The main contributions of this work include the following:

- A set of algorithms to automatically extract relevant features from the output of a patch-level DL-based PCa detection system.
- A 5-layer custom W&D model, trained and validated from scratch, which extracts independent features from the inputs and combines them to achieve a slide-level PCa screening method with high sensitivity.
- A comparative study between different widely-known ML algorithms for the patch-aggregation task on the same dataset, which shows that the proposed method achieves the highest sensitivity.

The rest of the paper is structured as follows: section 2 presents state-of-the-art works that are related to the goal of this paper. In section 3, the materials and methods are presented, focusing on the dataset that was used for this work (3.1), along with the neural network model (3.2), the details on the training and validation steps of the model with the aforementioned dataset (3.3), and the test methods (3.4) and statistical anaylisis (3.5). Then, section 4 presents the results obtained with the proposed model using different evaluation metrics. A comparison with other state-of-the-art ML techniques is also performed in the same section. In section 5, the results are discussed and some future works are presented. Finally, the conclusions of this work are presented in section 6.

2. Related work

Previous studies have also focused on cancer detection in histopathological images using different techniques to aid and support pathologists in their decision and also to serve as automatic screening alternatives, giving experts the possibility to prioritize higher cancer risk cases. Among them [12], presented a Region-based CNN model that was trained to classify patches extracted from WSIs as either stroma, benign, low Gleason Grading System (GGS) grade or high GGS grade. The authors achieved 88.78% standard mean accuracy. In Ref. [14], the authors designed, trained and evaluated a custom shallow CNN model for detecting between normal and malignant patches extracted from WSIs with the lowest possible latency. With this method, the authors achieved 99.98% accuracy and 0.999 AUC in the external test set.

The aforementioned studies report patch-wise metrics and results, with which useful information could be provided, such as the location and size of the malignant areas within the tissue. In these works, the authors combine the results obtained at patch-level in order to represent the output of the system, which is a heatmap highlighting the cancerous regions detected. However, for building an intelligent system for PCa screening purposes, a slide-wise result could summarize whether the sample is malignant or normal using a single label and, therefore, provide a faster and more convenient solution for this task.

In this regard, other authors have proposed different strategies for patch-aggregating features obtained from the classification into a single slide-wise label. Ström et al. [9] used boosted trees, which were trained to report the ISUP score and the cancer length from aggregated features obtained from the patch-wise probabilities predicted by 2 ensembles of 30 InceptionV3. The authors achieved 0.62 mean pairwise kappa and 0.986 AUC on the validation set when classifying between malignant and benign samples. Bulten et al. [13] proposed the use of a UNet architecture to perform the patch-level classification in histopathological images. In this case, the authors considered a simpler solution for the patch-aggregation task, which consisted of classifying a sample as malignant if at least 10% of the tissue was predicted as cancer by the system. This allowed achieving 0.990 AUC when classifying between malignant and benign biopsies. Litjens et al. [11] presented a custom CNN for patch-wise cancer detection in WSIs. Whole-slide likelihood images, which take between 5 and 10 min to generate, are obtained as output, and features extracted from them are calculated to report a slide-level label. A percentile analysis is used to determine if the corresponding WSI is malignant or not, achieving 0.99 AUC. The authors in Ref. [10] proposed different patch aggregation techniques, including max pooling, RF and RNNs. With max pooling, the slide is considered positive if a patch is predicted as positive, which resulted on a non-robust solution, since a single spurious missclasification can change the slide prediction. The RF alternative achieved 0.98 AUC, although low sensitivity was obtained. A more complex alternative using Multiple Instance Learning and RNNs was proposed, obtaining 0.991 AUC.

3. Materials and methods

3.1. Dataset

A set of Hematoxylin and Eosin (H&E)-stained slides were used (158 normal WSIs and 174 malignant WSIs), provided by the Pathological Anatomy Unit of Virgen de Valme Hospital (Seville, Spain). These images were preprocessed using different steps. First, small subimages, called patches (100 × 100 pixels at 10× magnification), were extracted from them. Next, background patches and patches corresponding to unwanted areas were discarded with a filter that discriminates them based on the amount of tissue that they contain, the percentage of pixels that are within H&E's hue range, and the dispersion of the saturation and brightness channels. Then, a color normalization process called Reinhard stain-normalization [16,17] was applied to patches in order to reduce stain variability between samples. Finally, color-normalized

patches were used as input to a CNN, called PROMETEO, which classifies them as either malignant or normal tissue with a certain probability. A deeper insight on these steps is given in Ref. [14] and can be seen in Fig. 1.

Different features were obtained from PROMETEO's output in order to create the dataset. The first feature considered to discriminate between malignant or normal WSI was the percentage of malignant tissue area, also called malignant tissue ratio (MTR), expressed between 0 and 1. This was calculated by dividing the number of patches classified as malignant by the total amount of tissue patches extracted from the WSI. This is the most representative data to perform a slide-level classification, since the more malignant patches the network detects on the WSI, the greater its likelihood being malignant. However, based on the error of the CNN when performing the patch classification, the percentage of malignant tissue of the WSI should not be the only input to be considered for the patch aggregation task, since there are some exceptions that do not meet the aforementioned rule (e.g., a malignant WSI with a small tumor in a specific region or a normal WSI with a relatively high percentage of incorrectly-classified malignant tissue area).

Therefore, another feature taken into account to distinguish between malignant and normal WSIs was the distribution of the prediction probability for malignant patches. When the CNN predicts a patch, it reports the probability of the patch for being either malignant or normal. If we only focus on the malignant probability, the network should have a higher confidence for patches corresponding to malignant tissue than for those corresponding to normal tissue that have been incorrectly predicted as malignant. Thus, a 10-bin histogram with the prediction probabilities of the patches classified as malignant for each WSI was calculated. These probabilities were distributed from 50% to 100%, with 5% range for each bin. The histogram was normalized with respect to the total number of tissue patches. Along with the malignant probability histogram (MPH), the least squares regression line (LSRL) of the histogram, defined as $y = mx + b$, was also calculated, where m and b, which refer to the slope and the Y-intercept, are described in equations (1) and (2), respectively. This line represents the best approximation of the set of probabilities for all malignant patches of the corresponding WSI. The mean histogram for both malignant and normal WSIs are shown in Fig. 2 together with their corresponding LSRLs, which are highlighted in red.

$$m = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2} \tag{1}$$

$$b = \frac{N\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2} \tag{2}$$

Where x and y represent the coordinates of the different values of the histogram.

As was previously mentioned, the error of the ML algorithm (a CNN in this case) leads to errors in the classification, which in a WSI is presented as sparse normal tissue patches being classified as malignant. Therefore, in a WSI diagnosed as normal, patches classified as malignant by the CNN are sparsely distributed through the tissue. On the other hand, in a cancerous WSI, malignant-classified patches tend to be focused around the tumor areas. Due to this reason, the dispersion factor of malignant-classified patches was also considered as another relevant input for the slide-level classification between normal and malignant WSIs. This factor was obtained by calculating the number of malignant connected components (MCC), which counts the isolated components (sets of malignant patches) in the classification result according to a specific distance D. Algorithm 1 details the method used to calculate the number of connected components based on the center coordinates of malignant patches and D. In this work, five different values were considered for D (142, 283, 425, 566 and 708 pixels), which correspond to the Euclidean distances (i.e., radii) from a patch to a range of 1 up to 5 patches-distance, taking into account that the distance between two patches is 100 pixels (patches are 100 × 100 pixels size). The number of connected components was normalized with respect to the total number of malignant patches for each WSI. In this way, normal samples with a low quantity of sparse misclassifications are penalized when compared to malignant samples with sparse tumoral tissue regions.

Algorithm 1. Connected components algorithm

Algorithm 1: Connected components algorithm

```
ConnectedPatches (centers, D)
    inputs : A list of center points from malignant patches (centers); a distance (D).
    output: A set of lists of connected patch centers with relative distance D.
    connected_components = [];
    current_component = [];
    while count(centers) > 0 do
        current_component = [];
        current_component.append(centers[0]);
        centers.remove(centers[0]);
        foreach center in current_component do
            foreach point in centers do
                if distance from center to point <= D then
                    current_component.append(point);
                    centers.remove(point);
        connected_components.append(current_component);
```

3.2. Wide & Deep network model

The dataset described in section 3.1 was used as input to a Neural Network (NN) model called W&D [18] to provide a slide-level classification between normal and malignant WSIs. The W&D model combines both wide and deep components. The wide component memorizes sparse interactions between features effectively, which can be defined as learning how the output responds to combinations of sparse input values. On the other hand, the deep component corresponds to the feed-forward neural network which represents the generalization, that is, the ability to handle unseen data. Therefore, the benefits from both memorization (wide) and generalization (deep) are combined and achieved in a single model [18].

In this work, the malignant tissue ratio was used as the wide element while the malignant probability histogram, the slope and Y-intercept of the LSRL and the number of malignant connected components were used

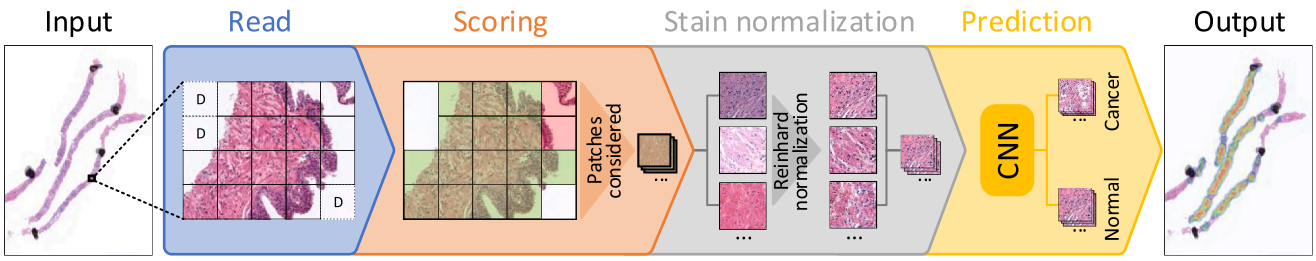

Fig. 1. Block diagram detailing each of the steps considered for processing a whole-slide image (WSI) in PROMETEO. First, in the step called *Read*, patches are extracted from the input WSI, and those corresponding to background are discarded (those identified as D). Then, in the next step, a score is given to each patch in order to discard patches corresponding to unwanted areas, such as pen marks and external agents. This score discriminates considering three factors: the amount of tissue that the patch contains, the percentage of pixels that are within H&E's hue range, and the dispersion of the saturation and brightness channels. Discarded patches in this step are shown in red, while those that pass the scoring filter are highlighted in green. The third step, called *Stain normalization*, applies a color normalization to the patches based on Reinhard's stain-normalization method in order to reduce color variability between samples. Finally, in the *Prediction* step, each of the patches are used as input to the CNN, which classifies them as either malignant or normal tissue. A deeper insight on these steps is given in Ref. [14].

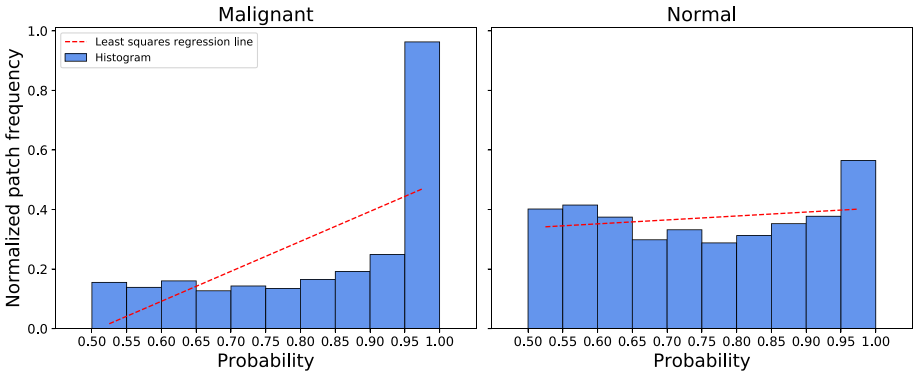

Fig. 2. Mean probability histogram of the normalized patch frequency across all the WSIs, distinguishing between malignant (left) and normal (right) samples. The least squares regression line is shown with a red dashed line. As can be seen, for malignant WSIs, the system tends to classify patches as malignant with a higher confidence. This produces a least squares regression line with a steeper slope. On the other hand, for the normal WSIs, the classification for malignant patches is not that accurate, which leads to a less steep regression line.

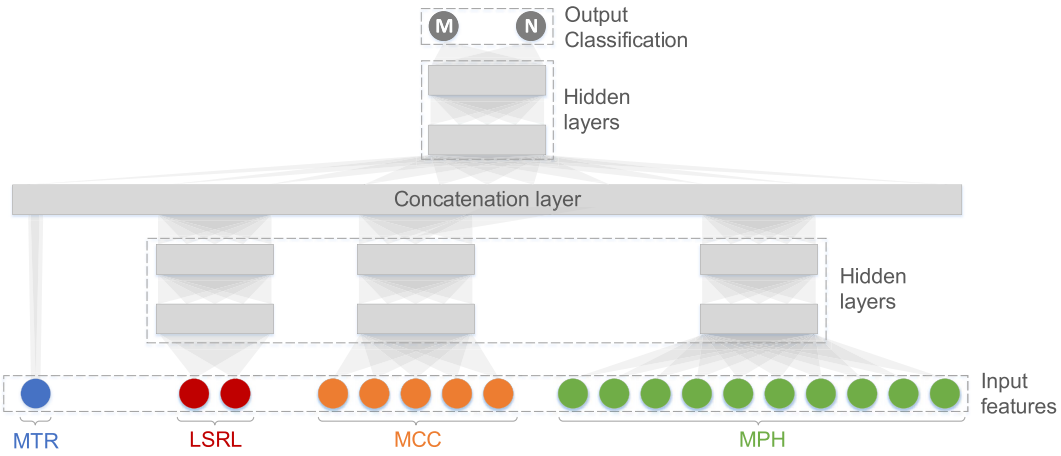

Fig. 3. Diagram of the W&D network model proposed in this work. Each hidden layer consists of 300 neurons. The input features, which are detailed in section 3.1, are: the malignant tissue ratio (MTR) of the WSI, the slope and Y-intercept of the least squares regression line (LSRL) of the histogram, the number of malignant connected components (MCC) with 5 different radii (from 1 to 5 malignant patch distance), and the 10-bin malignant probability histogram (MPH) between 50% and 100% with 5% ticks. These input features are used to classify the WSI as either malignant (M) or normal (N).

as the deep elements. Each of the deep data were separately connected to two hidden layers of 300 neurons. Then, these layers were concatenated together with the wide element to a hidden block of two hidden layers with 300 neurons each. The concatenation layer[1] takes a list of tensors as input and returns a single tensor, which is the result of combining all the inputs together. Finally, this hidden block was connected to the output layer, a SoftMax function which performs the classification of the WSI as either malignant or normal. In this way, complex features are extracted from combinations of sparse inputs and then concatenated together in order to perform the final decision. Each of the hidden layers in this model also contained a Rectified Linear Unit (ReLU) [19] activation layer at the output. The number of hidden layers and the number of neurons per layer were selected by means of Scikit-Learn's Grid Search algorithm (GridSearchCV) [20], where the configuration that achieved

[1] https://keras.io/api/layers/merging/_layers/concatenate (accessed 17th August 2021).

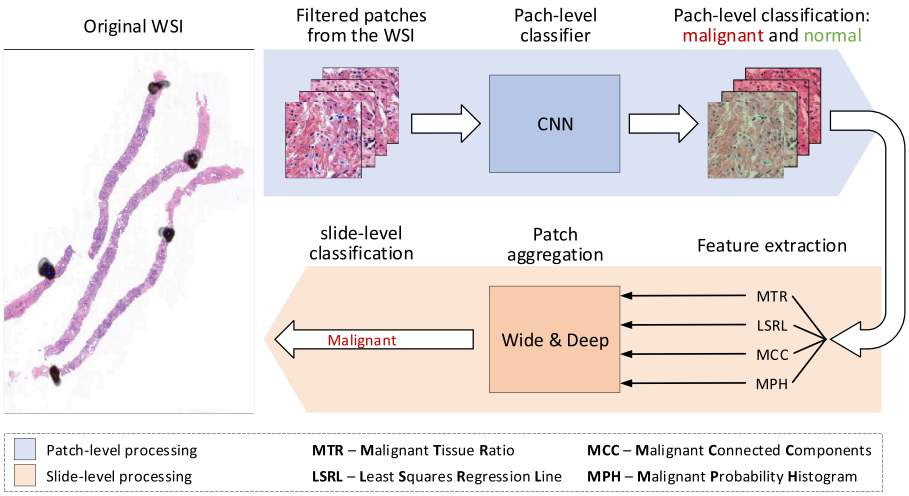

Fig. 4. Diagram of the whole processing step for the PCa screening task. First, the WSI is processed at patch level, following the same procedure presented in Fig. 1. Then, the output classification for each of the filtered patches from the original WSI is used to perform a slide-level prediction using the W&D model presented in Fig. 3, where the extracted features are used to classify the WSI as either malignant or normal.

the best result was selected. The number of hidden layers explored ranged from 1 to 5, and the number of neurons per layer were explored from 100 to 1000, in steps of 100 neurons.

Fig. 3 depicts the custom W&D model used in this work, where the different inputs and layers can be seen. Fig. 4 represents the whole processing step for the prostate screening task, highlighting both the patch-level and the slide-level components.

3.3. Training and validation

K-fold stratified cross-validation was performed to measure the generalization ability of the model. This technique consisted in dividing the dataset in 5 sets ($K = 5$). For each fold, the network was trained using four of the five sets (80% of the dataset) for 10,000 epochs and validated using the remaining one (20% of the dataset). Thus, for each experiment, the network was trained and validated a total of five times with different data. The EarlyStopping algorithm was used, which stopped the training step when the validation loss stopped improving after the last 10 epochs, which prevents the model from overfitting. The final results are presented as the mean accuracy calculated over the five cross-validation folds.

To validate the network, different evaluation metrics were used. These were the accuracy (eq. (3)), sensitivity (eq. (4)), precision (eq. (5)), F1 score (eq. (6)) and AUC of the Receiver Operating Characteristic (ROC) curve.

$$\text{Accuracy} = 100 \times \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3}$$

$$\text{Sensitivity} = 100 \times \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{Precision} = 100 \times \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{F1score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{6}$$

Where TP and FP denote true positive cases (when the model diagnoses a malignant WSI correctly) and false positive cases (the network diagnoses a normal WSI as malignant), respectively. TN and FN denote true negative cases (the system classifies a normal WSI as normal) and false negative cases (the network classifies a malignant WSI as normal), respectively.

For designing, training and validating the model, both TensorFlow[2] and Keras[3] were used.

3.4. Test methods

Pixel-wise annotations indicating the locations of cancerous tissue inside the WSIs provided by the expert pathologist panel were extracted. From malignant WSI, multiple overlapping patches were obtained from the tissue area contained within the extracted annotations. A patch dimension of 100 × 100 pixels at 10× magnification was used. From WSIs diagnosed as normal, similar patches were extracted from the tissue. Combining these normal and malignant patches and applying data augmentation, a total of 57 million patches were obtained. These were used to train a custom CNN called PROMETEO.

We used this trained model to detect cancer presence in 332 new WSIs at patch level. Patches were extracted from each of these images, discarding those corresponding to background based on a color filter. A patch scoring filter was also applied to remove unwanted areas, such as pen marks and other artifacts. All the patches from a WSI that passed these filters were stain-normalized to remove color variations caused by the H&E stain process and predicted by the trained CNN.

All the patches classified as malignant from a single WSI were collected, and different features (MTR, LSRL, MCC and MPH) were automatically extracted from them using Python based on the quantity of malignant patches and their position within the WSI. These features, together with the ground-truth label from the pathology report, were used to train and evaluate a custom W&D model.

After both the CNN and the W&D were trained, the process for evaluating and testing a new WSI consists of the following steps: first, patches are extracted, filtered and stain-normalized; then, these are used as input for the CNN; aggregated features obtained from malignant patches are then used as input for the W&D, which reports the final output of the system. The output is a single label for each WSI, which aggregates the patch-level labels obtained by the CNN. The system also reports the heatmap of the input image in case it is classified as malignant.

[2] https://www.tensorflow.org (accessed 17th August 2021).
[3] https://keras.io (accessed 17th August 2021).

Table 1
Validation results obtained with the proposed W&D model. The accuracy, sensitivity, precision, F1 score and AUC are shown for each of the different cross-validation folds. The average of the obtained metrics across the five folds is also presented.

Fold	Accuracy (%)	Sensitivity (%)	Precision (%)	F1 score (%)	AUC
1	93.93	100	89.74	94.59	0.93
2	93.93	97.29	92.30	94.73	0.93
3	95.45	100	90.32	94.91	0.96
4	93.93	100	87.09	93.10	0.94
5	93.93	97.05	91.66	94.28	0.93
Average	**94.24**	**98.87**	**90.23**	**94.33**	**0.94**

3.5. Statistical analysis

After being advised by the group of expert pathologists that supervised this work, they considered the number of cases and patients used as good for representing case diversity. Each of these cases was analyzed and labeled by a pathologist from the expert panel. In order to provide robust results, multiple evaluations were performed, following the 5-fold cross-validation presented in section 3.3, in which, samples from the same patient were not considered for both training and testing the model at the same time. In these experiments, we compared the cancer presence in each WSI with the report from pathologists following their supervision to evaluate how well the system agreed with them. Different widely-used metrics for the evaluation of DL-based systems were used and calculated to provide a complete analysis of the results and the performance of the proposed CAD model.

Statistical analysis was performed using Python 3.8 with the NumPy (1.19.5), scikit-learn (0.24.2), scipy (1.6.3) and TensorFlow (2.5.0) packages.

4. Results

4.1. Evaluation of the proposed system

After training the custom W&D model (section 3.2) with the dataset presented in section 3.1, all the different metrics described in section 3.3 were calculated and obtained in order to evaluate the proposed system. Table 1 summarizes the results for each cross-validation fold together with the average for all the evaluation metrics. With these, the average results were calculated, achieving an accuracy of 94.24%, a sensitivity of 98.87%, a precision of 90.23%, a F1 score of 94.33% and an AUC of 0.94.

As can be seen, the results obtained across the different folds are consistent and the proposed model achieves very high scores in all the different metrics studied for this classification task, particularly in terms of sensitivity. The sensitivity, which in this field is defined as the ability of the system to identify PCa, is of utmost importance for reporting and assessing the performance of the screening test [21]. The proposed system is able to achieve an average sensitivity of around 99%, where three of the folds achieved perfect sensitivity (100%). This means that our custom model makes almost no mistakes when predicting a malignant sample as such, making it a reliable patch aggregation method, together with PROMETEO, for PCa detection in WSIs. Fig. 5 shows some examples of correctly classified WSIs.

In order to evaluate the effect of the different inputs considered for the proposed W&D model on the results, a new experiment was carried out. This experiment consisted in removing each of the inputs and performing the same 5-fold cross-validation, obtaining the average of the metrics. These results can be seen in Table 2, where each row represents the average of the metrics obtained for a 5-fold cross-validation test where a specific input is not used. As can be seen, the sensitivity is considerably reduced when the MCC is not used, although none of the experiments achieved values as high as those obtained when all the proposed inputs are considered. Since the F1 score and the AUC depend on the sensitivity, these results are also lower in the same aforementioned cases. On the other hand, the precision does not vary that much between the different experiments, and a clear effect of removing an input can only be seen when the MPH is not used.

4.2. Comparison with other widely-used machine learning models

The results obtained in this study were compared with different ML-based methods and classifiers using the same dataset. The following well-known machine learning algorithms were used to classify the WSIs: an Artificial Neural Network (ANN) [22], a Support Vector Machine (SVM) [23], a RF [24] and a k-Nearest Neighbors (KNN) [25]. In order to perform a fair comparison, a similar Grid Search algorithm was used. For the RF and the KNN, the number of trees/neighbors explored were 5, 10, 20, 50, 100, 200 and 300. The ones that achieved the best results were 100 trees in the case of the RF, and 20 neighbors in the case of the KNN. For the ANN, the exact same search that was used for the W&D was considered, ending up with 3 hidden layers and 200 neurons per layer as the best model. For the SVM, the default value of the C parameter (1.0) was used.

Table 3 summarizes the results obtained for each method, which are represented as the average of the evaluation metrics (see section 3.3) obtained for each cross-validation fold.

As can be seen, the best results for accuracy, sensitivity, F1 score and AUC are obtained with the proposed W&D model, with the exception of precision, for which SVM achieves the highest value. As was previously mentioned, sensitivity is the most relevant metric for measuring the performance of a classifier when performing a screening test. In this case, the proposed architecture is the one achieving the highest sensitivity score among the different algorithms evaluated, with a difference of more than 6% with the second highest, i.e., the ANN. On the other hand, SVM achieves around 99% precision, which could be very relevant for other binary or multi-class classification tasks, but not as much as the sensitivity when developing a ML-based PCa screening method that could help experts to speed up the whole process.

Since the SVM reported the highest precision result, a more detailed evaluation of this architecture was performed by changing the value of the C hyperparameter. The selected values for this experiment ranged from 2^{-6} to 2^{6}, with step size of power of 2. For each of these values, the SVM was trained and evaluated using the aforementioned 5-fold cross-validation scheme, and the average of the results were computed. These results can be seen in Table 4, where the average accuracy, sensitivity, precision, F1 score and AUC are shown for each C. As can be seen, changing the C hyperparameter allows achieving 100% precision in some cases and around 84% sensitivity in others, but they do not differ that much compared to the results obtained with the default C value (1.0).

5. Discussion

The results presented in this work, which have been supervised and validated by a panel of expert pathologists, are promising, and we hope to incorporate these techniques in hospitals in the near future to support experts in real case scenarios. Section 2 presented related studies, which aimed to address the same problem in different ways [9–13]. Some of them have achieved very good results when aggregating the information from patch-wise predictions into a single slide-wise label, and have proposed interesting solutions to this problem, but none of them have considered the use of the W&D model, which was originally conceived for recommender systems. This model allows aggregating different independent features, benefitting from its wide component, as well as processing and extracting relevant information from them independently, thanks to its deep nature, which perfectly fits this task and could lead to a very robust system. Other simpler alternatives for patch aggregation like the one presented in Ref. [13], which classifies a WSI as

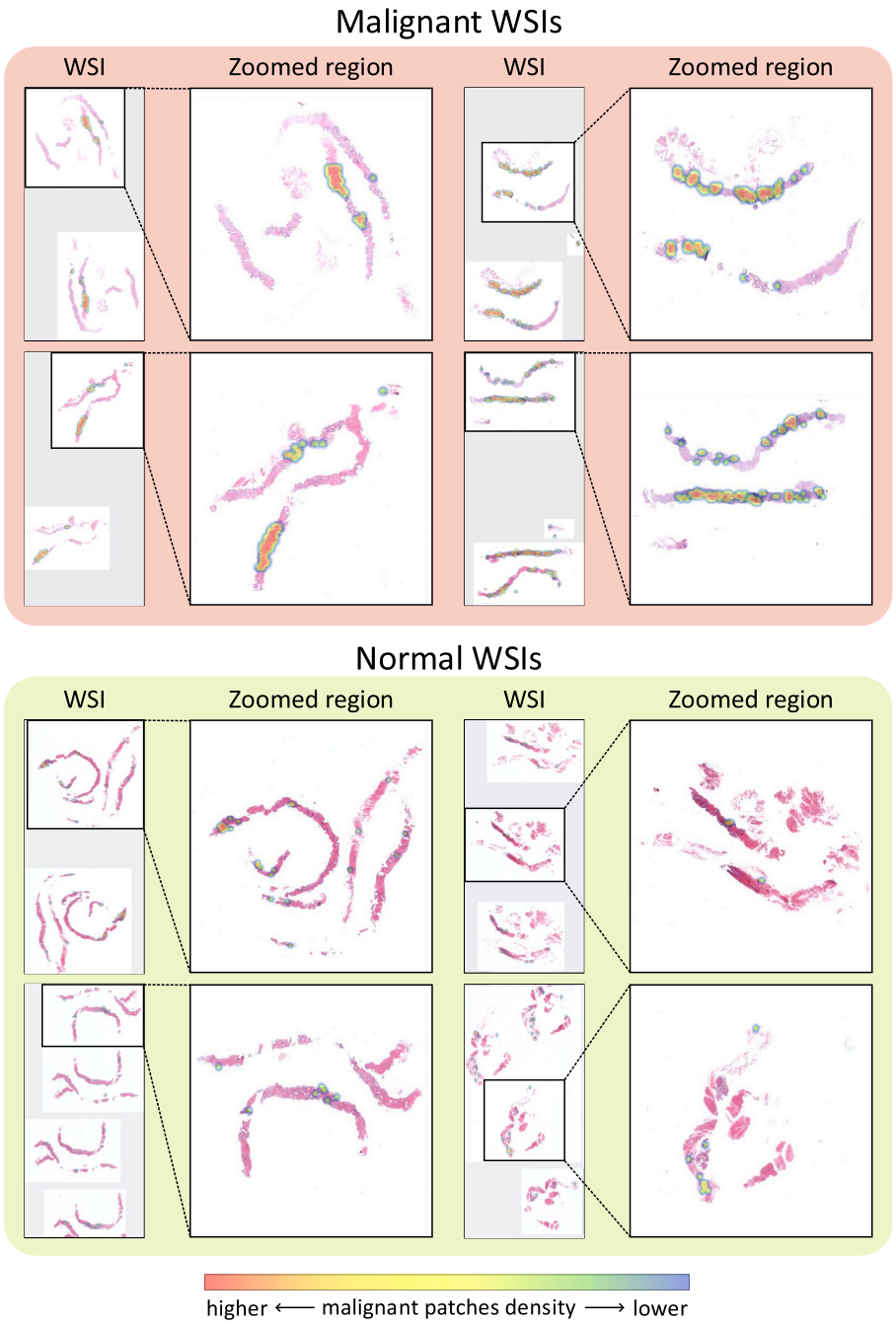

Fig. 5. Eight different WSI samples extracted from the dataset presented in section 3.1. A heatmap of the malignant patches predicted by PROMETEO is drawn on top of the WSI, and zoomed regions are presented for better visualization. Red regions represent higher concentrations of malignant patches, while blue represent the opposite. The examples presented were correctly classified by the proposed W&D model.

malignant if at least 10% of the tissue is predicted as malignant, could result in a high number of false negatives depending on the patient and sample. Smaller percentages of malignant tissue could also represent malignant cases, which is the reason why threshold-based alternatives may not be the best option, as seen in the max pooling solution presented in Ref. [10].

Furthermore, since each of the works mentioned in section 2 use a different dataset, results should not be directly compared neither between them nor with the ones obtained in our work if the aim is to conduct a strict and fair comparison. Therefore, our system was compared with different well-known machine learning alternatives that were used by the authors in the related work, such as ANNs, SVMs, RFs and KNNs, which were trained and tested on the same dataset that we used. The proposed W&D-based system achieved the highest sensitivity among the different machine learning alternatives evaluated. As was previously mentioned, the sensitivity is the most relevant metric to take into consideration in a screening test, since higher sensitivity means that there are fewer false negatives. An intelligent system with very high sensitivity could be implemented as a prescreening triage tool and help pathologists by prioritizing the most severe cases.

A malignant/normal slide-level label could also be used for the next steps in the analysis process. In this way, in case a WSI is predicted as malignant, the patches obtained from it could be passed to the next system, dedicated to assigning a GGS score. Therefore, WSIs predicted as normal would not need further processing, avoiding unnecessary operations and reducing the response time of the CAD system.

Table 2
Validation results obtained with the proposed W&D model when different inputs are used. The average accuracy, sensitivity, precision, F1 score and AUC calculated over the 5 different cross-validation folds are shown. Different inputs were used in order to evaluate their effect in the output metrics. The first column indicates which input was removed. The last row of the table corresponds to the results obtained using all the inputs considered in this work (also shown in Table 1).

Inputs removed	Accuracy (%)	Sensitivity (%)	Precision (%)	F1 score (%)	AUC
MTR	92.42	91.86	91.68	91.53	0.92
LSRL	90.91	93.06	88.63	90.64	0.91
MCC	90.91	92.10	89.06	90.29	0.90
MPH	89.69	93.55	86.42	89.48	0.89
None	**94.24**	**98.87**	**90.23**	**94.33**	**0.94**

Table 3
Validation results calculated from the average of the evaluation metrics (accuracy, sensitivity, precision, F1 score and AUC) for the 5 different cross-validation sets. The results obtained with the proposed W&D model are compared to other state-of-the-art ML-based algorithms, namely, ANN, SVM, RF and KNN. The best result for each specific evaluation metric is highlighted in bold.

Model	Accuracy (%)	Sensitivity (%)	Precision (%)	F1 score (%)	AUC
W&D (proposed)	**94.24**	**98.87**	90.23	**94.33**	**0.94**
ANN	89.69	92.47	87.29	89.54	0.89
SVM	88.18	80.78	**98.76**	88.79	0.89
RF	88.84	84.89	92.23	88.22	0.88
KNN	88.48	83.29	94.31	88.31	0.88

Table 4
Validation results calculated from the average of the evaluation metrics (accuracy, sensitivity, precision, F1 score and AUC) for the 5 different cross-validation sets obtained with an SVM. Different values were used for the C parameter of the SVM, ranging from 2^{-6} to 2^{6}. For each value of C, the 5-fold cross-validation was performed, and the average of the metrics are reported.

C	Accuracy (%)	Sensitivity (%)	Precision (%)	F1 score (%)	AUC
2^{-6}	80.91	71.47	100.0	83.25	0.86
2^{-5}	79.70	70.24	100.0	82.4	0.85
2^{-4}	81.52	72.03	100.0	83.68	0.86
2^{-3}	83.94	74.73	100.0	85.49	0.87
2^{-2}	86.06	77.35	100.0	87.19	0.89
2^{-1}	87.58	79.91	98.69	88.30	0.89
1.0	88.18	80.78	98.76	88.79	0.89
2^{1}	89.39	82.42	98.69	89.82	0.91
2^{2}	89.39	82.54	98.69	89.87	0.91
2^{3}	89.70	82.92	98.69	90.11	0.91
2^{4}	90.30	83.86	98.69	90.64	0.91
2^{5}	89.39	82.84	98.11	89.82	0.90
2^{6}	88.48	81.97	96.84	88.63	0.89

As a future work, our main goal is to develop and include a low-latency GGS recognition system and, as was previously mentioned, include it in the modular CAD system presented in this work. Thus, the whole system would be able to report whether a WSI is either normal or malignant, and, in the latter case, report both a global GGS score and a heatmap with the malignant areas highlighted. All of this will be designed with the focus on achieving very high sensitivity and very low computation time, being able to provide a fast response to an input image.

The proposed patch-aggregation technique based on the use of W&D models could be extrapolated to other tissue types within histopathology, allowing other new independent input features to be added to the model in a simple way with their corresponding specific hidden layers to extract relevant information from them.

6. Conclusions

In this work, the authors present a novel ML-based method to classify WSIs of prostate tissue as normal or malignant at global slide level based on a previous patch-level classification. This classification is based on a novel NN model called W&D, which combines both linear model components (wide) and neural network components (deep) in order to achieve both memorization and generalization in a single model. The custom W&D proposed model classifies each WSI as normal or malignant considering different processed inputs. This information was obtained using PROMETEO, a CAD system that extracts small patches from WSIs, which are first pre-processed and then classified, reporting a heatmap that shows where malignant areas are located inside the corresponding WSI. From the information obtained from malignant patches, different processed features are calculated, which are then used as input to the proposed W&D model. These are the malignant tissue ratio, the 10-bin malignant probability histogram between 50% and 100% with 5% ticks, the slope and Y-intercept of the least squares regression line of the histogram and the number of malignant connected components with 5 different radii. The network was trained and validated using 5-fold cross-validation. The average results obtained for the cross-validation sets with the W&D model achieved an accuracy of 94.24%, a sensitivity of 98.87%, a precision of 90.23%, a F1 score of 94.33% and an AUC of 0.94. The proposed model was compared with other state-of-the-art methods (ANN, SVM, RF and KNN) using the same dataset. The results show that the W&D model performs better in terms of accuracy, sensitivity, F1 score and AUC. The promising results obtained with this novel model show that the proposed system could aid pathologists when analyzing histopathological images as a screening method, discriminating between normal and malignant PCa slides.

Credit authorship contribution statement

L. Duran-Lopez: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Juan P. Dominguez-Morales:** Conceptualization, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing. **D. Gutierrez-Galan:** Validation, Visualization, Writing - review & editing. **A. Rios-Navarro:** Validation, Visualization, Writing - review & editing. **A. Jimenez-Fernandez:** Validation, Visualization, Writing - review & editing. **S. Vicente-Diaz:** Funding acquisition, Project administration, Resources, Writing - review & editing. **A. Linares-Barranco:** Funding acquisition, Project administration, Resources, Writing - review & editing.

Declaration of competing interest

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

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Acknowledgments

We would like to thank Antonio Felix Conde-Martin and the Pathological Anatomy Unit of Virgen de Valme Hospital in Seville (Spain) for their support in the PROMETEO project, together with VITRO S.A., as well as for providing us with annotated WSIs from the same hospital. The authors would also like to thank the Spanish Agencia Estatal de Investigación (AEI) for supporting this work. This work was partially funded by Spanish Agencia Estatal de Investigación (AEI) project MINDROB (PID2019- 105556 GB-C33/AEI/10.13039/501100011033), with support from the European Regional Development Fund, by the EU H2020 project CHIST-ERA SMALL (PCI2019-111841-2) and by the

Andalusian Regional Project PAIDI2020 (with FEDER support) PROMETEO (AT17_5410_USE).

References

[1] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, Freddie Bray, Global Cancer Statistics 2020: Globocan Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries, A Cancer Journal for Clinicians, CA, 2021.

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

[3] Nigel Borley, Mark R. Feneley, Prostate cancer: diagnosis and staging, Asian J. Androl. 11 (1) (2009) 74.

[4] Pavel Hamet, Johanne Tremblay, Artificial intelligence in medicine, Metabolism 69 (2017) S36–S40.

[5] S Ahuja Abhimanyu, The impact of artificial intelligence in medicine on the future role of the physician, PeerJ 7 (2019), e7702.

[6] Dinggang Shen, Guorong Wu, Heung-Il Suk, Deep learning in medical image analysis, Annu. Rev. Biomed. Eng. 19 (2017) 221–248.

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

[8] A. Fourcade, R.H. Khonsari, Deep learning in medical image analysis: a third eye for doctors, Journal of stomatology, oral and maxillofacial surgery 120 (4) (2019) 279–288.

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

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

[11] Geert Litjens, Clara I. Sánchez, Nadya Timofeeva, Meyke Hermsen, Iris Nagtegaal, Iringo Kovacs, Christina Hulsbergen-Van De Kaa, Peter Bult, Bram Van Ginneken, Jeroen Van Der Laak, Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis, Sci. Rep. 6 (1) (2016) 1–11.

[12] Wenyuan Li, Jiayun Li, Karthik V. Sarma, King Chung Ho, Shiwen Shen, Beatrice S. Knudsen, Arkadiusz Gertych, Corey W. Arnold, Path r-cnn for prostate cancer diagnosis and gleason grading of histological images, IEEE Trans. Med. Imag. 38 (4) (2018) 945–954.

[13] Wouter Bulten, Hans Pinckaers, Hester van Boven, Robert Vink, Thomas de Bel, Bram van Ginneken, Jeroen van der Laak, Christina Hulsbergen-van de Kaa, Geert Litjens, Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study, Lancet Oncol. 21 (2) (2020) 233–241.

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

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

[16] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, Peter Shirley, Color transfer between images, IEEE Computer Graphics and Applications 21 (5) (2001) 34–41.

[17] Derek Magee, Darren Treanor, Doreen Crellin, Mike Shires, Katherine Smith, Kevin Mohee, Philip Quirke, Colour normalisation in digital histopathology images, In*Proc. Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy (MICCAI Workshop)* 100 (2009) 100–111. Citeseer.

[18] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al., Wide & deep learning for recommender systems, in: Proceedings of the 1st Workshop on Deep Learning for Recommender Systems, 2016, pp. 7–10.

[19] Kazuyuki Hara, Daisuke Saito, Hayaru Shouno, Analysis of function of rectified linear unit used in deep learning, in: 2015 International Joint Conference on Neural Networks (IJCNN), IEEE, 2015, pp. 1–8.

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

[21] Matti Hakama, Anssi Auvinen, Nicholas E. Day, Anthony B. Miller, Sensitivity in cancer screening, J. Med. Screen 14 (4) (2007) 174–177.

[22] Bayya Yegnanarayana, Artificial Neural Networks, PHI Learning Pvt. Ltd., 2009.

[23] Lipo Wang, Support Vector Machines: Theory and Applications, 177, Springer Science & Business Media, 2005.

[24] Leo Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[25] Liangxiao Jiang, Zhihua Cai, Dianhong Wang, Siwei Jiang, Survey of improving k-nearest-neighbor for classification, in: Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007), 1, IEEE, 2007, pp. 679–683.

L. Duran-Lopez received the B.S. degree in Biomedical Engineering, the M.S. degree in Biomedical Research, and the Ph.D. degree in industrial computerized, robotic and neuromorphic systems from the University of Seville (Sevilla, Spain), in 2016, 2017 and 2021, respectively.

Since 2021, she has been working as Postdoctoral Researcher and Lecturer in the Robotics and Technology of Computers Lab. at 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.

D. Gutierrez-Galan received the B.S. degree in computer engineering in 2014 and the M.S. degree in computer engineering and networks in 2016, both from the University of Seville, Sevilla, Spain. Since September 2017, he has been a Ph.D. student in the Department of Computer Architecture and Technology, at University of Seville, with a research grant from the Spanish Ministry of Education and Science. His research interests include embedded systems programming, digital design, FPGA, spiking neural networks in embedded systems for audio processing, neuromorphic auditory sensors and neuromorphic robots.

A. Rios-Navarro received the B.S. degree in computer science engineering, the M.S. degree in computer engineering, and the Ph.D. degree in neuromorphic engineering from the University of Seville, Seville, Spain, in 2010, 2011, and 2017, respectively. He currently holds an Assistant Professor position at the Computer Architecture and Technology Department, University of Seville. His current research interests include neuromorphic systems, real-time spikes signal processing, field-programmable gate array design, and deep learning.

A. Jimenez-Fernandez received the B.S. Degree in Computer Engineering in 2005, the M.S. Degree in Industrial Computer Science in 2007 and the Ph.D. in Neuromorphic Engineering in 2010 from the University of Seville, Sevilla, Spain. In October 2007 he became Assistant Professor at the department of Computer Architecture and Technology of the University of Seville, and in 2021 has been promoted to Associate Professor of the same department. He is currently coordinator of the Biomedical Engineering degree of the University of Seville, and is cofunder of the Spin-Off COBER, mainly devoted to biomedical robotics. His research interests include neuromorphic engineering applied to robotics, real-time spikes signal processing, neuromorphic sensors, FPGA digital design, embedded intelligent systems development, and biomedical robotics.

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 cofunder 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, Airforce), he started as an Assistant Professor at the Architecture and Technology of Computers Department of the University of Sevilla in 2001. In 2021, he was promoted to Full Professor. 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 2021 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.