Computer Methods and Programs in Biomedicine 226 (2022) 107108

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Non-small cell lung cancer diagnosis aid with histopathological images using Explainable Deep Learning techniques

Javier Civit-Masot[a], Alejandro Bañuls-Beaterio[a], Manuel Domínguez-Morales[a,b,*], Manuel Rivas-Pérez[a], Luis Muñoz-Saavedra[a], José M. Rodríguez Corral[c]

[a] *Architecture and Computer Technology department (ATC), Robotics and Technology of Computers Lab (RTC), E.T.S. Ingeniería Informática, Avda. Reina Mercedes s/n, Universidad de Sevilla, Seville, 41012, Spain*

[b] *Computer Engineering Research Institute (I3US), E.T.S. Ingeniería Informática, Avda. Reina Mercedes s/n, Universidad de Sevilla, Seville, 41012, Spain*

[c] *Computer Science department, School of Engineering, Avda. Universidad de Cádiz 10, Universidad de Cádiz, Puerto Real (Cádiz), 11519, Spain*

ARTICLE INFO

Article history:
Received 12 April 2022
Revised 26 July 2022
Accepted 1 September 2022

Keywords:
Deep learning
Explainable deep learning
Convolutional neural networks
Lung cancer
Histopathology

ABSTRACT

Background: Lung cancer has the highest mortality rate in the world, twice as high as the second highest. On the other hand, pathologists are overworked and this is detrimental to the time spent on each patient, diagnostic turnaround time, and their success rate.

Objective: In this work, we design, implement, and evaluate a diagnostic aid system for non-small cell lung cancer detection, using Deep Learning techniques.

Methods: The classifier developed is based on Artificial Intelligence techniques, obtaining an automatic classification result between healthy, adenocarcinoma and squamous cell carcinoma, given an histopathological image from lung tissue. Moreover, a report module based on Explainable Deep Learning techniques is included and gives the pathologist information about the image's areas used to classify the sample and the confidence of belonging to each class.

Results: The results show a system accuracy between 97.11 and 99.69%, depending on the number of classes classified, and a value of the area under ROC curve between 99.77 and 99.94%.

Conclusions: The classification results obtain a substantial improvement according to previous works. Thanks to the given report, the time spent by the pathologist and the diagnostic turnaround time can be reduced.

1. Introduction

In 2020, according to the World Health Organization (WHO), lung cancer was the second with the highest number of new cases worldwide (2.21 million cases), only surpassed by breast cancer (2.26 million cases). However, in that same year, it was the most lethal cancer worldwide (1.80 million deaths) and, in the second place, colon and rectal cancer with half the deaths [1].

Although society's control and awareness campaigns help in the early detection of cancer, the most common and efficient way to verify the presence of a cancerous tumor is by tissue biopsy. For this goal, pathological anatomy professionals perform histopathological studies on tissue samples amplified under the microscope, to verify the existence or not of cancer cells in the sample.

The reports carried out by the pathologist, according to the National Cancer Institute (United States), are issued within 10 days after the biopsy [2]. This situation causes a problematic delay in the cancer diagnosis. Of special interest for this research team is the Spain country where, according to the latest report from the Ministry of Health, Consumption and Social Welfare of 2019, in the country there were 1367 specialists in pathological anatomy; however, according to the Spanish Society of Medical Oncology, approximately 1.5 million active cases of cancer are estimated in Spain. This implies that, on average, a pathologist can dedicate around 1 h and a half per year for each patient [3].

Because of that, it is useful to use diagnostic aid tools in this type of analysis to reduce pathologists' interventions, reducing the time needed to spend on each patient and freeing the pathologist

[*] Corresponding author.
E-mail address: mjdominguez@us.es (M. Domínguez-Morales).

https://doi.org/10.1016/j.cmpb.2022.107108
0169-2607/© 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

from analyzing multiple easy-to-diagnose samples. In this way, a preliminary report could be provided to the pathologist with the evidence detected, and the pathologist could carry out pertinent checks to validate or refute the preliminary diagnosis. This would be useful if the system provided a report in much less time than is required for manual analysis. Not surprisingly, there are multiple research works where diagnostic aid systems are designed and implemented for medical image analysis.

At this point, the application of Artificial Intelligence (AI) techniques is of great importance to designing classifier systems capable of extracting characteristics from images and differentiating between those that show some type of disease and those that represent a healthy patient [4].

The application of this type of technique in medical imaging provides three main benefits:

- Mass case filtering: easily diagnosable cases may have a quick analysis, reducing the time spent by the specialist.
- Specialists' workload reduction: as a consequence of the above, the specialist can spend more time on severe and/or difficult to diagnose cases. As a secondary implication, false negatives could be reduced.
- Diagnosis time reduction: as a consequence of the previous benefits, both the specialist and the patient could know the diagnosis in advance and, therefore, the action plan in case of detection of the disease could be streamlined.

AI fields specialized in developing classifiers to help in diagnostic tasks are Machine Learning (ML) and Deep Learning (DL). These techniques have been applied in multiple research works in recent years, obtaining very positive results with a higher rate of correct diagnoses than 80% [5–7]; even reaching, in multiple cases, values higher than 95% accuracy [8–10].

According to previous statistical studies in cancer diagnosis such as the one carried out by Mark Priebe and Markin [11], generally in medical imaging guided diagnosis, the mean percentage of discrepancies in the diagnosis reports is 12%; So, any diagnosis-aided system that exceeds 88% success, in theory, would have a greater success rate than the pathologist. However, the main objective of these systems is not to replace the pathologist, but to serve as an aid tool to reduce their workload, always taking into account a final intervention by the pathologist to validate the results.

However, very few automatic diagnostic systems developed to date achieve a 100% accuracy. Still, such systems are tested with a subset of samples from the dataset itself used to train it; this means that it may be mistaken in future classifications if samples from other medical centers are introduced, or if they have been captured with other digitizing devices (among other possible variables involved) [12].

These circumstances are added to the fact that, when a DL system is trained, the weights of the neural network connections do not provide information understandable by the user that helps him/her to know the objective criteria used by the system to perform the classification. For this reason, these systems are referred to be as “black boxes” [13].

Due to this, in recent years the use of Explainable Artificial Intelligent (xAI) technologies has taken on enormous importance, which, through various and varied subsequent analyzes, provide information about the objective classification criteria used in the automatic system [14,15]. This objective information obtained after these analyzes is of great importance, not only to detect possible classification errors, but also to understand the decisions made in the correct classifications. That is why this type of analysis is essential in diagnostic aid systems [16,17].

This research group has extensive experience in the field of Machine and Deep Learning applied to e-Health. This experience can be appreciated, for example, in the physiological signal processing field [18], biomechanical gait studies [19], fall detectors [20,21], etc. Moreover, in addition to this, this group provides experience in the medical imaging processing field using convolutional neural networks (CNN), developing multiple aid systems for cancer and other illnesses diagnosis [22–27].

Therefore, in accordance with what was previously presented and the demonstrated experience of this group, in this work a diagnosis aid system for non-small cell lung cancer diagnosis is designed, implemented, and tested using histopathological images. In addition to explaining the procedure and the results obtained, a comparison will be made with previous work in the area. And, finally, to provide information to the healthcare professional, Explainable Deep Learning (xDL) techniques will be used to study the areas of histopathologycal images in which the classifier has focused to carry out the decisions; in this way, not only the possible diagnosis but also the areas of images in which the classifier has been fixed to obtain the diagnosis can be provided to the pathologist.

The rest of the manuscript is structured as follows: The methods used to develop and test the diagnosis aid system are presented in Section 2. The results obtained after testing the classifier are detailed in Section 3. Finally, in Section 4, the discussion regarding the results obtained is included, as well as a comparison with previous works and the final report provided is discussed too, including the final conclusions of this work and future research lines.

2. Methods

To carry out the previously exposed study, this section presents the dataset used to train and test the classification system, the preprocessing stage for the input data, the designed classifiers, the metrics used to evaluate them and the postprocessing stage applying xAI techniques to extract more information from the classifier behavior.

Fig. 1 shows the general processing scheme of the system proposed in this work.

2.1. Dataset

To develop a robust and reliable system, a dataset with a good number of histopathological images was sought. The dataset used is the so-called LC25000 [28]. It contains lung and colon tissue images, but this work has focused only on the lung cancer images contained in the dataset. Those images represent zoomed sections of biopsied tissue observed under a microscope. All malignant tumour cases in this dataset are non-small cell lung cancer (NSCLC). It is important to mention that the samples taken for this dataset are previously treated with himnunistochemistry techniques to enhance the features that distinguish cancer cells from healthy cells. This is a fundamental difference to other datasets, and is now an essential practice in automatic sample analysis devices.

NSCLC represents 85% of all lung cancers and includes all types of epithelial lung cancer except small cell lung cancer (SCLC). The most common types of NSCLC are squamous cell carcinoma, large cell carcinoma, and adenocarcinoma. NSCLC is usually a less sensitive type of cancer to chemotherapy and radiotherapy compared to SCLC. It is therefore imperative to develop a reliable and rapid early detection mechanism. Among the subtypes of NSCLC listed above, the dataset includes two of them: adenocarcinoma and squamous cell carcinoma.

Therefore, this dataset contains 15,000 lung images: 5000 from healthy tissue, 5000 contains adenocarcinoma, and 5000 contains squamous cell carcinoma. All images have a resolution of 768 x 768 pixels in jpeg format. The distribution of images for each

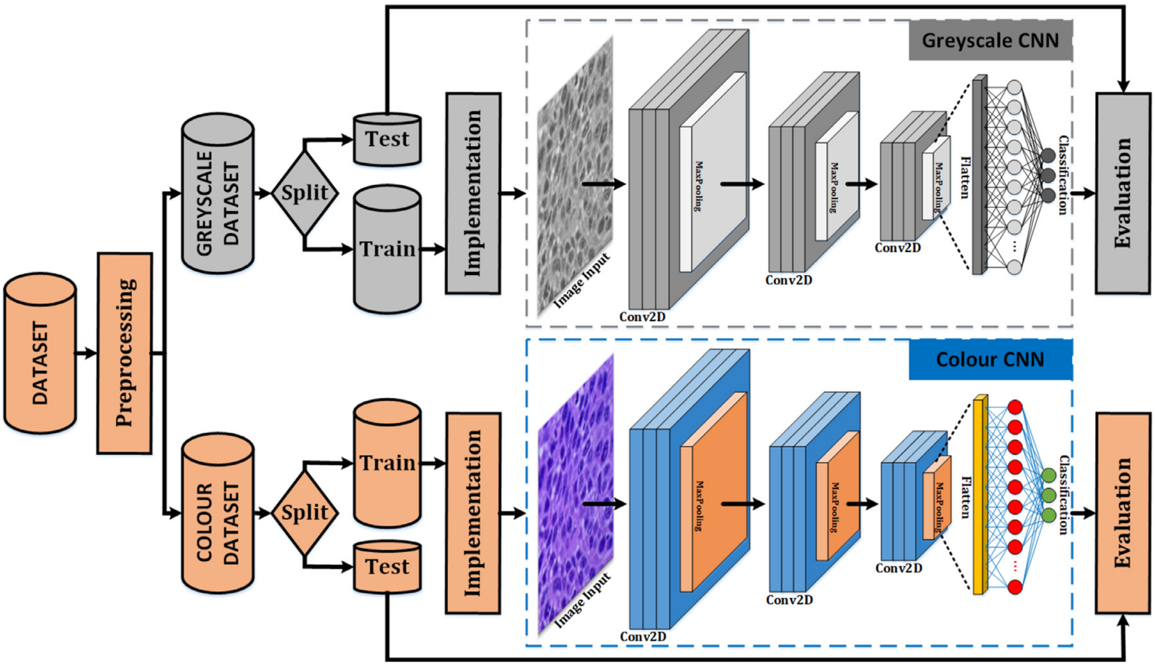


Fig. 1. Full system scheme.

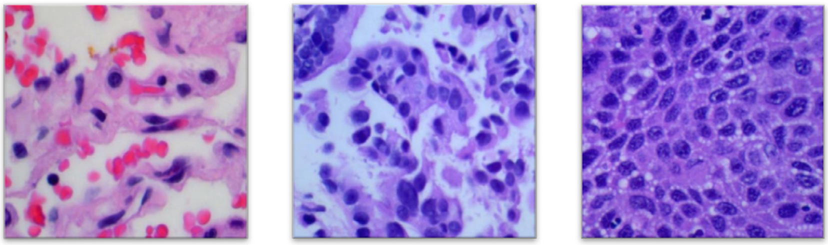


Fig. 2. Images example from dataset LC25000. (left) Healthy tissue; (middle) Adenocarcinoma; (right) Squamous Cell Carcinoma.

Table 1
Number of images of each class and distribution between train (80%) and test (20%).

Class	Train	Test	Total
Healthy	4000	1000	5000
Adenocarcinoma	4000	1000	5000
Squamous-cell carcinoma	4000	1000	5000
TOTAL	**12000**	**3000**	**15000**

tagged class and the subsets' size used to train and test the classifiers are detailed in Table 1.

Carcinoma is a type of cancer that starts in cells that cover the outside or inside of an organ (epithelial). On the other hand, adenocarcinoma originates in glandular tissue and is made up of cells that are capable of secreting substances into the body, and may be a differentiating parameter to squamous cell carcinomas.

Looking at some selected images from the dataset, some differentiating parameters can be detected between the classes (not in all of them). See Fig. 2.

In the benign tissue (Fig. 2-left), lung parenchyma tissue (alveolar septa made up of fibrous connective tissue) can be seen. In addition, blood vessels and flat cells are present. Regarding adenocarcinoma, it is usually located in areas where cells would normally secrete substances such as mucus. This presence of mucus is perhaps the most noticeable feature to the naked eye, distinguishable by the white areas in the image (see Fig. 2-middle). Finally, squamous cell carcinoma would be formed by flat cells lining the inside of the airways (see Fig. 2-right).

2.2. Preprocessing stage

In the preprocessing stage, images are adapted to the classifier's input by size and colour adaptation, to reduce the computational requirements (reducing the time spent in the classification) and facilitate the feature extraction process (increasing the system accuracy). The full preprocessing stage is shown in Fig. 3.

Firstly, the original resolution of the images (768x768 pixels) is compressed to a lower resolution (180x180 pixels) to reduce the computational complexity of the classifier. After that, the colours of the images are then normalised to a floating range between 0 and 1 (originally, the images have an integer value range of 0 to 255).

At this point, the preprocessing of the dataset takes two different paths to study the impact of colour on the classifier: in the first path, we proceed with the colour images, while in the second path, we work with the greyscale images.

In the colour image path, a final processing is carried out to change the colour map from RGB to YUV to separate the luminosity from the colour components. And, in the greyscale image path, a histogram equalisation process is performed.

Finally, before entering the classifier, the images of the dataset are grouped in batches of size 32.

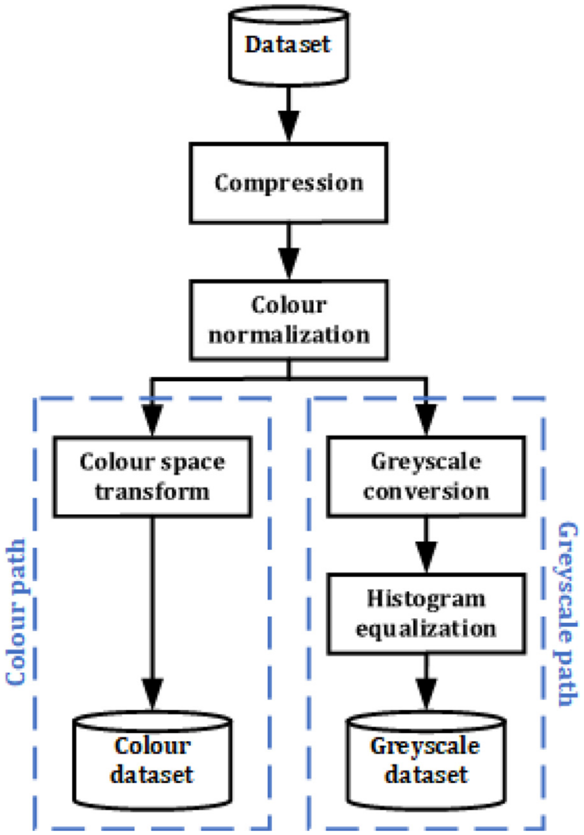


Fig. 3. Preprocessing chain.

2.3. Classifiers

In this work, three classifiers are implemented: the first one for the color image path and 3-class classification, the second one for the greyscale image path and the 3-class classification, and the last one for the image path with a 2-class classification (cancer or not cancer). The results provided by the colour and greyscale dataset are studied because the immunhistochemistry techniques enhance the colour of the samples; and it is also the purpose of this study to compare the difference that the inclusion of these techniques makes to the classification results.

Both 3-class classifiers use the same custom convolutional neural network (CNN) model, which consists of three convolution layers with a maxpolling clustering layer after each. Thus, each pair of layers (convolution + maxpooling) achieves a compression of the input image dimension by grouping each subset of pixels into a single one containing the maximum value of all of them. This process helps to avoid overfitting of the classifier and reduces the computational load.

Resolution compression achieved by each layer pair is:

- First pair: from 180x180 to 89x89 pixels.
- Second pair: from 89x89 to 43x43 pixels.
- Third pair: from 43x43 to 20x20 pixels.

After the last Maxpooling layer, a flattening of the input is performed, which means eliminating all dimensions except one to perform a final Multi-Layer Perceptron (MLP) classic layer. Therefore, next, a fully connected layer is placed with 128 neurons that are activated by a relu activation function, extracting the most relevant features from the flattened image.

At the end, a final dense layer is included as the output layer with a number of neurons equal to the number of classified classes; this is three output neurons (benign tissue, adenocarcinoma and squamous carcinoma) with numerical references 0, 1 and 2, respectively.

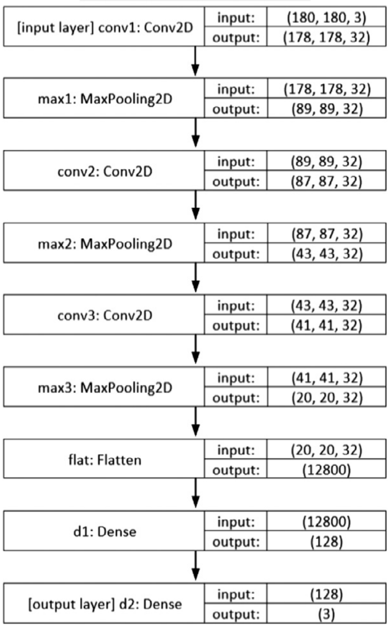

[input layer] conv1: Conv2D input: (180, 180, 3) output: (178, 178, 32)
max1: MaxPooling2D input: (178, 178, 32) output: (89, 89, 32)
conv2: Conv2D input: (89, 89, 32) output: (87, 87, 32)
max2: MaxPooling2D input: (87, 87, 32) output: (43, 43, 32)
conv3: Conv2D input: (43, 43, 32) output: (41, 41, 32)
max3: MaxPooling2D input: (41, 41, 32) output: (20, 20, 32)
flat: Flatten input: (20, 20, 32) output: (12800)
d1: Dense input: (12800) output: (128)
[output layer] d2: Dense input: (128) output: (3)

Fig. 4. Convolutional Neural Network architecture.

So, summarizing the CNN structure used, each layer is detailed next:

- Input layer: bi-dimensional convolutional layer that receives a RGB 180x180 pixels image and applies a 3x3 convolution matrix, eliminating the image borders.
- 1st hidden layer: bi-dimensional maxpooling layer that reduce the input layer to half rows and columns, keeping the maximum value of each 2x2 square.
- 2nd hidden layer: bi-dimensional convolutional layer that works in the same way that the first one but, for this case, it receives 89x89 pixels images.
- 3rd hidden layer: bi-dimensional maxpooling layer that works in the same way that the previous one.
- 4th hidden layer: bi-dimensional convolutional layer that works in the same way that the previous ones but, for this case, it receives 43x43 pixels images.
- 5th hidden layer: bi-dimensional maxpooling layer that works in the same way that the previous ones.
- 6th hidden layer: flatten operation that transform the bi-dimensional information in an unique row that contains the information of all the pixels.
- 7th hidden layer: dense layer that contains 128 simple neurons and receives 12,800 connections from the previous layer.
- Output layer: dense layer that contains 3 simple neurons (for the three classification classes of the system) and receives the connections from the previous 128-neuron layer.

A global network schema for the 3-class classifiers can be found in Fig. 4. Finally, the 2-class classifier uses a similar network model, only changing the last layer (2 output cells instead of 3).

2.4. Evaluation metrics

To evaluate the effectiveness of the classification systems, it is common to use different and well-known metrics: accuracy (most-used metric), sensitivity (also known as recall), specificity, precision and $F1_{score}$ [29].

Table 2
Values used for the different hyperparameters of the developed classifier.

Hyperparameter	Values evaluated
Learning Rate	1e-4, 1e-3, 1e-2
Batch Size	16, 24, 32
Training Epochs	5, 10, 50, 200

To apply them, the classification results obtained for each class must be tagged individually as ”True Positive” (TP: belonging to a class and classified as the same class), ”True Negative” (TN: belonging to another class and classified as that class), ”False Positive” (FP: belonging to another class and classified to the evaluated class) or ”False Negative” (FN: belonging to the class and classified as other class). According to them, the high-level metrics are presented in the next equations:

$$Accuracy = \sum_{c} \frac{TP_c + TN_c}{TP_c + FP_c + TN_c + FN_c}, c \in classes \tag{1}$$

$$Specificity = \sum_{c} \frac{TN_c}{TN_c + FP_c}, c \in classes \tag{2}$$

$$Precision = \sum_{c} \frac{TP_c}{TP_c + FP_c}, c \in classes \tag{3}$$

$$Sensitivity = \sum_{c} \frac{TP_c}{TP_c + FN_c}, c \in classes \tag{4}$$

$$F1_{score} = 2 * \frac{precision * sensitivity}{precision + sensitivity}. \tag{5}$$

About those metrics:

- Accuracy: all samples classified correctly compared to all samples (see Eq. (1)).
- Specificity: proportion of “true negative” values in all cases that do not belong to this class (see Eq. (2)).
- Precision: proportion of “true positive” values in all cases that have been classified as it (see Eq. (3)).
- Sensitivity (or Recall): proportion of “true positive” values in all cases that belong to this class (see Eq. (4)).
- $F1_{score}$: It considers both the precision and the sensitivity (recall) of the test to compute the score. It is the harmonic mean of both parameters (see Eq. (5)).

There are other commonly used metrics, but not all works use them. However, the ROC curve (Receiver Operating Characteristic) [30] is of particular interest in diagnostic systems, because it is the visual representation of the True Positives Rate (TPR) versus the False Positives Rate (FPR) as the discrimination threshold is varied. Usually, when using the ROC curve, the area under the curve (AUC) is used as a value of the system’s goodness-of-fit.

Finally, it is important to take into account the values of the hyperparameters used in the evaluation of the classifier developed for this work. These are described in Table 2.

2.5. Classifier comparison

Many works have been performed in the last years regarding classification systems for lung cancer detection, as it is the most lethal type of cancer.

Among all of them, three types can be distinguished: those that classify between benign and malignant tumours; those that differentiate the type of cancer once it is known to be a malignant tumour; and those that detect whether it is a cancerous tumour or not, and indicate the type of cancer if it is.

In this work, once the classifier has been developed and evaluated, a comparison is made with previous work. For this purpose, the main searching platforms have been used (IEEExplore, Scopus, and Google Scholar), using the keywords *Lung Cancer* and *Deep Learning*. However, to make a realistic comparison, the founded works have been filtered using the next criteria:

- It must classify between benign or malignant tumours (with or without cancer type differentiation).
- They must use histopathological images. Other works use X-Ray lung imaging, but these will not be taken into account to make a comparison with this work.
- They must implement a classifier based on the application of Artificial Intelligence (AI) techniques.
- They must present the results using metrics similar to those used to evaluate this work (global accuracy and/or AUC)
- They must be published from 2015 onwards.

With these restrictions, the number of works to be compared with this classifier is reduced to 12. The results and a detailed explanation of the comparison are presented in the Results section.

2.6. Postprocessing stage

As mentioned in the introduction section, systems trained using DL techniques are, in short, black boxes that receive an input sample and provide an output resulting from the internal transformations of the network and the coefficients generated after training. However, these coefficients do not follow a path that can be easily deduced and/or understood.

This fact is compounded by the sensitivity of this type of diagnostic aid systems, which usually work with data from patients with potentially very aggressive diseases. It is therefore essential that the health professional has the possibility of accessing the justifications that have led the classifier to give a certain answer.

To this end, various tools and/or mechanisms have been developed in recent years for use in systems based on CNNs, and which allow various aspects related to the network’s decision-making to be appreciated. Looking at the compilation of updated tools and algorithms carried out by [31], the most common mechanisms can be observed depending on the field of application and the use that is to be made of them.

Among all of them, due to the extent of application, the information provided, and the possibilities it offers, the use of the Grad-CAM algorithm for CNN-based systems is very widespread [32], ranking among the visual utilities of xAI based on backpropagation. Moreover, it can be adapted to classification problems (as is the case in this work), visual question answering, and captioning.

The Gradient-weighted Class Activation Mapping (Grad-CAM) uses the gradients of any target concept, flowing into the final convolutional layer, to produce a coarse localization map highlighting the important regions in the image for predicting the concept. Unlike previous approaches, Grad-CAM is applicable to a wide variety of CNN model families: CNNs with fully connected layers, CNNs with structured outputs, CNNs with multimodal input, or reinforcement learning, and needs no architectural changes or retraining. This algorithm is combined with existing fine-grained visualizations to create a high-resolution class-discriminative visualization and apply it to image classification, image captioning, and visual question answering (VQA) models. An example of using this algorithm can be observed in Fig. 5.

Therefore, after the classification carried out by the diagnostic aid system developed in this work, the information provided to the healthcare professional is completed with an explanation of the decision taken based on the use of the Grad-CAM algorithm on the image evaluated. In this way, the healthcare professional can

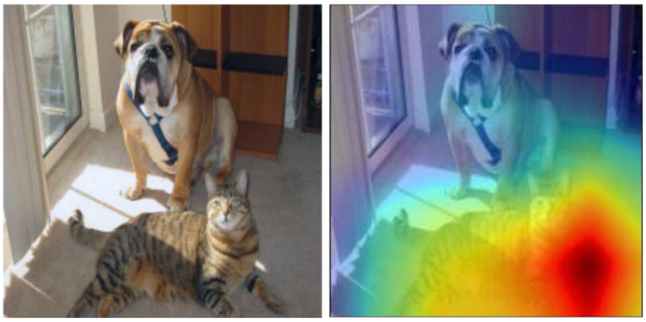


Fig. 5. Grad-CAM algorithm output example: class-discriminative regions of class 'Cat'. Images taken from Selvaraju et al. [32].

observe the regions of the analysed image that have been decisive for the classification; this avoids having to analyse the entire image to verify the results.

In addition to the GradCAM mechanism (classified as a *back-propagation* mechanism), there are other mechanisms widely used in this field. Another type of mechanisms are those based on *perturbations*, which provoke changes in the input and observe the variation of the output. In this order, we find the mechanism proposed by [33], called "Occlusion Sensitivity". In this work, in addition to the application of GradCAM, the *Occlusion Sensitivity* mechanism is also applied to evaluate the robustness of the classifier to input perturbations.

In the Results section, Grad-CAM and Occlusion Sensitivity are applied to the testing dataset images to demonstrate the correct cancer detection and the robustness of the classifier.

3. Results

This section presents the results in a progressive manner, following the same order detailed in the previous section. First, the training and validation results of the classifiers developed in this work (both for two and three classes) will be detailed. Next, a comparison will be made with previous work on lung cancer detection using anatomic pathology images. Finally, some results of the application of the Grad-CAM algorithm over the classifier developed to complete the diagnostic aid system will be shown.

3.1. 3-Class classifier

Implementation and training is carried out in Google Colab with the Tensorflow library. The learning rate is set to 0.001 and the loss function used is Categorical Cross-Entropy (common in multiclass classification problems). The classified classes are for the 3-class system: Benign (BNG), Adenocarcinoma (ADE), and Squamous Cell Carcinoma (SCC).

With 80% of the initial dataset, 5, 10, 50, and 200 epochs are trained and a GPU is used to accelerate the training (reducing the average time from 3.8 seconds on average per epoch with CPU to 0.044 seconds on average per epoch with GPU, obtaining an acceleration of more than 90).

To obtain more accurate results, the systems have been tested using Bootstrap Sampling [34], dividing each dataset initially into ten subsets. After that, eight of them are used for training and two for testing, changing this distribution after each evaluation process. Therefore, each training and testing processes (for 5, 10, 50, and 200 epochs) were repeated 90 times (variations without repetition of 10 subsets taken 2 by 2; applying $V_{m,n} = m \times (m-1) \times (m-2) \ldots (m-n+1)$, where m = 10 and n = 2). The results presented in each subsection are therefore the result of the arithmetic mean of the 90 tests performed for each case, including the standard deviation obtained.

Table 3
Results obtained for the testing subset after training the colour CNN for the 3-class classification during 5 epochs.

Class	Accuracy	Precision	Sensitivity	$F1_{score}$
BNG	0.996 ± 0.020	0.994 ± 0.021	0.994 ± 0.019	0.994 ± 0.021
ADE	0.912 ± 0.034	0.807 ± 0.033	0.981 ± 0.032	0.886 ± 0.038
SCC	0.917 ± 0.023	0.982 ± 0.025	0.759 ± 0.023	0.857 ± 0.025

Table 4
Results obtained for the testing subset after training the colour CNN for the 3-class classification during 10 epochs.

Class	Accuracy	Precision	Sensitivity	$F1_{score}$
BNG	0.998 ± 0.018	0.998 ± 0.019	0.997 ± 0.018	0.998 ± 0.019
ADE	0.944 ± 0.022	0.974 ± 0.021	0.862 ± 0.021	0.915 ± 0.024
SCC	0.948 ± 0.022	0.871 ± 0.026	0.978 ± 0.024	0.922 ± 0.027

Table 5
Results obtained for the testing subset after training the colour CNN for the 3-class classification during 50 epochs.

Class	Accuracy	Precision	Sensitivity	$F1_{score}$
BNG	0.999 ± 0.009	1.0 ± 0.010	0.997 ± 0.010	0.999 ± 0.013
ADE	0.971 ± 0.015	0.954 ± 0.012	0.963 ± 0.013	0.958 ± 0.015
SCC	0.972 ± 0.012	0.961 ± 0.011	0.954 ± 0.011	0.957 ± 0.015

Table 6
Results obtained for the testing subset after training the colour CNN for the 3-class classification during 200 epochs.

Class	Accuracy	Precision	Sensitivity	$F1_{score}$
BNG	0.996 ± 0.018	0.997 ± 0.018	0.997 ± 0.019	0.994 ± 0.019
ADE	0.957 ± 0.027	0.939 ± 0.022	0.939 ± 0.025	0.938 ± 0.023
SCC	0.961 ± 0.025	0.938 ± 0.023	0.943 ± 0.024	0.941 ± 0.024

3.1.1. Colour CNN

In the first case, using the colour CNN with 3 classes and 5 epochs, an overall accuracy of 91.25% was obtained. The metrics results are detailed in Table 3. The confusion matrix obtained is shown in Fig. 6-top-left.

In the second case, using the colour CNN with 3 classes and 10 epochs, an overall accuracy of 94.43% was obtained. The metrics results are detailed in Table 4. The confusion matrix obtained is shown in Fig. 6-top-right.

In the third case, using the colour CNN with 3 classes and 50 epochs, an overall accuracy of 97.11% was obtained. The metrics results are detailed in Table 5. The confusion matrix obtained is shown in Fig. 6-bottom-left.

Finally, for the colour CNN model, using 3 classes and 200 epochs, an overall accuracy of 95.75% was obtained. The metrics results are detailed in Table 6. The confusion matrix obtained is shown in Fig. 6-bottom-right.

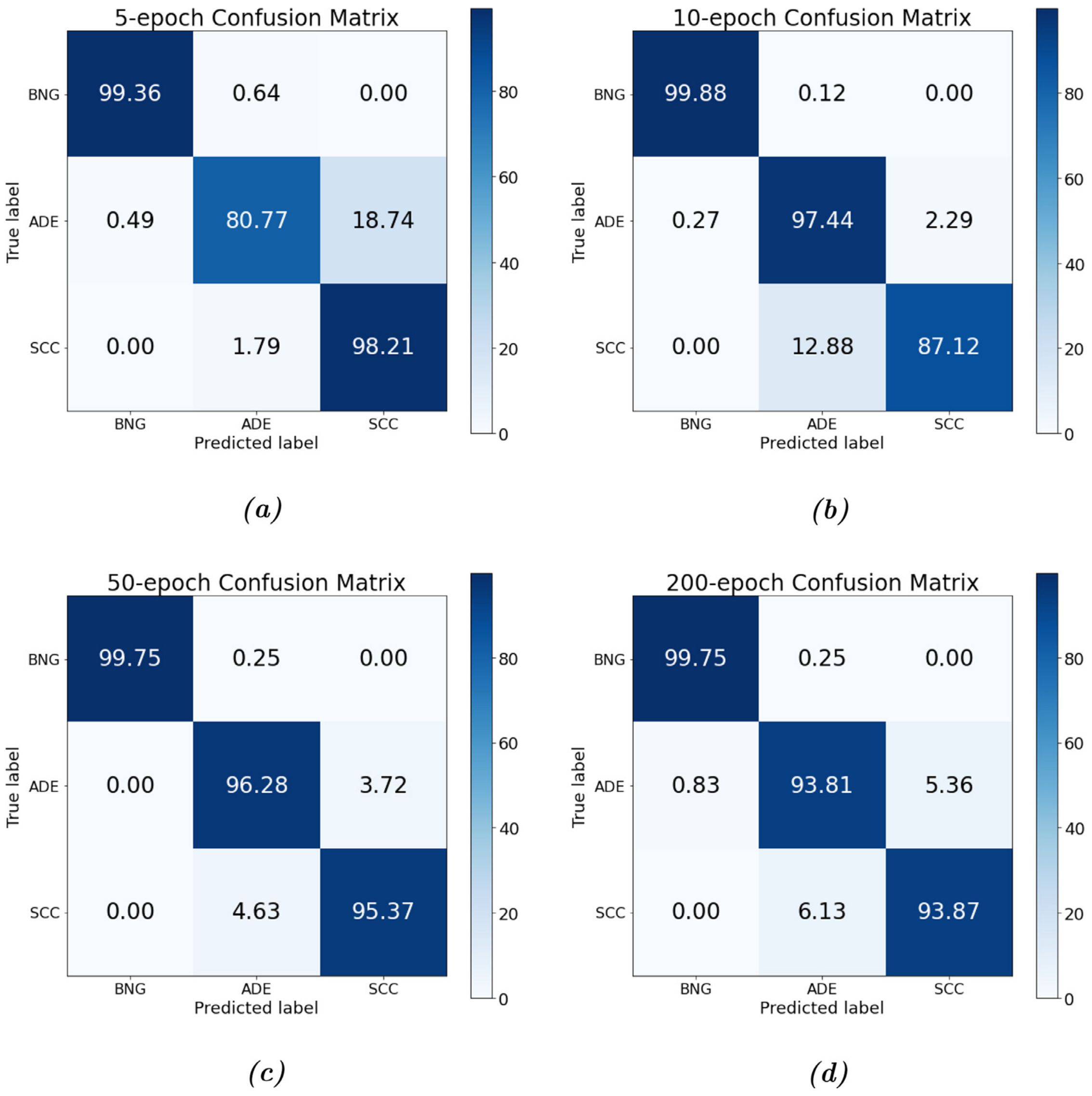
3.1.2. Greyscale CNN

The same process is repeated with the greyscale classifier, evaluating 5, 10, 50, and 200 epoch training.

In the first case, using the greyscale CNN with 3 classes and 5 epochs, an overall accuracy of 91.46% was obtained. The metrics results are detailed in Table 3, and the confusion matrix obtained is shown in Fig. 7-top-left.

In the second case, using the greyscale CNN with 3 classes and 10 epochs, an overall accuracy of 92.73% was obtained. The metrics results are detailed in Table 8, and the confusion matrix obtained is shown in Fig. 7-top-right.

In the third case, using the greyscale CNN with 3 classes and 50 epochs, an overall accuracy of 94.01% was obtained (the best accuracy of all cases in greyscale CNN). The metrics results are de-



(a) *(b)* *(c)* *(d)*

Fig. 6. Confusion Matrixes for the colour CNN with 3 classes: (a) 5 epoch training, (b) 10 epoch training, (c) 50 epoch training, (d) 200 epoch training.

Table 7
Results obtained for the testing subset after training the greyscale CNN for the 3-class classification during 5 epochs.

Class	Accuracy	Precision	Sensitivity	**F1**$_{score}$
BNG	0.971 ± 0.023	0.981 ± 0.022	0.997 ± 0.020	0.954 ± 0.024
ADE	0.915 ± 0.038	0.891 ± 0.035	0.929 ± 0.039	0.875 ± 0.039
SCC	0.942 ± 0.041	0.878 ± 0.040	0.957 ± 0.036	0.916 ± 0.040

Table 8
Results obtained for the testing subset after training the greyscale CNN for the 3-class classification during 10 epochs.

Class	Accuracy	Precision	Sensitivity	**F1**$_{score}$
BNG	0.982 ± 0.021	0.979 ± 0.021	0.967 ± 0.019	0.973 ± 0.022
ADE	0.928 ± 0.032	0.897 ± 0.033	0.896 ± 0.030	0.896 ± 0.035
SCC	0.943 ± 0.031	0.908 ± 0.029	0.920 ± 0.030	0.914 ± 0.031

Table 9
Results obtained for the testing subset after training the greyscale CNN for the 3-class classification during 50 epochs.

Class	Accuracy	Precision	Sensitivity	F1$_{score}$
BNG	0.984 ± 0.018	0.973 ± 0.017	0.978 ± 0.017	0.976 ± 0.020
ADE	0.942 ± 0.024	0.913 ± 0.025	0.920 ± 0.022	0.916 ± 0.025
SCC	0.953 ± 0.021	0.935 ± 0.019	0.923 ± 0.020	0.929 ± 0.022

Table 10
Results obtained for the testing subset after training the greyscale CNN for the 3-class classification during 200 epochs.

Class	Accuracy	Precision	Sensitivity	F1$_{score}$
BNG	0.975 ± 0.022	0.963 ± 0.021	0.962 ± 0.020	0.962 ± 0.023
ADE	0.929 ± 0.030	0.890 ± 0.031	0.908 ± 0.033	0.899 ± 0.034
SCC	0.948 ± 0.028	0.930 ± 0.030	0.911 ± 0.032	0.921 ± 0.032

tailed in Table 9, and the confusion matrix obtained is shown in Fig. 7-bottom-left.

Finally, for the greyscale CNN model, using 3 classes and 200 epochs, an overall accuracy of 92.69% was obtained. The metrics results are detailed in Table 10, and the confusion matrix obtained is shown in Fig. 7-bottom-right.

The full comparison between the colour and greyscale CNNs is detailed in Table 11, and the best results for each metric are highlighted in blue.

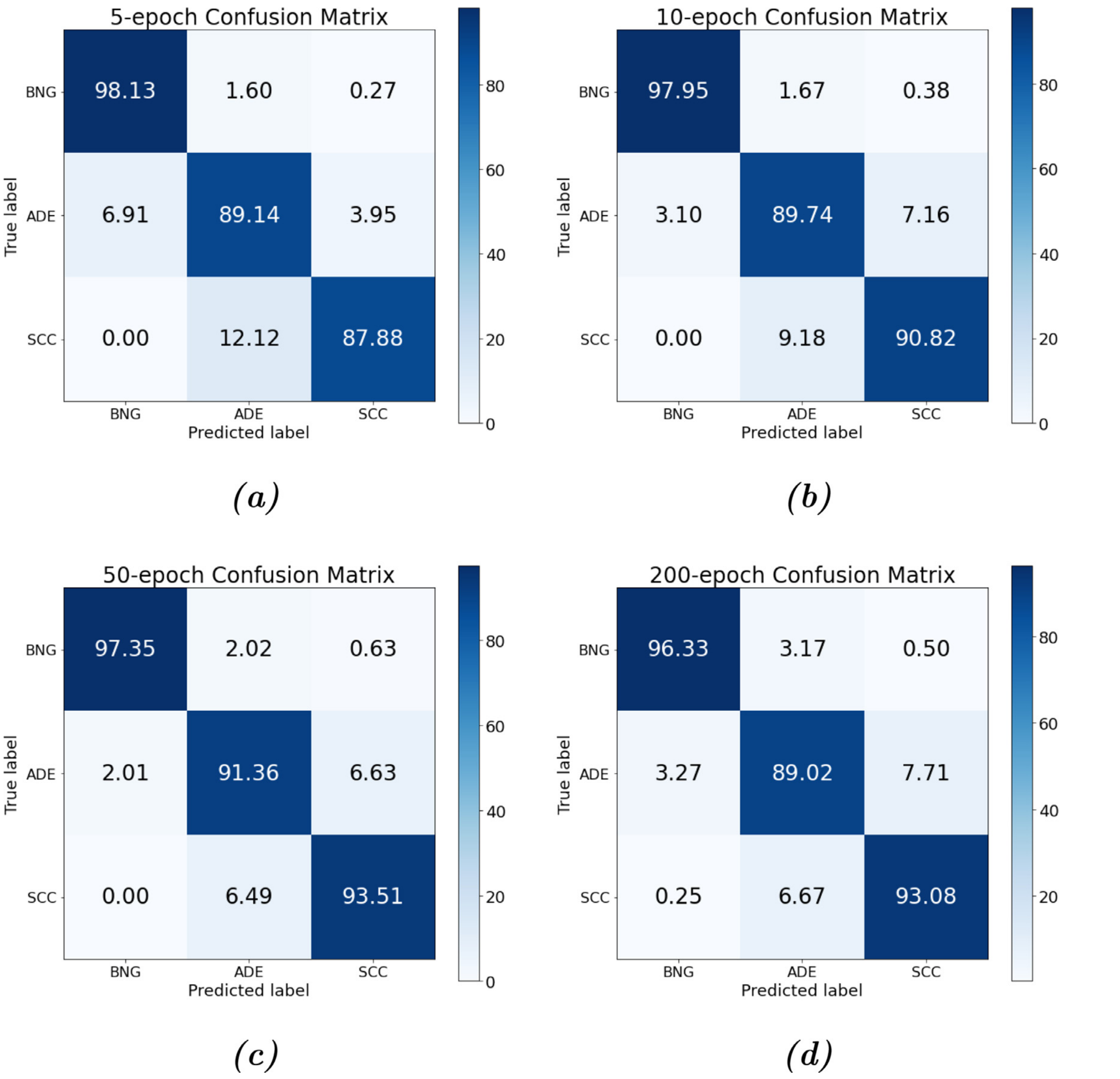


Fig. 7. Confusion Matrixes for the greyscale CNN with 3 classes: (a) 5 epoch training, (b) 10 epoch training, (c) 50 epoch training, (d) 200 epoch training.

Table 11
Comparison between the Colour CNN and the Greyscale CNN results for the 3-class classification.

		Colour CNN				Greyscale CNN			
Class	Metric	*5-epoch*	*10-epoch*	*50-epoch*	*200-epoch*	*5-epoch*	*10-epoch*	*50-epoch*	*200-epoch*
BGN	*Accuracy*	0.998 ± 0.020	0.998 ± 0.018	**0.999 ± 0.009**	0.996 ± 0.018	0.971 ± 0.023	0.982 ± 0.021	0.984 ± 0.018	0.975 ± 0.022
	Precision	0.994 ± 0.021	0.998 ± 0.019	**1.0 ± 0.010**	0.997 ± 0.018	0.981 ± 0.022	0.979 ± 0.021	0.973 ± 0.017	0.963 ± 0.021
	Sensitivity	0.994 ± 0.019	**0.997 ± 0.018**	**0.997 ± 0.010**	**0.997 ± 0.019**	0.997 ± 0.020	0.967 ± 0.019	0.978 ± 0.017	0.962 ± 0.020
	$F1_{score}$	0.994 ± 0.021	0.998 ± 0.019	**0.999 ± 0.013**	0.994 ± 0.019	0.954 ± 0.024	0.973 ± 0.022	0.976 ± 0.020	0.962 ± 0.023
ADE	*Accuracy*	0.912 ± 0.034	0.944 ± 0.022	**0.971 ± 0.015**	0.957 ± 0.027	0.915 ± 0.038	0.928 ± 0.032	0.942 ± 0.024	0.929 ± 0.030
	Precision	0.807 ± 0.033	**0.974 ± 0.021**	0.954 ± 0.012	0.939 ± 0.022	0.891 ± 0.035	0.897 ± 0.033	0.913 ± 0.025	0.890 ± 0.031
	Sensitivity	0.981 ± 0.032	0.862 ± 0.021	**0.963 ± 0.013**	0.939 ± 0.025	0.929 ± 0.039	0.896 ± 0.030	0.920 ± 0.022	0.908 ± 0.033
	$F1_{score}$	0.886 ± 0.038	0.915 ± 0.024	**0.958 ± 0.015**	0.938 ± 0.023	0.875 ± 0.039	0.896 ± 0.035	0.916 ± 0.025	0.899 ± 0.034
SCC	*Accuracy*	0.917 ± 0.023	0.948 ± 0.022	**0.972 ± 0.012**	0.961 ± 0.025	0.942 ± 0.041	0.943 ± 0.031	0.953 ± 0.021	0.948 ± 0.028
	Precision	**0.982 ± 0.025**	0.871 ± 0.026	0.961 ± 0.011	0.938 ± 0.023	0.878 ± 0.040	0.908 ± 0.029	0.935 ± 0.019	0.930 ± 0.030
	Sensitivity	0.759 ± 0.023	**0.978 ± 0.024**	0.954 ± 0.011	0.943 ± 0.024	0.957 ± 0.036	0.920 ± 0.030	0.923 ± 0.020	0.911 ± 0.032
	$F1_{score}$	0.857 ± 0.025	0.922 ± 0.027	**0.957 ± 0.015**	0.941 ± 0.024	0.916 ± 0.040	0.914 ± 0.031	0.929 ± 0.022	0.921 ± 0.032
Global	*Accuracy*	91.25%	94.43%	**97.11%**	95.75%	91.46%	92.73%	94.01%	92.69%
	*Cancer FN**	0.30%	0.12%	**0%**	0.42%	3.34%	1.58%	1.04%	1.83%

*Cancer FN: False Negatives from any of the cancer classes (ADE or SCC), classified as benign tissue. These cases are of special interest as they are the ones that can potentially cause the most problems in the long term (most dangerous).

Table 12
Results obtained for the testing subset after training the colour CNN for the 2-class classification during 5 epochs.

Class	Accuracy	Precision	Sensitivity	$\mathbf{F1}_{score}$
BNG	0.994 ± 0.040	0.993 ± 0.039	0.993 ± 0.014	0.993 ± 0.028
MLG	0.994 ± 0.015	0.994 ± 0.015	0.994 ± 0.040	0.994 ± 0.028

Table 13
Results obtained for the testing subset after training the colour CNN for the 2-class classification during 10 epochs.

Class	Accuracy	Precision	Sensitivity	$\mathbf{F1}_{score}$
BNG	0.997 ± 0.011	0.993 ± 0.024	1 ± 0.005	0.997 ± 0.014
MLG	0.997 ± 0.003	1 ± 0.005	0.994 ± 0.025	0.997 ± 0.014

Table 14
Results obtained for the testing subset after training the colour CNN for the 2-class classification during 50 epochs.

Class	Accuracy	Precision	Sensitivity	$\mathbf{F1}_{score}$
BNG	0.997 ± 0.008	0.996 ± 0.009	0.997 ± 0.005	0.997 ± 0.007
MLG	0.997 ± 0.004	0.997 ± 0.005	0.996 ± 0.008	0.997 ± 0.006

Table 15
Results obtained for the testing subset after training the colour CNN for the 2-class classification during 200 epochs.

Class	Accuracy	Precision	Sensitivity	$\mathbf{F1}_{score}$
BNG	0.993 ± 0.026	0.995 ± 0.030	0.991 ± 0.027	0.993 ± 0.028
MLG	0.993 ± 0.026	0.991 ± 0.025	0.995 ± 0.028	0.993 ± 0.027

3.2. 2-class classifier

Following the same procedure and structure than the 3-class classifier subsection, this implementation and training are carried out in Google Colab with the Tensorflow library. The learning rate is set to 0.001 and the loss function used is Categorical Cross-Entropy. In this case, there are only 2 classified classes: Benign (BNG) and Malignant (MLG). For that purpose, several GPU trainings with 5, 10, 50, and 200 epochs are performed and the systems are tested using Bootstrap Sampling (explained deeply in the previous subsection). As demonstrated above, the implementation of the color CNN shows slightly better results in classification accuracy and significantly better results in standard deviation (higher precision). Therefore, in this case, the 2-class classifier is only implemented and tested for a Colour CNN.

In the first case, using the colour CNN with 2 classes and 5 epochs, an overall accuracy of 99.38% was obtained. The metrics results are detailed in Table 12, and the confusion matrix obtained is shown in Fig. 8-top-left.

In the second case, using the colour CNN with 2 classes and 10 epochs, an overall accuracy of 99.69% was obtained. The metrics results are detailed in Table 13, and the confusion matrix obtained is shown in Fig. 8-top-right.

In the third case, using the colour CNN with 2 classes and 50 epochs, an overall accuracy of 99.69% was obtained. The metrics results are detailed in Table 14, and the confusion matrix obtained is shown in Fig. 8-bottom-left.

Finally, for the colour CNN model, using 2 classes and 200 epochs, an overall accuracy of 99.31% was obtained. The metrics results are detailed in Table 15, and the confusion matrix obtained is shown in Fig. 8-bottom-right.

The summary of the 2-class Colour CNN classifier is shown in Table 16. In this work, both types of systems have been developed in order to compare the processing performed in previous works, as many of them only distinguish benign from malignant.

At this point, the results of all classifiers developed in this work have been presented. A table summarizing the best cases of every system presented is included (see Table 17). The ROC curve and the AUC (area under curve) regarding the 3-class colour CNN with 50-epoch training can be observed in Fig. 9. As the AUC is a metric defined for 2-class classifiers, it is widespread to use pseudo-ROC curves for multiclass systems, where a ROC curve is represented for each single class (using multiple 2-class classifiers, one per each class, where the class itself is compared with the rest of the classes). Thus, in the figure three ROC curves are represented, and the AUC results show values higher than 97.77% for all the classes.

And, finally, the ROC curve and the AUC regarding the 2-class colour CNN with 50-epoch training can be observed in Fig. 10. In this case, the AUC is 99.75%.

4. Discussion

The previous results will be discussed in detail following the same order as the results were presented. With this discussion, the work developed will be compared with previous works. Finally, the final report provided to the pathologists is presented and discussed.

4.1. 3-class colour classifier

With a 5-epoch training there is a significant drop in the precision of the ADE class (80.7%) as 19% of the samples were erroneously classified as squamous cell carcinoma; this fact also caused a drop in the sensitivity of the SCC class. Increasing the epochs to 10, there is an improvement in the precision and sensitivity metrics (because of the increase of ADE class accuracy); moreover, the cases that are classified as BNG but belong to ADE or CCE (false positive cases of BNG, that are the most dangerous cases) are decreased to half the previous cases.

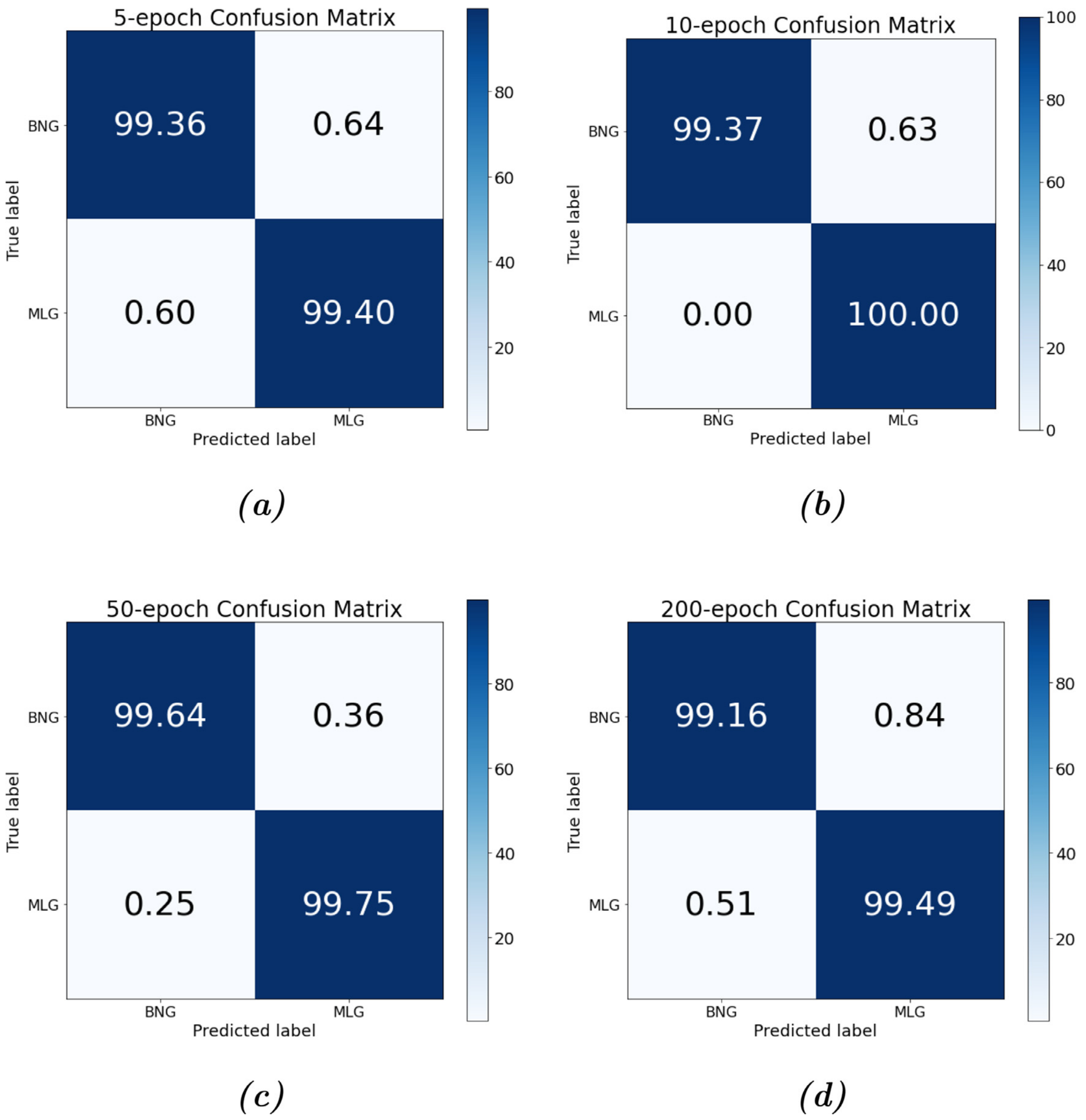
If we continue increasing the epochs to 50, accuracy is improved and it is the first time that all metrics obtain results above 95%. Moreover, the most dangerous cases (cancer classified as BNG) are zero, obtaining a precision value of 1. The substantial increase in accuracy is due to a better distinction between ADE and SCC.

And, finally, with 200 epochs, the system seems to achieve an asymptote, since the metrics worsen slightly in all aspects. The results are not bad at all, improving the overall accuracy compared to the 10-epoch training, but the 50-epoch training obtains better results.

As can be observed, the best results are obtained with the 50-epoch training, with an standard deviation around 1-1.5%. It is important to mention that, in addition to this improvement, there are no false positive cases of benign tissue (the precision value is 1), and those are the most dangerous cases for the diagnose. However, a 0.12% of the cases obtained in the 10-epoch training means 6 cases from 5000 samples, which is a much lower percentage than the average for pathologists. Moreover, the results of the 10-epoch training reflect a higher standard deviation (around 2-2.7%), so it is a less reliable system. Thus, the selected system is the one obtained after the 50-epoch training: it has the best accuracy, the best standard deviation, and zero false positive cases of benign tissue.

4.2. 3-class greyscale classifier

With a 5-epoch training, precision metrics of ADE and SCC classes are low (less than 90%) because of the several mistakes produced between those two classes; and, moreover, in this case there are too many false negatives in cancer cases (around 6% of adenocarcinoma cases are classified as benign tissue), so it is very



(a) *(b)* *(c)* *(d)*

Fig. 8. Confusion Matrixes for the colour CNN with 2 classes: (a) 5 epoch training, (b) 10 epoch training, (c) 50 epoch training, (d) 200 epoch training.

Table 16
Summary of the Colour CNN for the 2-class classifier.

Class	Metric	*5-epoch*	*10-epoch*	*50-epoch*	*200-epoch*
BGN	*Accuracy*	0.994 ± 0.040	0.997 ± 0.011	0.997 ± 0.008	0.993 ± 0.026
	Precision	0.993 ± 0.039	0.993 ± 0.02	0.996 ± 0.009	0.995 ± 0.030
	Sensitivity	0.993 ± 0.014	1 ± 0.005	0.997 ± 0.005	0.991 ± 0.027
	$F1_{score}$	0.993 ± 0.028	0.997 ± 0.014	0.997 ± 0.007	0.993 ± 0.028
MLG	*Accuracy*	0.994 ± 0.015	0.997 ± 0.003	0.997 ± 0.004	0.993 ± 0.026
	Precision	0.994 ± 0.015	1 ± 0.005	0.997 ± 0.005	0.991 ± 0.025
	Sensitivity	0.994 ± 0.040	0.994 ± 0.025	0.996 ± 0.008	0.995 ± 0.028
	$F1_{score}$	0.994 ± 0.028	0.997 ± 0.014	0.997 ± 0.006	0.993 ± 0.027
Global	*Accuracy*	99.38%	99.69%	99.69%	99.31%
	*Cancer FN**	0.60%	0%	0.25%	0.51%

*Cancer FN: False Negatives from MLG class classified as benign tissue.

dangerous. Increasing the training epochs to 10, precision metric is improved slightly in ADE and SCC classes, although it decreases in BNG class; although, the most important improvement is the reduction of false negatives in cancer (around 3% of adenocarcinoma samples are classified as benign tissue, half of the obtained with the 5-epoch training).

If we continue increasing the epochs to 50, the precision metric and $F1_{score}$ are over 90% for all the classes, and the false negative cases for cancer are reduced too (only 2% of the adenocarcinoma cases are classified as benign tissue). And, finally, for the 200-epoch training, there is a worsening of the results, with a slight decrease of all parameters (indeed, false negatives for cancer in-

Table 17
Classifiers summary.

Classifier	Epochs	Accuracy	STD	Cancer FN
3-class Colour CNN	50	0.9711	0.9-1.5%	0%
3-class Greyscale CNN	50	0.9402	1.7-2.5%	1.04%
2-class Colour CNN	50	0.9969	0.4-0.9%	0.25%

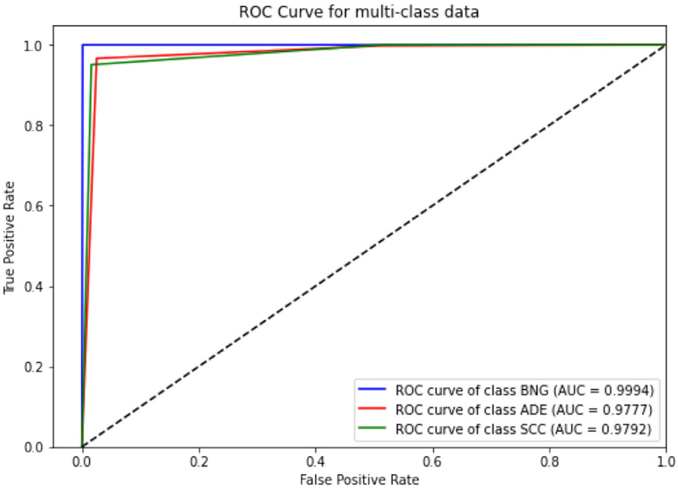


Fig. 9. ROC Curve and AUC for multi-class data giving the 3-class Colour CNN after 50-epoch training.

crease to a 3.27% of the adenocarcinoma samples and a 0.25% of the squamous cell carcinoma cases).

As can be observed, the best results are obtained with the 50-epoch training. However, the false negative cases for cancer are much higher than the one obtained with the colour CNN (2.10% compared to 0.36%, a difference of nearly 6 times). Comparing the overall accuracy, the greyscale path (with its best training) obtains 94.01%, and the colour path obtains 96.49%. The difference is not too high, compared to the reduction in computational load due to working with a single color component; but the substantial increase in the false negative cases for cancer is of great concern, so it is preferred to use the colour CNN. Moreover, these diagnostic aid systems do not need to run in a hard real-time environment, so a small delay (seconds) does not affect the result; and it is important to take into account too that the results of the greyscale CNN reflect a higher standard deviation (between 2 and 2.7%) compared to the colour CNN (between 0.9 and 1.5%), so the colour one seems more reliable.

It can be observed that the best options are Color CNN with 10-epoch training and with 50-epoch training. The best global accuracy is obtained with 50 epochs (97.11%), and the best ”Cancer FN” metric is obtained for the 50 epochs training too (0%). Moreover, the 10-epoch option shows more variability in the standard deviation compared to the 50-epoch option: between 1.8 and 2.7% for 10 epochs, and between 0.9 and 1.5% for 50 epochs (about half). Therefore, the best option for the 3-class classified among those studied is the Colour CNN with 50-epoch training.

Accuracy metric evolution during the increasing of the training epochs (with its standard deviation) for Colour CNN is represented in Fig. 11 for BNG class, in Fig. 12 for ADE class, and in Fig. 13 for SCC class. As can be observed, the cases of ADE and SCC classes are very clear, as the best absolute results and the minor standard deviation are obtained for the 50-epoch training. However, for BGN class, the results obtained for 10 and 50 epochs are the same, with the main difference that the minor standard deviation is obtained with 50 epochs.

4.3. 2-class classifier

With a 5-epoch training, the absolute values of the metrics analyzed are acceptable, but there are two issues than may be taken into account: the false negative cases for cancer are the worst among all trainings performed (0.6%), and also the standard de-

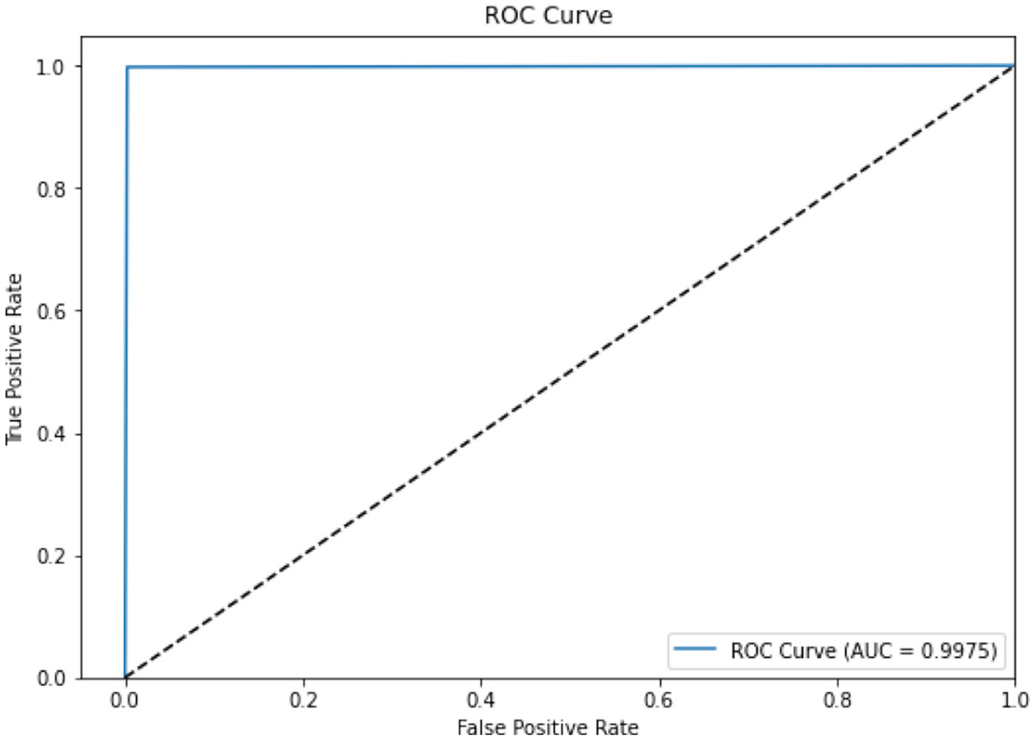


Fig. 10. ROC Curve and AUC giving the 2-class Colour CNN after 50-epoch training.

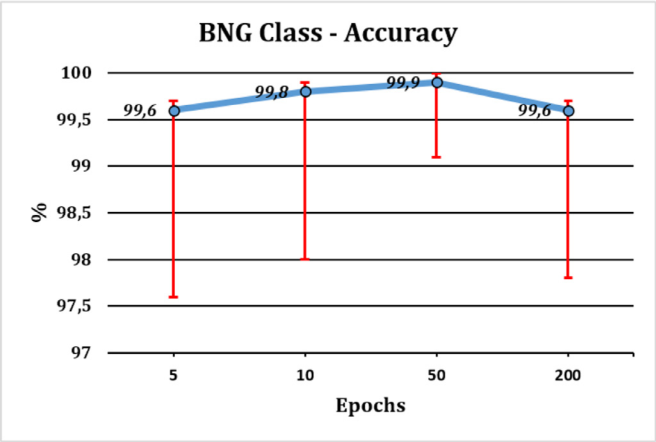


Fig. 11. Accuracy evolution (absolute value and standard deviation) for class BNG of Colour CNN.

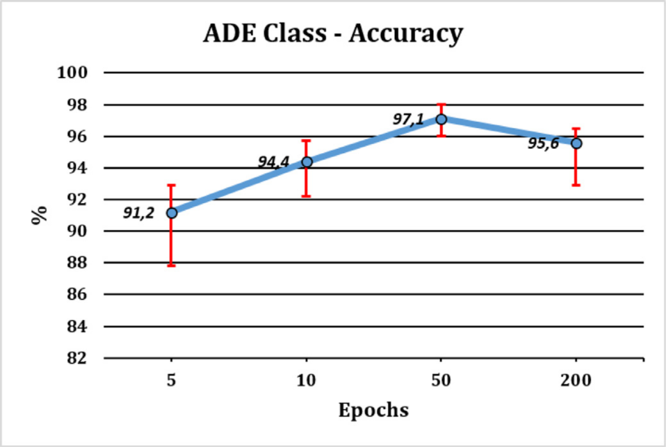


Fig. 12. Accuracy evolution (absolute value and standard deviation) for class ADE of Colour CNN.

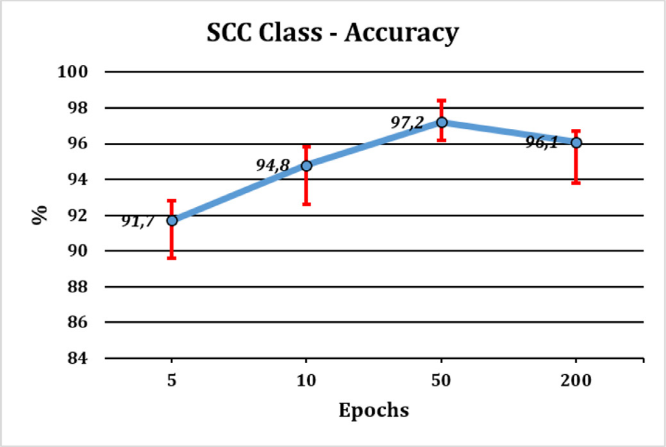


Fig. 13. Accuracy evolution (absolute value and standard deviation) for class SCC of Colour CNN.

viation is too high (between 1.4 and 4.0%); these values cause the system not to be adequate to solve the problem detailed in this work. Increasing the training epochs to 10, an important goal is achieved: there is no false negative case of cancer among the testing dataset, so the most dangerous cases are avoided; although the standard deviation of some metrics is not as good as desired, with some cases over 2%.

If we continue increasing the epochs to 50, the best accuracy result is achieved (the same as the one obtained with the 10-epoch training), but there is an increase of false negative cases for cancer (0.25%). However, the standard deviation has been significantly reduced in this case, to values between 0.4 and 0.9% (less than half of what was obtained in the previous case). And, finally, for the 200-epoch training, not only the accuracy is reduced: precision and sensitivity decreased too, provoking an increase in the false negatives of both classes.

As can be observed, the best accuracy results are obtained with the 10-epoch and 50-epoch trainings. In the 50-epoch training, the false positive cases of benign tissue are not zero (as happened with the 10-epoch training). However, the standard deviation obtained for the 50-epoch training is half of the presented after the 10-epoch training (0.4-0.9% versus 0.3-2.4%). Thus, the 50-epoch training is the most reliable option.

As could be observed with the 3-class systems, the class with the best accuracy results is that of benign tissue, since most of the failures were caused by distinguishing the type of cancer. In the 2-class case, by not having to make such a distinction, the results benefit.

In terms of use as a diagnostic aid, 3-class systems provide more information to the pathologist and allow him/her to tailor treatment to the type of cancer. In addition, if cancer is detected in a 2-class system, the pathologist would have to perform a second test to distinguish the type, which would not benefit him as much as it should. However, 2-class systems allow mass screening mainly for benign tissue cases.

4.4. Works comparison

Following the guidelines indicated in the methods section, a deep search of lung cancer detection systems using histopathological images has been done. It is important to note that, within the range of years used for the search (2015 to 2021), the publicly available works that meet all the requirements indicated previously are concentrated in recent years (2018 to 2021). There are some works related to lung cancer detection in the first years, but most of them use different image types, classifiers, and/or metrics to evaluate. In that case, it would be difficult to compare this work with them. Moreover, the works with the best classification results are found in the last years.

Therefore, taking into account the restrictions provided and the above explanation, twelve works have been selected: four published in 2018, four published in 2019, three published in 2020, and one published in 2021.

The summary of the selected work with its main attributes and results are shown in Table 18. In this table, it can be seen that the selected papers use convolutional neural networks in their classifier (some even include other types of additional classifiers). However, the main differences are centered on the classes detected by each classifier. As, in this work, two different classifiers are developed and evaluated (2-class classifier and 3-class classifier), the comparison will be divided in two parts: on the one hand, works with 2-class classifiers will be compared; and, on the other hand, works that classify more than 2 classes will be compared.

In addition to the characteristics of the classifiers, it is important to look at the datasets used by previous works in order to understand the choice made in our work. Briefly, our work uses a dataset of lung histological images, and only three of these previous works also use histological images (most of them use computer tomography images). Therefore, for our case, we could use the datasets from these three works only; however, these datasets are not publicly available and that is the reason why we had to work with the dataset described above. Even so, the differences can be seen in Table 19.

Table 18
List of works published in recent years on neural network-based classifiers for lung cancer detection using histopathological images. The evaluation metrics presented are related to the test subset results (not the training one).

Work	Classifier	Classes	Evaluation Results
Li et al. [35]	SVM, CNN (AlexNet,VGG16, ResNet50, SqueexeNet)	4: BNG, ADE, SCC, SCLC	AUC (SVM): 71.69% AUC (CNN): 91.19%
Wang et al. [36]	CNN	2: BNG, MLG	Accuracy: 89.8%
Coudray et al. [37]	CNN (InceptionV3)	3: BNG, ADE, SCC	AUC (ADE): 91.9% AUC (SCC): 97.7%
Wang et al. [38]	CNN (VGG16, ScanNet)	2: BNG, MLG	Accuracy: 53–97.1%
Bilaloglu et al. [39]	CNN (InceptionV3)	3: BNG, ADE, SCC	AUC (BNG): 99.85% AUC (ADE): 93.31% AUC (SCC): 93.24% Precision: 95.65 Sensitivity: 88.12
Noorbakhsh et al. [40]	CNN (InceptionV3)	2: BNG, MLG	AUC: 98% Accuracy: 91% Precision: 97% Sensitivity: 90% Specificity: 86%
Sha et al. [41]	CNN (ResNet18)	2: ADE, SCC 2: BNG, MLG	AUC (ADE, SCC): 83% AUC (BNG, MLG): 80%
Wang et al. [42]	CNN (VGG16, ScanNet) + RF	4: BNG, ADE, SCC, SCLC	AUC: 85.6% Accuracy: 82%
Kanavati et al. [43]	CNN (EfficientNet-B3)	2: BNG, MLG	AUC: 97.4–98.5%
Yu et al. [44]	CNN (AlexNet, GoogleNet, VGG16, ResNet50)	2: BNG, ADE 2: BNG, SCC	AUC (BNG, ADE): 97.1% AUC (BNG, SCC): 98.5%
Kriegsmann et al. [45]	CNN (InceptionV3, VGG16, InceptionResNetV2)	3: ADE, SCC, SCLC	Accuracy: 60–89%
Guo et al. [46]	CNN (ProNet, RadNet)	3: ADE, SCC, SCLC	AUC: 78.9-84% Accuracy: 71.6–74.7% $F1_{score}$: 72.2–73.2%
This Work (2021)	CNN (Custom)	2: BNG, MLG 3: BNG, ADE, SCC	AUC (BNG, MLG): 99.75% AUC (BNG, ADE, SCC): 99.94, 97.77, 97.92% Accuracy (2, 3 classes): 99.69, 97.11% Precision (2, 3): 99.69, 97.15% Sensitivity (2, 3): 99.69, 97.13% $F1_{score}$ (2, 3): 99.69, 97.14%

* CNN: Convolutional Neural Network, SVM: Support Vector Machine, RF: Random Forest, SCLC: Small-Cell Lung Cancer.

Table 19
Dataset comparison with previous works.

Work	Dataset	Publicly available	Image type
Li et al. [35]	Own	No	Histology
Wang et al. [36]	NLST	Yes	CT
Coudray et al. [37]	TCGA	Yes	CT
Wang et al. [38]	Own and TCGA	No/Yes	CT
Bilaloglu et al. [39]	TCGA	Yes	CT
Noorbakhsh et al. [40]	TCGA	Yes	CT
Sha et al. [41]	Own	No	Histology
Wang et al. [42]	Own and TCGA	No/Yes	CT
Kanavati et al. [43]	TCGA	Yes	CT
Yu et al. [44]	TCGA	Yes	CT
Kriegsmann et al. [45]	Own	No	Histology
Guo et al. [46]	Own	No	CT
This Work	LC25000	Yes	Histology

In Table 19, it can be observed that most of the works use images from the TCGA (The Cancer Genome Atlas Program) dataset. This dataset consists of more than 18 GigaBytes of chest computed tomography (CT) images. Additionally, one of these works [36] uses the NLST (National Lung Screening Trial) dataset, which has fewer images, but also CT images. Another point to take into account is that those studies that use histopathological images work with their own dataset. That is why, this work used the publicly available dataset LC25000.

4.4.1. Works with a 2-class classifier

Works included in this category are: Wang et al. [36,38], Noorbakhsh et al. [40], Sha et al. [41], Kanavati et al. [43] and Yu et al. [44].

Most of these works present a classifier that distinguishes between benign and malign tissue. However, there are two special cases that need to be evaluated:

- In the work [41], two classifiers are developed. One of them classifies benign and malign tissue, and this one is included in the comparison. However, the other one distinguishes between adenocarcinoma (ADE) and squamous cell carcinoma (SCC); so, this last classifier cannot be compared as it does not detect benign tissue (like ours).
- In the work [44], two classifiers are developed too. Both of them classify between benign tissue and one specific type of cancer: the first one distinguishes benign from adenocarcinoma; and the second one distinguishes benign from squamous cell carcinoma. Both cases are more beneficial because they only include a specific type of cancer and, theoretically, it should be easier to find common features that distinguish cancer from benign tissue. For this case, we will compare our system with the system that provides better results among the two classifiers developed in this specific work. Although this is a detrimental comparison for our work, we will later observe that we obtain better results.

Therefore, a comparison summary for the 2-class classifiers is presented in Table 20. The two metrics included in this table are the most used among the compared works: accuracy (ACC) and the area under ROC curve (AUC).

It can be observed that not every work analyzed uses both metrics to evaluate the system. The most used metric in the last years is AUC, although the works published in 2018 only present the accuracy. As we obtained both values for each classifier, we can compare our work with all of them.

Taking into account the previous works that use the accuracy metric, the most high result is achieved by [38] with a 97.1%. In our case, we achieve a 99.69% accuracy with a custom CNN classifier that has a computational complexity lower than the CNN used by [38] (VGG16). While VGG16 structure includes five maxpooling

Table 20
Comparison of 2-class classifier works using the most common metrics: accuracy (ACC) and area under curve (AUC).

Work	ACC	AUC
[36]	89.8%	
[38]	97.1%	
[40]	91%	98%
[41]		80%
[43]		98.5%
[44]		98.5%
This Work (2021)	99.69%	99.75%

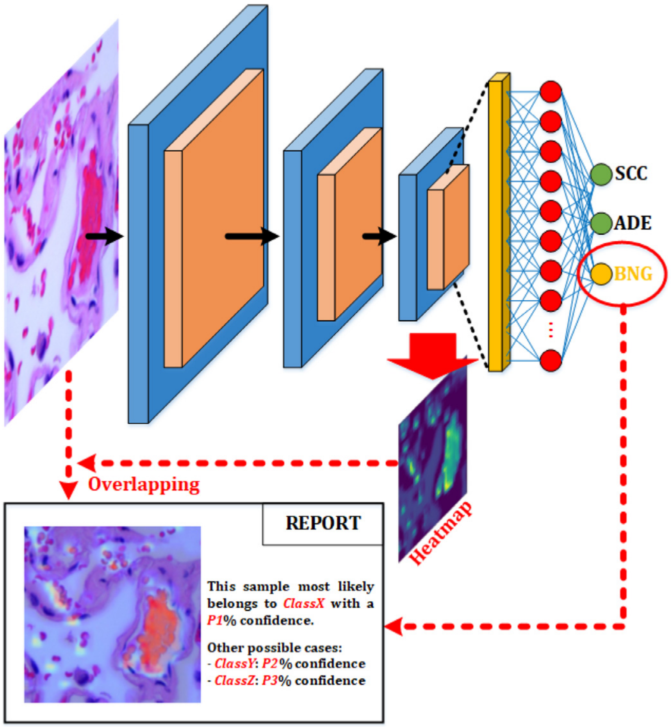


Fig. 14. System's final report given to the healthcare professional.

stages (each one preceded by two or three convolutional layers) and three dense layers, our custom CNN includes three maxpooling stages (preceded only by one convolutional layer) and two dense layers.

Finally, observing the previous works that use the AUC metric, the most high result is achieved by [43] and [44] with a 98.5%. In our case, we achieve a 99.75%. As happened in the previous case, the CNN used in our work is less complex that the CNN used in those two works: [43] uses EfficientNet model (more than 15 convolutional layers); and [44] uses several models (the least complex is AlexNet with four convolutional layers, three maxpooling layers and three dense layers; but this is not the model than obtains the best results).

Summarizing, our 2-class lung cancer classifier based on a custom CNN model achieves the best results among all the works published in recent years regarding lung cancer detection on histopathologycal images with convolutional neural networks.

4.4.2. Works with classifiers for more than 2 classes

Regarding the works that classify more than 2 classes, we can distinguish the next ones: Li et al. [35], Coudray et al. [37], Bilaloglu et al. [39], Wang et al. [42], Kriegsmann et al. [45], and Guo et al. [46]. There are some cases that need to be analyzed in detail:

Table 21
Comparison more than 3 classes classifiers using accuracy (ACC) and area under curve for each class.

Work	ACC	A_{BNG}	A_{ADE}	A_{SCC}
[35]		91.19%		
[37]			91.9%	97.7%
[39]		99.85%	93.31%	93.24%
[42]	82%	85.6%		
This Work (2021)	97.11%	99.94%	97.77%	97.92%

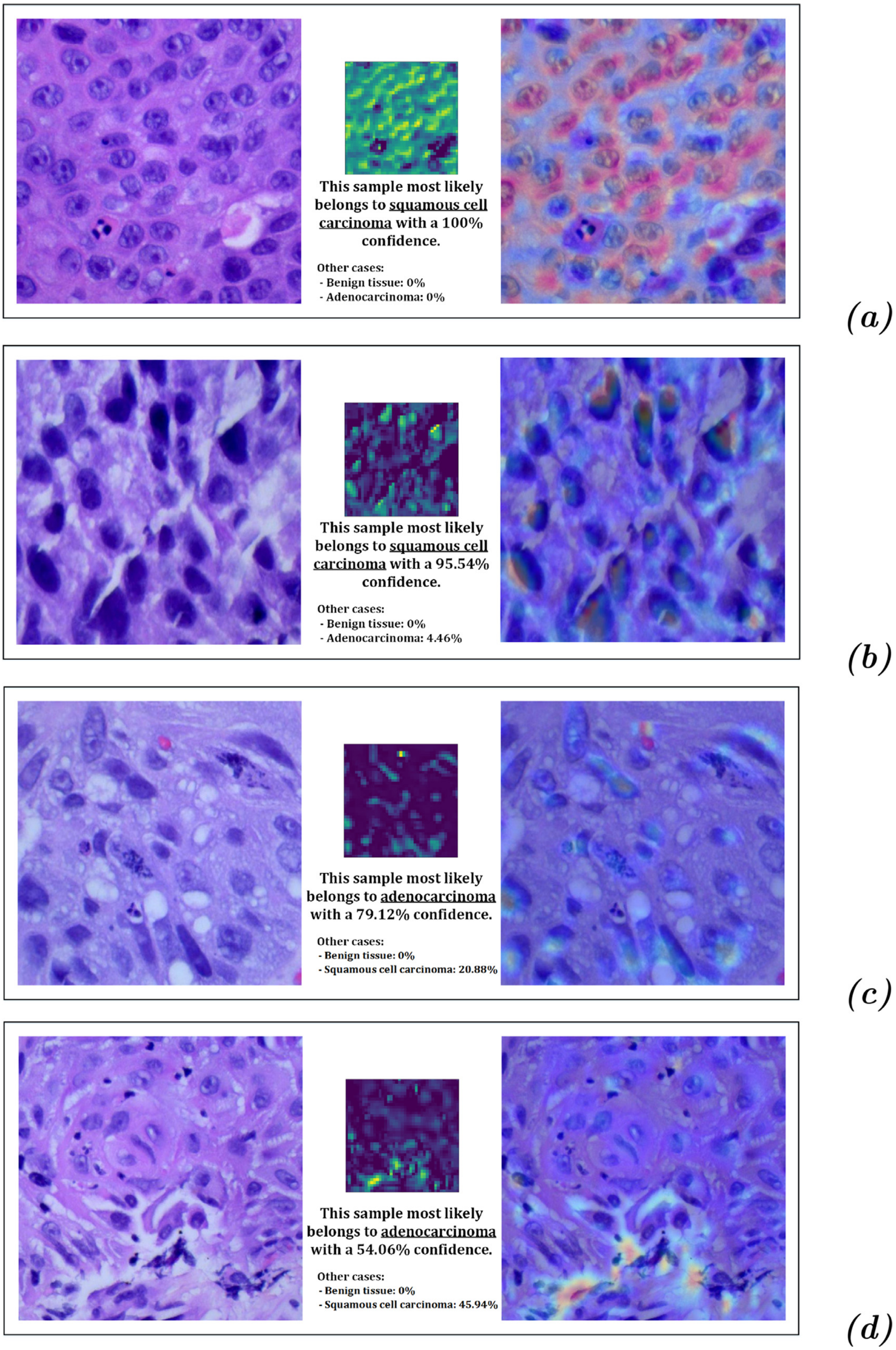
- In the works Kriegsmann et al. [45] and Guo et al. [46], the developed classifiers distinguish between three classes: adenocarcinoma, squamous cell carcinoma, and small-cell lung cancer. Thus, these two works aim to classify the lung cancer type in a previously detected malignant tissue, without distinguishing a benign from a malign sample. As our work includes benign samples, we aim to distinguish benign from malign tissue and, in the same classifier, the type of cancer detected (among non-small cell types). Therefore, the goal of these two systems is different than ours; so it is the complexity of the classifier developed. That is why these two works will not be taken into account in this specific comparison. However, looking at the results, both systems obtain worse classification results than ours: [45] obtains 89% accuracy (versus 97.11% obtained by our system); and [46] obtains a 74.7% accuracy and an AUC of 84% (versus 97.77–99.94% obtained by our system).
- The classifiers of the works [35] and [42] include small-cell lung cancer class (SCLC), classifying between four classes (one more than our system). Thus, although both are included in this specific comparison, this fact must be taken into account (as we suppose that, if they do not include the fourth class, the results can be slightly improved).

Therefore, a comparison summary for the classifiers than distinguish between three or four classes is presented in Table 21. The two metrics included in this table are the most used among the compared works: accuracy (ACC) and the AUC for each class (A_{BNG}, A_{ADE}, A_{SCC}).

As happened in the 2-class comparison, not all the works included present every metric analyzed in this work. Except for the work Wang et al. [42], which includes the accuracy, the others show only the AUC. Moreover, it is not common to include a multiclass analysis for the ROC curve (as detailed in this work), and many of them only present one AUC value (as it is not specified in each work, we assume that it is the AUC global value). The case of the work presented by Coudray et al. [37] is particular, as the AUC results are presented only for cancer classes (not for the benign-tissue class).

The only work that includes the accuracy metric is [42], obtaining a result of 82%; while our work obtains a 97.11% (+15%). Moreover, this work also analyzes the global AUC value, obtaining a result of 85.6%; while the multiclass ROC curves evaluated in our work obtain values from 97.77% to 99.94% (depending on the class analyzed).

Regarding the case of the work presented by [35], the only result given is the global AUC value. In this case, the work obtains a result of 91.19%, less than the worst class results in our work (97.77%). However, it is important to remember that this work and the previous one are the one that use a 4-class classifier (including both the SCLC class, that is not used in our work). This fact is important to be mentioned because, although we are comparing ourselves with the results indicated in their work (and our system significantly outperforms them), we do not know whether their results would improve if they did not use this fourth class. In our case, the reason for not using this class in our classifier is because the dataset used does not include it; but in future works we in-



This sample most likely belongs to squamous cell carcinoma with a 100% confidence.

Other cases:
- Benign tissue: 0%
- Adenocarcinoma: 0%

This sample most likely belongs to squamous cell carcinoma with a 95.54% confidence.

Other cases:
- Benign tissue: 0%
- Adenocarcinoma: 4.46%

This sample most likely belongs to adenocarcinoma with a 79.12% confidence.

Other cases:
- Benign tissue: 0%
- Squamous cell carcinoma: 20.88%

This sample most likely belongs to adenocarcinoma with a 54.06% confidence.

Other cases:
- Benign tissue: 0%
- Squamous cell carcinoma: 45.94%

Fig. 15. Some reports obtained from SCC images: (a,b) correct classified cases with high confidence; (c,d) wrongly classified cases with lower confidence.

Table 22
Classifier complexity regarding the number of layers used in the CNN. In the case of works that use more than one model, the less-complex one is indicated.

Work	Model	N_{CL}	N_{ML}	N_{DL}
[35]	AlexNet	5	3	3
[37]	InceptionV3	40+	10+	3
[39]	InceptionV3	40+	10+	3
[42]	VGG16	13	5	3
This Work (2021)	Custom	3	3	2

tend to use other datasets that include it to be able to compare ourselves with all works under equal conditions.

And, finally, the other remaining works [37,39] present AUC values for each class (with the exception detailed previously for the work [37]), obtaining a maximum value of 99.85% for BNG class (our work obtains a 99.94%), 93.31% for ADE class (our work obtains a 97.77%), and a 97.7% for SCC class (our work obtains a 97.92%). Thus, our classifier obtains the best results in all cases and classes evaluated.

However, as some results are very close, a summary of the complexity of each classifier is included in Table 22. In this table, three new columns are included: the number of convolution layers (N_{CL}), the number of maxpooling layers (N_{ML}, and the number of dense layers (N_{DL}) used by each classifier.

As can be seen in Table 22, the works whose results are closer to ours use a much more complex network model (InceptionV3) than the others. In fact, the number of convolutional layers (those that require more computational workload) is more than 10 times higher than those used in our work. According to these information, the high results of these two works seem to be related with the increased complexity of the network; whereas, in our case, a custom lightweight network has been combined with a more specific preprocessing of the images.

At this point, it is important to remember that the dataset used in this work has been compiled after applying immunohistochemistry techniques to the samples, which enhances the colours of the most remarkable features of the images. Even so, if we look at the results of the greyscale training (where the previous treatments would no longer have an effect), the results are still very high (with a decrease of less than 3%, reaching accuracy values of over 94% in general, and between 94 and 98.5% for each individual class). Under these circumstances, using the results obtained by the greyscale CNN, the only work that would achieve better results would be the one developed by [39]. However, the computational requirements of that classifier are more than 10 times those needed in our work.

Nevertheless, immunohistochemistry techniques are continuously applied and, as the aim of these systems is to improve their accuracy, it is essential to work with a dataset already treated with these techniques. Moreover, these results also demonstrate the enormous usefulness of all these feature enhancement techniques.

To conclude the comparison section, we can summarize that our system obtains significantly better results than those obtained in previous works for 2-class classifiers. While, in the case of the 3-class classifier, the results are still better but closer; but, even so, it seems that the complexity of the network model plays an important role in those results.

Finally, the application of xAI techniques to the developed classifier are detailed in the next two subsections. First, GradCAM mechanism is detailed; and, secondly, Occlusion Sensitivity is described.

4.5. GradCAM results

The above results and comparison show that the system developed in this work obtains better classification results than the previous work and requires less computational load.

However, for the report provided to the pathologist, this work provides additional information about the confidence (in percentage value) of the results provided, and a heat map specifying those aspects or areas of the image that have been taken into account for the classification.

In this final report for the health professional is of utmost importance as it will be thoroughly checked to assess the reliability of the result and to consider whether a reevaluation of the sample is necessary. Thus, the final report of the classification system is shown in Fig. 14.

As can be seen, the implemented Explainable Deep Learning algorithm (custom Grad-CAM) extracts the resulting information after the last convolution (numerical weight matrix) and converts it to a heat map. This map shows the areas on which the classifier has focused to obtain the verdict. This heat map is overlapped with the original image so that the health professional can appreciate the areas that determine the verdict. In addition, the numerical result of the classifier is extracted from the last layer of the system before applying the softmax process (activation of the class with the highest value and inhibition of the remaining ones); in this way, a percentage of reliability of the result can be provided.

As a final report, as shown in Fig. 14, the original image with the overlapped heat map, the classification obtained, and the percentage of reliability of this classification are provided. Based on these parameters, the healthcare professional can make the final verdict, which could be to validate these results or to proceed with a more thorough study of the sample.

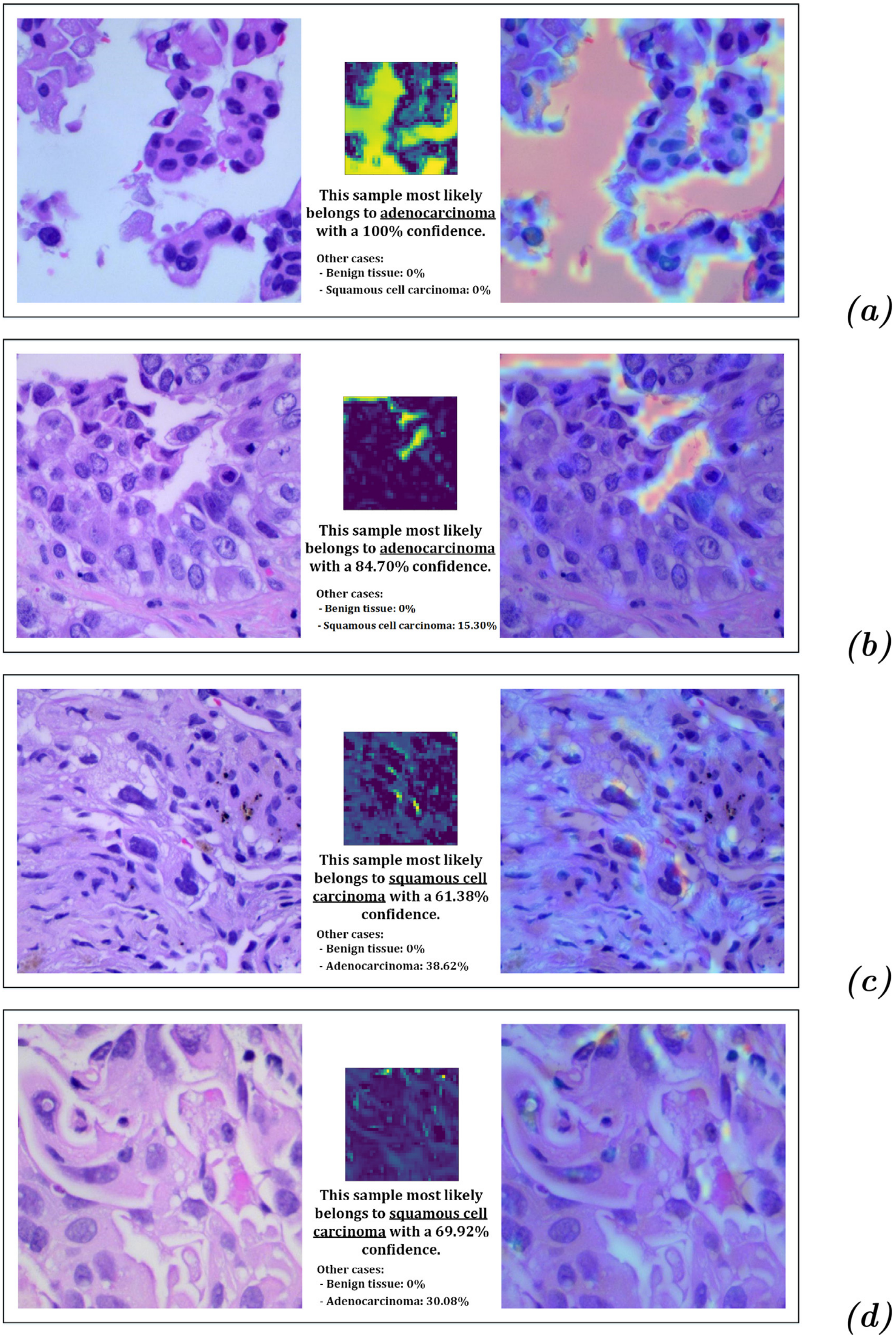
It is important to note that the original image has a resolution of 180x180 pixels, while the heat map has a resolution of 41x41 pixels (the result of the last convolution layer before maxpooling). Because of this, the heat map image must be overscaled before overlapping it to the original. This causes that, due to the decimals obtained during this process of resolution increase, some parts of the heat map do not fit perfectly with the original; however, when observing them, it is clear which parts of the image it refers to.

Reporting results for the training dataset will be shown below. Several cases will be shown for each class, with special emphasis on cases with low percentage reliability and classifier system confusions.

4.5.1. SCC cases reports

The accuracy of this class is high (97.2%), although its sensitivity is lower (95.4%). This means that there are some cases that are not classified correctly (4.63% of the cases are classified as adenocarcinoma, but no cases are classified as benign tissue). The vast majority of the cases classified correctly give confidence percentages between 90 and 100%, and the cases with confidence lower than 90% have a high probability of being wrongly classified. Thus, in this case, some samples with high result confidence are evaluated to check what the system focuses on to make the classification; and after that, two of the most unfavourable cases (classification failures) are evaluated.

In Fig. 15, four example cases are shown. Cases *a* and *b* represent two correct classifications with a high confidence value. As can be observed, for case *a* there is 0% probability of belonging to the other classes; and, for case *b*, there is only a 4.5% probability of belonging to adenocarcinoma class. On the other hand, for more than 4% of the cases, the classifier fails, and cases *c* and *d* represent two of those cases: in both of them, the classifier misses and confuses the samples with adenocarcinoma. Even so, the percentage of



This sample most likely belongs to adenocarcinoma with a 100% confidence.

Other cases:
- Benign tissue: 0%
- Squamous cell carcinoma: 0%

This sample most likely belongs to adenocarcinoma with a 84.70% confidence.

Other cases:
- Benign tissue: 0%
- Squamous cell carcinoma: 15.30%

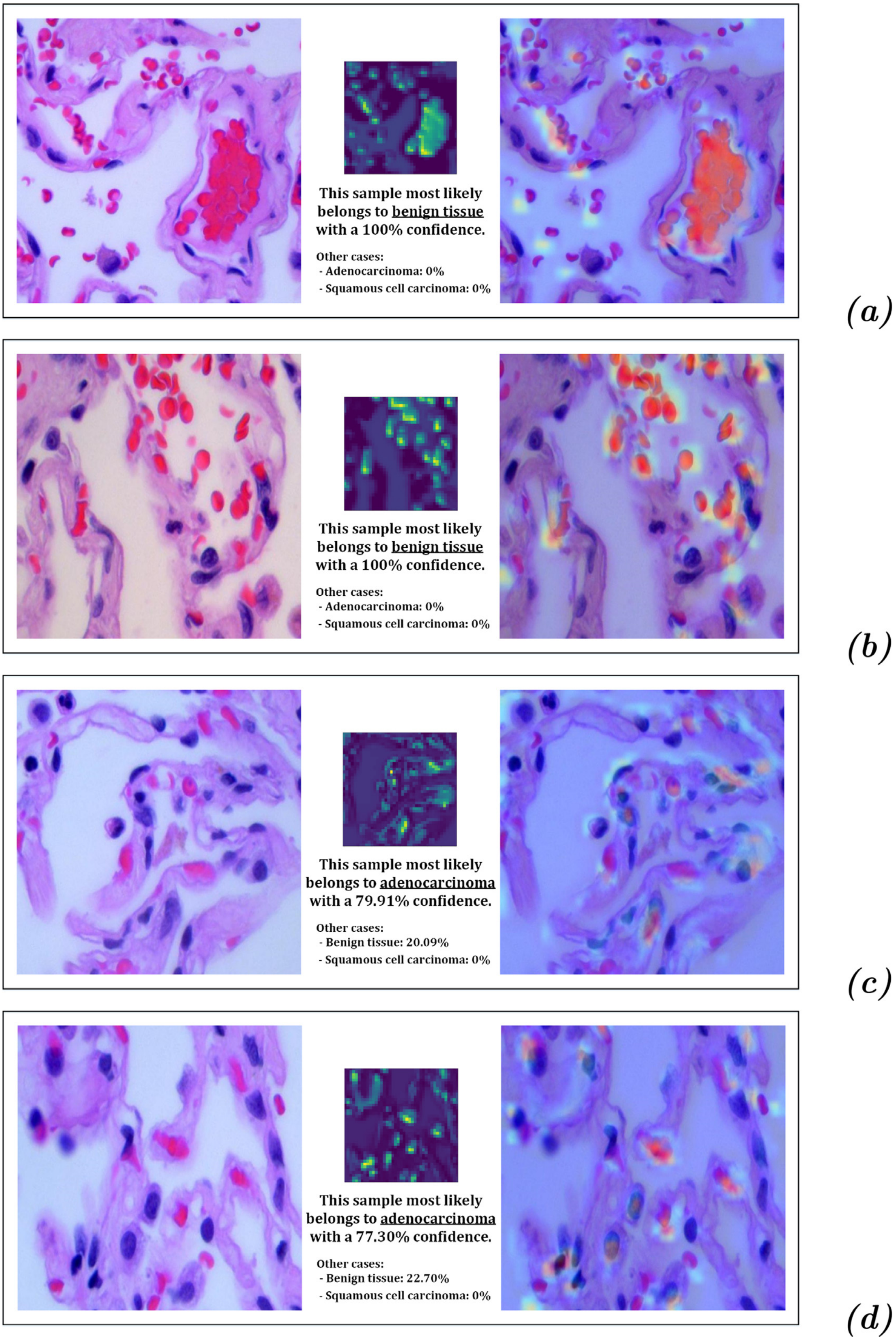
This sample most likely belongs to squamous cell carcinoma with a 61.38% confidence.

Other cases:
- Benign tissue: 0%
- Adenocarcinoma: 38.62%

This sample most likely belongs to squamous cell carcinoma with a 69.92% confidence.

Other cases:
- Benign tissue: 0%
- Adenocarcinoma: 30.08%

Fig. 16. Some reports obtained from ADE images: (a,b) correct classified cases with high confidence; (c,d) wrongly classified cases with lower confidence.


This sample most likely belongs to benign tissue with a 100% confidence.

Other cases:
- Adenocarcinoma: 0%
- Squamous cell carcinoma: 0%

This sample most likely belongs to benign tissue with a 100% confidence.

Other cases:
- Adenocarcinoma: 0%
- Squamous cell carcinoma: 0%

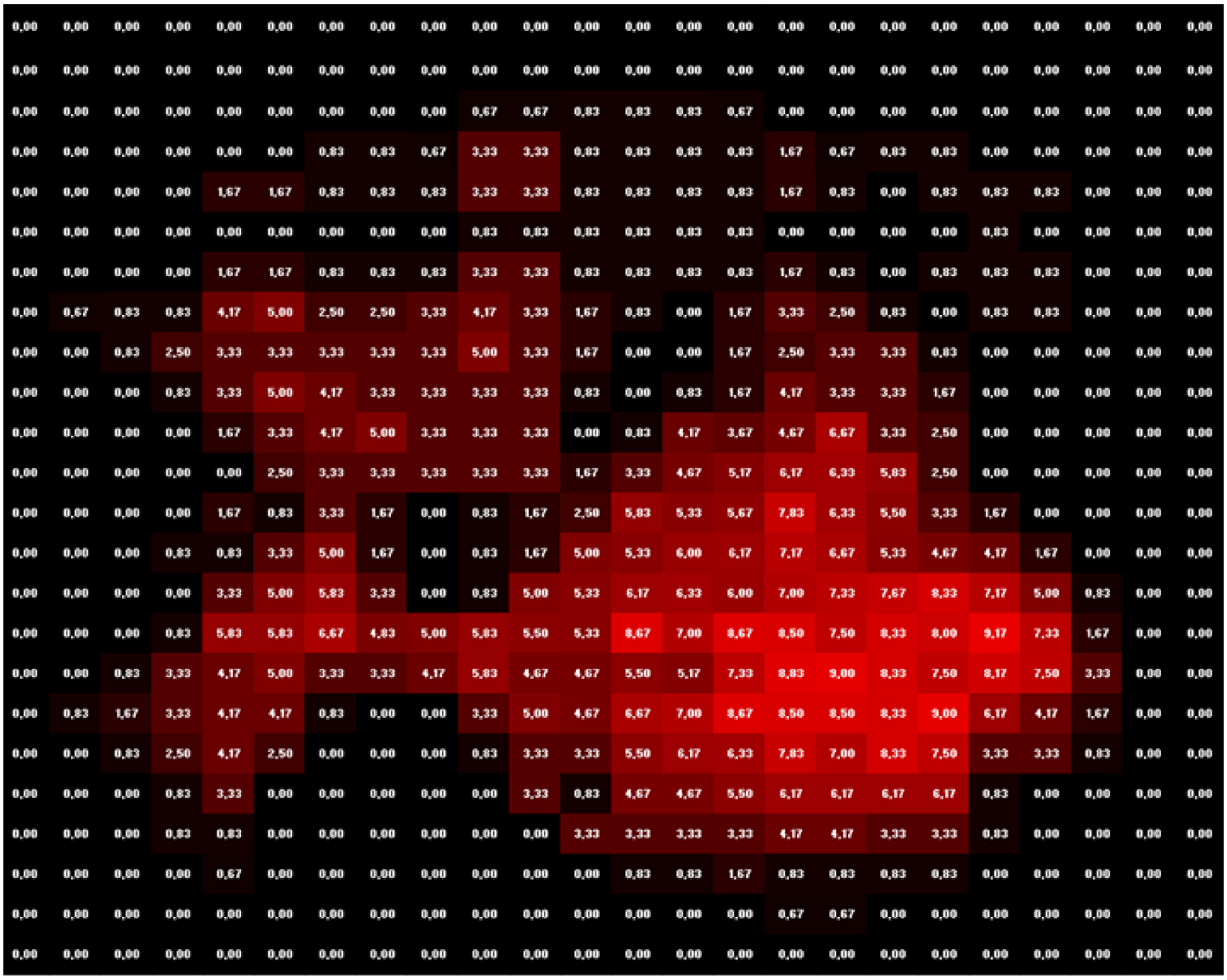
This sample most likely belongs to adenocarcinoma with a 79.91% confidence.

Other cases:
- Benign tissue: 20.09%
- Squamous cell carcinoma: 0%

This sample most likely belongs to adenocarcinoma with a 77.30% confidence.

Other cases:
- Benign tissue: 22.70%
- Squamous cell carcinoma: 0%

Fig. 17. Some reports obtained from BNG images: (a,b) correct classified cases with high confidence; (c,d) wrongly classified cases with lower confidence.



0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,67	0,67	0,83	0,83	0,83	0,67	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,83	0,83	0,67	3,33	3,33	0,83	0,83	0,83	0,83	1,67	0,67	0,83	0,83	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	1,67	1,67	0,83	0,83	0,83	3,33	3,33	0,83	0,83	0,83	0,83	1,67	0,83	0,00	0,83	0,83	0,83	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,83	0,83	0,83	0,83	0,83	0,83	0,00	0,00	0,00	0,00	0,83	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	1,67	1,67	0,83	0,83	0,83	3,33	3,33	0,83	0,83	0,83	0,83	1,67	0,83	0,00	0,83	0,83	0,83	0,00	0,00	0,00
0,00	0,67	0,83	0,83	4,17	5,00	2,50	2,50	3,33	4,17	3,33	1,67	0,83	0,00	1,67	3,33	2,50	0,83	0,00	0,83	0,83	0,00	0,00	0,00
0,00	0,00	0,83	2,50	3,33	3,33	3,33	3,33	3,33	5,00	3,33	1,67	0,00	0,00	1,67	2,50	3,33	3,33	0,83	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,83	3,33	5,00	4,17	3,33	3,33	3,33	3,33	0,83	0,00	0,83	1,67	4,17	3,33	3,33	1,67	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	1,67	3,33	4,17	5,00	3,33	3,33	3,33	0,00	0,83	4,17	3,67	4,67	6,67	3,33	2,50	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	2,50	3,33	3,33	3,33	3,33	3,33	1,67	3,33	4,67	5,17	6,17	6,33	5,83	2,50	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	1,67	0,83	3,33	1,67	0,00	0,83	1,67	2,50	5,83	5,33	5,67	7,83	6,33	5,50	3,33	1,67	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,83	0,83	3,33	5,00	1,67	0,00	0,83	1,67	5,00	5,33	6,00	6,17	7,17	6,67	5,33	4,67	4,17	1,67	0,00	0,00	0,00
0,00	0,00	0,00	0,00	3,33	5,00	5,83	3,33	0,00	0,83	5,00	5,33	6,17	6,33	6,00	7,00	7,33	7,67	8,33	7,17	5,00	0,83	0,00	0,00
0,00	0,00	0,00	0,83	5,83	5,83	6,67	4,83	5,00	5,83	5,50	5,33	8,67	7,00	8,67	8,50	7,50	8,33	8,00	9,17	7,33	1,67	0,00	0,00
0,00	0,00	0,83	3,33	4,17	5,00	3,33	3,33	4,17	5,83	4,67	4,67	5,50	5,17	7,33	8,83	9,00	8,33	7,50	8,17	7,50	3,33	0,00	0,00
0,00	0,83	1,67	3,33	4,17	4,17	0,83	0,00	0,00	3,33	5,00	4,67	6,67	7,00	8,67	8,50	8,50	8,33	9,00	6,17	4,17	1,67	0,00	0,00
0,00	0,00	0,83	2,50	4,17	2,50	0,00	0,00	0,00	0,83	3,33	3,33	5,50	6,17	6,33	7,83	7,00	8,33	7,50	3,33	3,33	0,83	0,00	0,00
0,00	0,00	0,00	0,83	3,33	0,00	0,00	0,00	0,00	0,00	3,33	0,83	4,67	4,67	5,50	6,17	6,17	6,17	6,17	0,83	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,83	0,83	0,00	0,00	0,00	0,00	0,00	0,00	3,33	3,33	3,33	3,33	4,17	4,17	3,33	3,33	0,83	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,67	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,83	0,83	1,67	0,83	0,83	0,83	0,83	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,67	0,67	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

Fig. 18. Occlusion Sensitivity heat map obtained for BNG class.

confidence shows a significant drop, which may be an indication to be taken into account by the health professional. And, moreover, it can be observed that the probability of belonging to the squamous cell carcinoma class is around 21% for case *c* and 46% for case *d*.

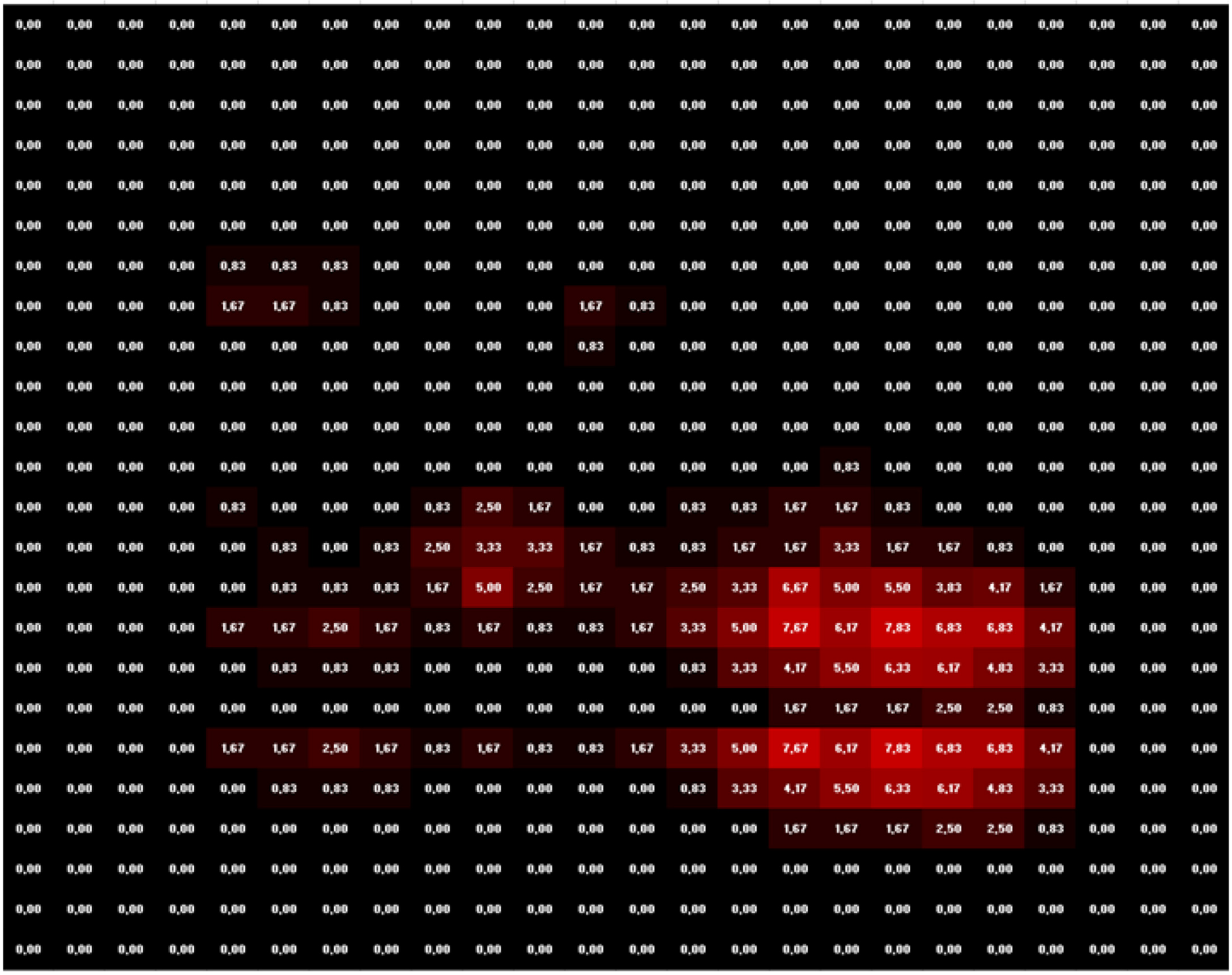
Looking at the heat maps, it can be interpreted that the classifier has used the concentration of dark cells as a differentiating element of this class. While in the first two cases they are clearly observed, in the last two cases there are parts of the samples where mucosal and/or connective tissue are present (white parts); so, this element causes the classifier to fail because it is a differentiator from the other classes (in the other classes, the presence of connective tissue and/or mucus is more present).

4.5.2. ADE cases reports

The accuracy of this class is the lowest (97.1%), although it is a high result. Regarding the other parameters, this class presents a low value of precision (95.4%); this means that several samples from other classes are classified as it (false positives). Moreover, the sensitivity has a value of 96.8%, so there are some samples from this class classified as other class (false negatives). These two events occur between SCC and ADE classes, as both represent cancer tissue and, in some cases, both are interchanged. It is important to mention that, among the three classes, the correctly classified samples for this class give the lowest confidence values (between 80 and 100%). In contrast, misclassified cases show lower confidence levels than cases detected in the other classes (being easier to differentiate). Thus, following the same progression as the previous class, some samples with high result confidence are evaluated to check what the system focuses on to make the classification; and after that, two unfavourable cases (classification failures) are evaluated.

In Fig. 16, four example cases are shown. Cases *a* and *b* represent two correct classifications with a high confidence value: for case *a* there is 0% probability of belonging to the other classes (100% confidence classification); and, for case *b*, there is a 15.30% probability of belonging to squamous cell carcinoma class. On the other hand, cases *c* and *d* represent two of the classification failures: in both of them, the classifier misses and confuses the samples with squamous cell carcinoma. But, for this class, the percentage of confidence is very low, which makes it easier to detect them by the health professional. And, moreover, it can be observed that the probability of belonging to the adenocarcinoma class is around 38% for case *c* and 30% for case *d*.

Looking at the heat maps, it can be interpreted that the classifier has used the mucosal and/or connective tissue as a differentiating element of this class. While in the first two cases it is



0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,83	0,83	0,83	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	1,67	1,67	0,83	0,00	0,00	0,00	0,00	1,67	0,83	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,83	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,83	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,83	0,00	0,00	0,00	0,83	2,50	1,67	0,00	0,00	0,83	0,83	1,67	1,67	0,83	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,83	0,00	0,83	2,50	3,33	3,33	1,67	0,83	0,83	1,67	1,67	3,33	1,67	1,67	0,83	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,83	0,83	0,83	1,67	5,00	2,50	1,67	1,67	2,50	3,33	6,67	5,00	5,50	3,83	4,17	1,67	0,00	0,00	0,00
0,00	0,00	0,00	0,00	1,67	1,67	2,50	1,67	0,83	1,67	0,83	0,83	1,67	3,33	5,00	7,67	6,17	7,83	6,83	6,83	4,17	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,83	0,83	0,83	0,00	0,00	0,00	0,00	0,00	0,83	3,33	4,17	5,50	6,33	6,17	4,83	3,33	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,67	1,67	1,67	2,50	2,50	0,83	0,00	0,00	0,00
0,00	0,00	0,00	0,00	1,67	1,67	2,50	1,67	0,83	1,67	0,83	0,83	1,67	3,33	5,00	7,67	6,17	7,83	6,83	6,83	4,17	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,83	0,83	0,83	0,00	0,00	0,00	0,00	0,00	0,83	3,33	4,17	5,50	6,33	6,17	4,83	3,33	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,67	1,67	1,67	2,50	2,50	0,83	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

Fig. 19. Occlusion Sensitivity heat map obtained for ADE class.

clearly observed, in the last two cases there are parts of the samples where mucosal and/or connective tissue appears in a low concentration, and several dark cells are present; so, this fact causes the classifier to fail because it confuses the dark cells concentration with the squamous cell carcinoma differenciator.

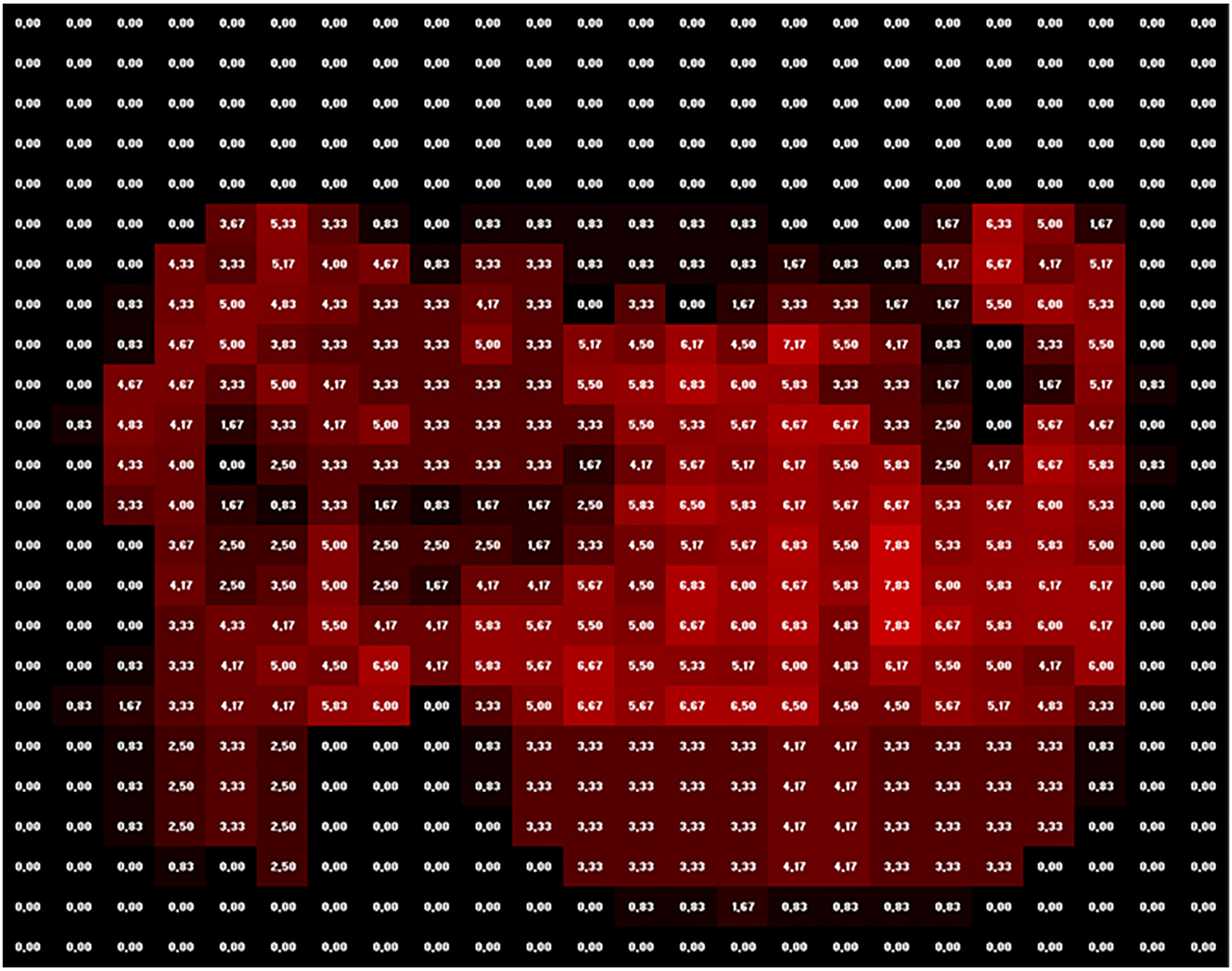
4.5.3. BNG cases reports

This last class has the highest accuracy in the system (close to 100%), so it is difficult to find cases where the report's confidence percentage is not close to 100%. Even so, some cases with high result confidence are evaluated to check what the system focuses on to make the classification; and after that, the most unfavourable cases (the only two cases in which the classification of this class fails) are evaluated.

In Fig. 17, four example cases are shown. Cases *a* and *b* represent two correct classifications with a high confidence value. As can be observed, there is 0% probability of belonging to the other classes. In fact, the classifier is correct in the vast majority of cases with a confidence level of 100% (this is the situation in more than 97% of the cases). For the remaining 3%, the classifier is also correct in almost all cases (although indicating confidence percentages between 99 and 100%). And, for only two specific cases (shown in Fig. 17 as cases *c* and *d*), the classifier misses and confuses the samples with adenocarcinoma. Even so, the percentage of confidence shows a significant drop, which may be an indication to be taken into account by the health professional. And, moreover, it can be observed that the probability of belonging to the benign tissue class is around 20–22%.

Looking at the heat maps, it can be interpreted that the classifier has used the appearance of pink cells as a differentiating element of this class. While in the first two cases they are clearly observed, in the last two cases pink cells appear in a low percentage and are mixed with darker cells. Because of that, in two cases, this low concentration of pink cells and mucosal/connective tissue makes the classifier to wrongly tag those two samples as adenocarcinoma.

After analysing the customised reports provided by the Explainable Deep Learning system, it can be seen that the classification carried out is correct in most of the cases, and the reports provided are very useful for the healthcare professional. In the case that the system does not match, the pathologist will be able to detect it due to the low confidence percentage of the report and, thanks to this, analyse those samples more closely. Moreover, if we look at the additional information in the reports where the classifier fails, there is always a significant percentage of belonging to another class (and it is indeed that class that is correct). Therefore,



0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	3,67	5,33	3,33	0,83	0,00	0,83	0,83	0,83	0,83	0,83	0,83	0,00	0,00	0,00	1,67	6,33	5,00	1,67	0,00	0,00
0,00	0,00	0,00	4,33	3,33	5,17	4,00	4,67	0,83	3,33	3,33	0,83	0,83	0,83	0,83	1,67	0,83	0,83	4,17	6,67	4,17	5,17	0,00	0,00
0,00	0,00	0,83	4,33	5,00	4,83	4,33	3,33	3,33	4,17	3,33	0,00	3,33	0,00	1,67	3,33	3,33	1,67	1,67	5,50	6,00	5,33	0,00	0,00
0,00	0,00	0,83	4,67	5,00	3,83	3,33	3,33	3,33	5,00	3,33	5,17	4,50	6,17	4,50	7,17	5,50	4,17	0,83	0,00	3,33	5,50	0,00	0,00
0,00	0,00	4,67	4,67	3,33	5,00	4,17	3,33	3,33	3,33	3,33	5,50	5,83	6,83	6,00	5,83	3,33	3,33	1,67	0,00	1,67	5,17	0,83	0,00
0,00	0,83	4,83	4,17	1,67	3,33	4,17	5,00	3,33	3,33	3,33	3,33	5,50	5,33	5,67	6,67	6,67	3,33	2,50	0,00	5,67	4,67	0,00	0,00
0,00	0,00	4,33	4,00	0,00	2,50	3,33	3,33	3,33	3,33	3,33	1,67	4,17	5,67	5,17	6,17	5,50	5,83	2,50	4,17	6,67	5,83	0,83	0,00
0,00	0,00	3,33	4,00	1,67	0,83	3,33	1,67	0,83	1,67	1,67	2,50	5,83	6,50	5,83	6,17	5,67	6,67	5,33	5,67	6,00	5,33	0,00	0,00
0,00	0,00	0,00	3,67	2,50	2,50	5,00	2,50	2,50	2,50	1,67	3,33	4,50	5,17	5,67	6,83	5,50	7,83	5,33	5,83	5,83	5,00	0,00	0,00
0,00	0,00	0,00	4,17	2,50	3,50	5,00	2,50	1,67	4,17	4,17	5,67	4,50	6,83	6,00	6,67	5,83	7,83	6,00	5,83	6,17	6,17	0,00	0,00
0,00	0,00	0,00	3,33	4,33	4,17	5,50	4,17	4,17	5,83	5,67	5,50	5,00	6,67	6,00	6,83	4,83	7,83	6,67	5,83	6,00	6,17	0,00	0,00
0,00	0,00	0,83	3,33	4,17	5,00	4,50	6,50	4,17	5,83	5,67	6,67	5,50	5,33	5,17	6,00	4,83	6,17	5,50	5,00	4,17	6,00	0,00	0,00
0,00	0,83	1,67	3,33	4,17	4,17	5,83	6,00	0,00	3,33	5,00	6,67	5,67	6,67	6,50	6,50	4,50	4,50	5,67	5,17	4,83	3,33	0,00	0,00
0,00	0,00	0,83	2,50	3,33	2,50	0,00	0,00	0,00	0,83	3,33	3,33	3,33	3,33	3,33	4,17	4,17	3,33	3,33	3,33	3,33	0,83	0,00	0,00
0,00	0,00	0,83	2,50	3,33	2,50	0,00	0,00	0,00	0,83	3,33	3,33	3,33	3,33	3,33	4,17	4,17	3,33	3,33	3,33	3,33	0,83	0,00	0,00
0,00	0,00	0,83	2,50	3,33	2,50	0,00	0,00	0,00	0,00	3,33	3,33	3,33	3,33	3,33	4,17	4,17	3,33	3,33	3,33	3,33	0,00	0,00	0,00
0,00	0,00	0,00	0,83	0,00	2,50	0,00	0,00	0,00	0,00	0,00	3,33	3,33	3,33	3,33	4,17	4,17	3,33	3,33	3,33	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,83	0,83	1,67	0,83	0,83	0,83	0,83	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

Fig. 20. Occlusion Sensitivity heat map obtained for SCC class.

the pathologist can analyse the doubtful cases and look at the second option provided in the report (class with the second highest confidence).

4.6. Occlusion sensitivity results

In this subsection, the second xAI technique is applied in order to evaluate the robustness of the system. Unlike GradCAM, this time modifications will be made to the input images.

So, as detailed before, by eliminating some parts of the input information (images), the accuracy reduction occasioned by these modifications can be evaluated and, because of that, the system's robustness can be evaluated.

For this case, several tests are carried out with the test subset, eliminating 32x32 pixel areas from each image and observing the result obtained in each case. As the images are 768x768 pixels in resolution, 24x24 combinations are obtained (576 combinations in total).

The results are shown as the average of the variation in the accuracy of the classifier after the elimination of each 32x32 pixel sector from all the images for each class independently. In order to make it easier to observe, it is represented visually by a heat map and class by class. Results are analysed class by class.

4.7. BNG class

A maximum reduction of 9.17% is obtained in the classification accuracy of this class, and an average reduction of 1.73%. Fig. 18 represents numerically in each box the percentage reduction of the accuracy of this class when eliminating that 32x32 pixels box from all the images of the class; in the same way, the colour represents the most critical areas to eliminate. It can be observed that the areas close to the edges do not cause any problems, and it is in the lower central area where most of the information used by the system to classify this class is concentrated.

4.8. ADE class

A maximum reduction of 7.83% is obtained in the classification accuracy of this class, and an average reduction of 0.54%, being a more robust class than the previous one in terms of perturbations. In Fig. 19 the information is represented numerically and visually for each box. It can be seen that the areas near the edges still do not cause any type of problem and, on this occasion, it is in the lower right corner where most of the information used by the system to classify this class is concentrated.

4.9. SCC class

A maximum reduction of 7.83% is obtained in the classification accuracy of this class, and an average reduction of 2.13%, being the least robust class of the three in terms of perturbations. Fig. 20 represents numerically and visually the information for each box. It can be observed that the areas close to the edges still do not cause any type of problem and, on this occasion, something similar to the first class occurs and the information is distributed very homogeneously throughout the central area of the image.

The previous results demonstrate the robustness of the system to perturbations, with an average reduction in accuracy between 0.54 and 2.13% for perturbations affecting 32x32 pixel squares.

To conclude this work, we are able to state that several possibilities for the implementation of a diagnosis aid system for lung cancer detection with histopathological images are analysed, using classifiers based on convolutional neural networks (CNN). Among the possibilities and combinations designed and evaluated, the colour image classifier with differentiation between three classes (benign tissue, adenocarcinoma, and squamous cell carcinoma) is the one selected for providing more information (three classes) and having a very high accuracy rate (97.11%), as well as an area under the ROC curve (AUC) value higher than 97.7% for all classes.

Next, this work has been compared with previous works related to the same topic and developed in recent years, and focused on using classifiers based on CNNs with histopathological images too. The comparison indicates that our system obtains the best results in both accuracy and AUC. In addition, it is shown that the computational workload of the developed classifier in this work is significantly lower than other works in which similar results are achieved. All this has been detailed in a reasoned manner.

Ultimately, a reporting module is added to the classifier for the healthcare professional, extracting the information from the last convolutional layer to generate a heat map that, overlapped to the original image, clearly shows the highlighted areas of the image on which the classifier has focused to perform the classification. In the same way, and based on the final numerical results of the system, the confidence percentage of belonging to each of the three classes is extracted. All this information is provided as a final report so that, with this information, the pathologist can make the decision to accept the classification or carry out a test personally. Based on this module, various reports are shown for each of the classes, with special emphasis on analysing the most unfavourable cases.

Summarizing, the system designed, developed, and evaluated in this work represents a significant improvement both in classification results and in the information provided to the specialist with respect to the previous work with which it has been compared. In the same way, the inclusion of explainable artificial intelligence (xAI) or explainable deep learning (xDL) techniques applied to the healthcare field proves to be necessary in diagnostic aid systems because, although these classification systems normally fail less than a healthcare professional, they are not infallible; and it is strictly necessary in cases of serious diseases (as in this case) that the pathologist has at his disposal all the information and justifications available to make the decision that best benefits the patient. Moreover, these techniques help to analyze the robustness of the classifier.

Acknowledgments

This work was partially supported by the Andalusian Regional I+D+i FEDER Project under Grant DAFNE US-1381619 and partially supported by "Robotics and Technology of Computers Lab." (TEP-108) of the University of Seville (Spain).

Funding

Open Access funding provided thanks to the CRUE-CSIC agreement with Elsevier.

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.

Ethical statement

No ethical committee has been consulted as this work uses a public available dataset.

CRediT authorship contribution statement

Javier Civit-Masot: Conceptualization, Writing – original draft. **Alejandro Bañuls-Beaterio:** Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft. **Manuel Domínguez-Morales:** Conceptualization, Formal analysis, Investigation, Writing – original draft, Funding acquisition, Supervision. **Manuel Rivas-Pérez:** Formal analysis, Investigation, Writing – review & editing, Resources. **Luis Muñoz-Saavedra:** Methodology, Writing – review & editing. **José M. Rodríguez Corral:** Writing – review & editing, Resources, Supervision.

References

[1] H. Sung, et al., Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J Clin 71 (3) (2021) 209–249.

[2] M. Morra, E. Potts, Choices, Harper Collins, 2003.

[3] B. Pérez, et al., Estimación de la oferta y demanda de médicos especialistas: España 2018–2030 (2019).

[4] A. Madabhushi, G. Lee, Image analysis and machine learning in digital pathology: challenges and opportunities, Med Image Anal 33 (2016) 170–175.

[5] Z. Liu, et al., Deep learning framework based on integration of s-mask r-cnn and inception-v3 for ultrasound image-aided diagnosis of prostate cancer, Future Generation Computer Systems 114 (2021) 358–367.

[6] C. Syrykh, et al., Accurate diagnosis of lymphoma on whole-slide histopathology images using deep learning, NPJ digital medicine 3 (1) (2020) 1–8.

[7] C. Roncato, et al., Colour doppler ultrasound of temporal arteries for the diagnosis of giant cell arteritis: a multicentre deep learning study, Clin Exp Rheumatol 38 (Suppl 124) (2020) S120–25.

[8] R. Kundu, et al., Pneumonia detection in chest x-ray images using an ensemble of deep learning models, PLoS ONE 16 (9) (2021) e0256630.

[9] W. Lotter, et al., Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach, Nat. Med. 27 (2) (2021) 244–249.

[10] S.M. Thomas, et al., Interpretable deep learning systems for multi-class segmentation and classification of non-melanoma skin cancer, Med Image Anal 68 (2021) 101915.

[11] M. Mark Priebe, R. Markin, Review of anatomic pathology and diagnostic radiology quality assurance tools to reduce diagnostic discordance in cancer, Acta Scientific Cancer Biology 3 (2019) 04–11.

[12] A. Wright, et al., Clinical decision support alert malfunctions: analysis and empirically derived taxonomy, Journal of the American Medical Informatics Association 25 (5) (2018) 496–506.

[13] W.J. Von-Eschenbach, Transparency and the black box problem: why we do not trust ai, Philosophy & Technology (2021) 1–16.

[14] A. Singh, et al., Explainable deep learning models in medical image analysis, Journal of Imaging 6 (6) (2020) 52.

[15] P. Angelov, E. Soares, Towards explainable deep neural networks (xdnn), Neural Networks 130 (2020) 185–194.

[16] Q. Xue, M.C. Chuah, Explainable deep learning based medical diagnostic system, Smart Health 13 (2019) 100068.

[17] L. Brunese, et al., Explainable deep learning for pulmonary disease and coronavirus covid-19 detection from x-rays, Comput Methods Programs Biomed 196 (2020) 105608.

[18] L. Muñoz-Saavedra, et al., Affective state assistant for helping users with cognition disabilities using neural networks, Electronics (Basel) 9 (11) (2020) 1843.

[19] M. Domínguez-Morales, et al., Smart footwear insole for recognition of foot pronation and supination using neural networks, Applied Sciences 9 (19) (2019) 3970.

[20] F. Luna-Perejón, et al., Wearable fall detector using recurrent neural networks, Sensors 19 (22) (2019) 4885.

[21] F. Luna-Perejón, et al., Ankfall falls, falling risks and daily-life activities dataset with an ankle-placed accelerometer and training using recurrent neural networks, Sensors 21 (5) (2021) 1889.
[22] J. Civit-Masot, et al., Multi-dataset training for medical image segmentation as a service, in: UCCI 2019: 11th International Joint Conference on Computational Intelligence (2019), pp. 542–547., ScitePress Digital Library, 2019.
[23] J. Civit-Masot, et al., Deep learning system for covid-19 diagnosis aid using X-ray pulmonary images, Applied Sciences 10 (13) (2020) 4640.
[24] J. Civit-Masot, et al., Dual machine-learning system to aid glaucoma diagnosis using disc and cup feature extraction, IEEE Access 8 (2020) 127519–127529.
[25] J. Civit-Masot, et al., A study on the use of edge tpus for eye fundus image segmentation, Eng Appl Artif Intell 104 (2021) 104384.
[26] I. Amaya-Rodríguez, et al., Glioma diagnosis aid through CNNS and fuzzy-c means for mri, in: UCCI 2019: 11th International Joint Conference on Computational Intelligence (2019), pp. 528–535., ScitePress Digital Library, 2019.
[27] L. Durán-López, et al., Breast cancer automatic diagnosis system using faster regional convolutional neural networks, in: IJCCI 2019: 11th International Joint Conference on Computational Intelligence (2019), pp. 444–448., ScitePress Digital Library, 2019.
[28] A.A. Borkowski, et al., LC25000 Lung and colon histopathological image dataset (2019). https://github.com/tampapath/lung_colon_image_set/.
[29] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Inf. Process. & Manag. 45 (4) (2009) 427–437.
[30] Z.H. Hoo, J. Candlish, D. Teare, What is an roc curve?, 2017.
[31] G. Ras, N. Xie, M. Van-Gerven, D. Doran, Explainable deep learning: a field guide for the uninitiated (2021). arXiv preprint arXiv: 1207.0580.
[32] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad–cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.
[33] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European Conference on Computer Vision, Springer, 2014, pp. 818–833.
[34] P.J. McCarthy, C.B. Snowden, The bootstrap and finite population sampling (1985).
[35] Z. Li, et al., Computer-aided diagnosis of lung carcinoma using deep learning-a pilot study, arXiv preprint arXiv:1803.05471 (2018).
[36] S. Wang, et al., Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome, Sci Rep 8 (1) (2018) 1–9.
[37] N. Coudray, et al., Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning, Nat. Med. 24 (10) (2018) 1559–1567.
[38] X. Wang, et al., Weakly supervised learning for whole slide lung cancer image classification (2018).
[39] S. Bilaloglu, et al., Efficient pan-cancer whole-slide image classification and outlier detection using convolutional neural networks, bioRxiv (2019) 633123.
[40] J. Noorbakhsh, et al., Pan-cancer classifications of tumor histological images using deep learning, BioRxiv (2019) 715656.
[41] L. Sha, et al., Multi-field-of-view deep learning model predicts nonsmall cell lung cancer programmed death-ligand 1 status from whole-slide hematoxylin and eosin images, J Pathol Inform 10 (2019).
[42] X. Wang, et al., Weakly supervised deep learning for whole slide lung cancer image analysis, IEEE Trans Cybern 50 (9) (2019) 3950–3962.
[43] F. Kanavati, et al., Weakly-supervised learning for lung carcinoma classification using deep learning, Sci Rep 10 (1) (2020) 1–11.
[44] K.-H. Yu, et al., Classifying non-small cell lung cancer types and transcriptomic subtypes using convolutional neural networks, Journal of the American Medical Informatics Association 27 (5) (2020) 757–769.
[45] M. Kriegsmann, et al., Deep learning for the classification of small-cell and non-small-cell lung cancer, Cancers (Basel) 12 (6) (2020) 1604.
[46] Y. Guo, et al., Histological subtypes classification of lung cancers on ct images using 3d deep learning and radiomics, Acad Radiol 28 (9) (2021) e258–e266.