Digital Pathology Image Classification using Deep Learning

Diego Caballero García-Alcaide, Máximo Rodríguez Herrero, Lucía Guijarro Martínez
Universidad Carlos III de Madrid
Av. de la Universidad, 30, 28911, Leganés, Madrid, España
dicaball@inf.uc3m.es, maxrodri@inf.uc3m.es, 100455351@alumnos.uc3m.es

***Abstract*—Histological image analysis plays a crucial role in pathological anatomy, aiding in disease diagnosis and characterization. In this work, we implement a state-of-the-art DenseNet deep learning model to address two colorectal tissue classification tasks. The first is a binary classification task to distinguish between tumor and non-tumor tissue, while the second is a multi-class classification task to identify various tissue types, including tumorous tissue. To enhance data quality and model performance, we employ transfer learning and preprocessing techniques. The results demonstrate the potential of these approaches to support automated diagnosis in digital pathology, achieving state-of-the-art AUC metrics.**

I. Introduction

A. Context and Motivation

Medical diagnosis based on imaging has been a fundamental pillar in modern medicine, with significant advances in recent decades thanks to artificial intelligence (AI). Machine learning algorithms have enabled the automation and enhancement of medical image analysis across various fields, from radiology to histology, among others. In particular, deep neural networks have proven to be effective tools for the classification of digital pathology images, facilitating disease detection and greatly assisting healthcare professionals [1].

In this field, pathological anatomy plays a crucial role in characterizing diseased tissues at the microscopic level. Its study is based on the analysis of histological slides, which are thin sections of tissue adhered to a slide and stained with specific dyes to highlight cellular structures. Traditionally, the diagnostic process using these images relied on the expertise of pathologists. However, recent advances have enabled the development of models capable of automating this task with high precision, providing significant support in diagnosis [2].

Despite significant advances, histological image classification still faces several challenges, including staining variability, differences in image acquisition devices, and potential class imbalances in the data. To address these issues, convolutional neural networks (CNNs) have been developed to learn meaningful representations directly from data, eliminating the need for manual feature extraction. Additionally, transfer learning has proven to be a powerful strategy for enhancing model performance by leveraging knowledge from large-scale datasets, such as ImageNet [3], and adapting pretrained models to more specific histological classification tasks.

This work focuses on the implementation of deep neural networks for the classification of histological images of colorectal tissue. Pretrained architectures and fine-tuning techniques have been employed to enhance model generalization. Additionally, preprocessing strategies, including normalization and data augmentation, have been applied to improve the system's robustness and accuracy [4]. Our approach closely follows that of [4], which utilized VGG19; however, we opted for a more lightweight DenseNet model with only 7 million parameters. Despite its significantly lower size, our model achieved a similar AUC metric on the test set to their cross-validation results. While further improvements are needed, this outcome highlights the efficiency of DenseNet as a lighter yet highly capable alternative. Furthermore, our implementation leverages the MONAI framework, demonstrating its ease of use and its potential to make AI-driven medical solutions, helping address the shortage of medical professionals.

B. Objectives

With this context in mind, the objective of this work is to develop and evaluate deep learning-based classification models for histological images of colorectal tissue. Specifically, we aim to explore the effectiveness of lightweight architectures while maintaining strong classification performance. To achieve this, two classifiers have been trained on "NCT-CRC-HE-100K" dataset [5]:

- A binary classifier designed to distinguish between tumor and non-tumor tissue.
- A multiclass classifier capable of identifying nine different tissue types within histological images, including adipose tissue, normal mucosa, lymphocytes, and tumor epithelium, see Fig. 2.

Both models have been evaluated using clinically relevant metrics, e.g. the area under the Receiver Operating Characteristic curve (AUC-ROC) [6]. Additionally, the obtained results have been analyzed to assess their clinical applicability and identify potential improvements for both classifiers. Aiming to achieve clinically relevant sensitivity metric given a minimum value of 85% for specificity.

Sensitivity (true positive rate) is the probability that a positive prediction is truly positive, while specificity (true negative rate) is the probability that a negative prediction is truly negative. These metrics are defined in terms of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), as seen in equations 1 and 2.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2}$$

Furthermore, we explored AI explainability by generating Class Activation Maps (CAMs) for each DenseBlock of the model, as well as the final output ReLU layer [7]. This analysis provides insights into the model's decision-making process, enhancing interpretability and supporting its potential integration into clinical practice.

II. DenseNets

For tissue image classification, the DenseNet121 architecture has been used. Dense Convolutional Networks (DenseNets) [8] are a convolutional neural network architecture in which all layers are directly connected within each block. Unlike traditional networks, where each layer only receives information from the previous one, in a DenseNet, each layer receives as input the feature maps of all preceding layers and passes its output to all subsequent layers, Fig. 1.

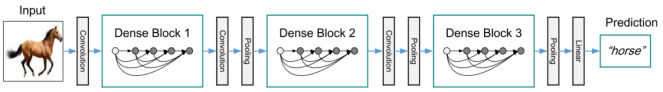

Fig. 1: Example of DenseNet with three dense blocks [8]. Hereafter, layers between two adjacent blocks will be referred to as transition layers.

Compared to ResNets [9], which introduced residual connections that allow feature summation between consecutive layers, DenseNets implement a denser connectivity pattern through feature concatenation. This approach enhances feature reuse and improves information propagation through the net.

To understand how DenseNet works, it is important to analyze its key components:

- **Composite Function**: Each layer in a DenseNet applies a nonlinear transformation defined as a function composed of three consecutive operations:
  - **Batch Normalization (BN)**: Normalizes layer activations to improve training stability.
  - **Rectified Linear Unit (ReLU)**: An activation function that introduces non-linearity into the network.
  - **Convolution** 3×3: Extracts spatial features using convolutional filters.
- **Pooling Layers**: In convolutional networks, it is necessary to progressively reduce the spatial resolution of feature maps. To achieve this, DenseNet employs transition layers between dense blocks, which include:
  - **Convolution** 1×1: Reduces the dimensionality of feature maps to improve computational efficiency.
  - **Average Pooling** 2×2: After convolution, the spatial resolution of feature maps is reduced, decreasing the amount of data processed in subsequent layers.
- **Growth Rate**: The growth rate in DenseNet is a hyperparameter, k, that controls how much new information each layer adds to the "global state" of the network. One distinguishing feature of DenseNet compared to other architectures is that it can have a relatively small growth rate. However, this growth is sufficient to achieve high performance, as each layer has access to all previously generated feature maps, allowing it to learn more efficiently without redundant information propagation.
- **Bottleneck Layers**: To improve computational efficiency, DenseNets incorporate bottleneck layers. As mentioned earlier, although each layer in DenseNet produces a relatively small number of feature maps (k), it typically has many more input maps. To reduce the number of input feature maps and enhance computational efficiency, a 1×1 convolution is introduced before a 3×3 convolution. The 1×1 convolution acts as a "bottleneck," reducing the dimensionality of feature maps and making the subsequent 3×3 convolution more efficient.
- **Compression**: To prevent the number of feature maps from growing excessively, DenseNet introduces a compression strategy in transition layers. In these layers, if a dense block generates m feature maps, the transition layer retains only a percentage θ of them, where $0 < \theta \leq 1$.

A. DenseNet-121

The DenseNet-121 architecture is an optimized variant of DenseNet. Its structure consists of the following elements:

- **Number of Layers**: The architecture contains 121 layers, organized into dense blocks and transition layers.
- **Growth Rate**: Each layer adds $k = 32$ new feature maps to the network.
- **Dense Blocks**: The network is structured into four dense blocks, separated by transition layers that reduce feature map dimensionality.
- **Initial Layers**: A 7×7 convolution is applied, followed by a 3×3 max pooling operation to reduce spatial resolution and improve feature extraction.
- **Classification Layer**: After the final dense block, a global average pooling operation is performed, followed by a fully connected layer with a softmax activation function for final classification.

DenseNet-121 achieves a good balance between performance and model size. This architecture is used for both binary and multiclass classification tasks.

III. Dataset

For this work we use an openly published dataset containing more than 100,000 histological images of human colorectal tissue. The data is available on Zenodo [5], a general-purpose open-access repository developed under the European OpenAIRE program. This dataset is divided into two subsets, one with 100,000 images which we used for training and another one composed of about 7,000 images which we used for the evaluation of the models.

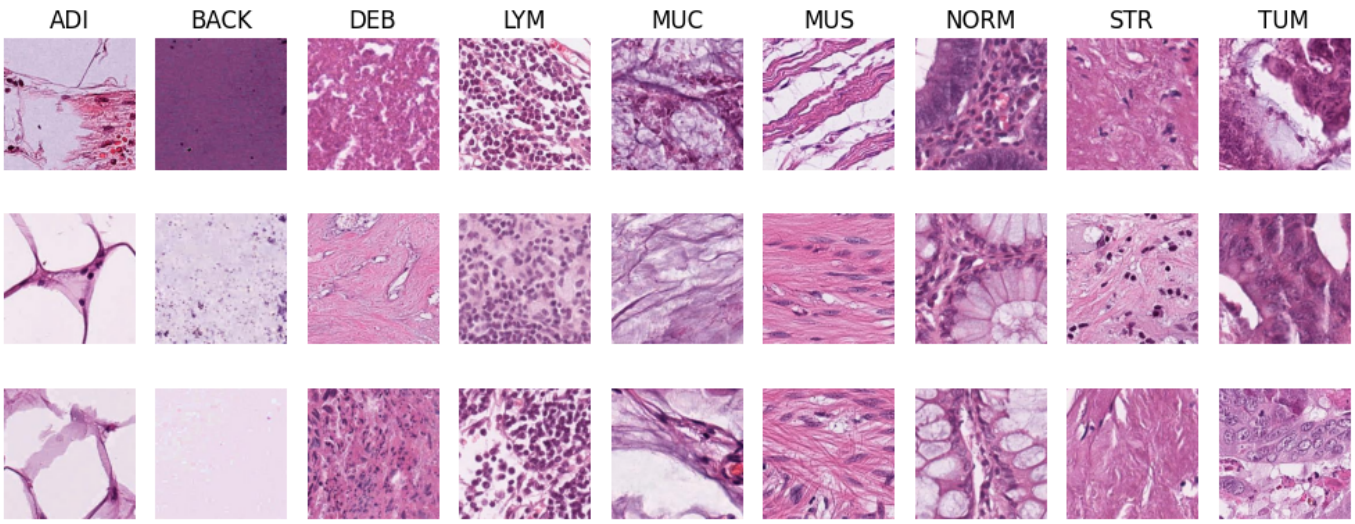

Fig. 2: Examples of images belonging to the nine different classes of the dataset. ADI, adipose tissue; BACK, background; DEB, debris; LYM, lymphocytes; MUC, mucus; MUS, smooth muscle; NORM, normal colon mucosa; STR, cancer-associated stroma; TUM, colorectal adenocarcinoma epithelium.

A. Training set

This set, named "NCT-CRC-HE-100K", contains 100,000 non-overlapping image patches from hematoxylin & eosin (H&E) stained histological images of human colorectal cancer (CRC) and normal tissue. All images are 224x224 pixels at 0.5 MPP and are color-normalized using Macenko's method. It includes nine classes, shown in Fig. 2: Adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR) and colorectal adenocarcinoma epithelium (TUM). This set was further split into train and validation sets.

B. Test set

The set used for evaluating the final models, named "CRC-VAL-HE-7K", contains 7180 image patches from $N = 50$ patients with colorectal adenocarcinoma (no overlap with patients in NCT-CRC-HE-100K). Like in the training data set, images are 224x224 pixels at 0.5 MPP and were provided by the NCT tissue bank.

IV. Implementation

A. Image Preprocessing

To obtain a balanced and well-prepared dataset for model training, the images undergo a preprocessing pipeline consisting of several steps:

- **Format Conversion**: The original .tif images are converted to .jpg format in order to take up less space. This is done by reducing image quality.
- **Class Balancing**: For the multiclass classifier, since the number of images per class may vary, to prevent over-representation of dominant classes during training, class balancing is performed by selecting an equal number of images from each class. This is done by determining the minimum number of samples across all classes and selecting that number of images per class. Doing this, the balanced dataset for the multiclass classifier was composed of 8763 of each class. On the other hand, for the binary classifier, the number of tumor images is divided by the number of the rest of the tissue types (8), and this resulting number is used to select an equal amount of images from each of these tissue types. These selected images are then combined into a single healthy class, while the tumor tissue remain as a separate class. This ensures that the total number of healthy and tumor images is balanced.
- **Dataset Splitting**: Once the images are balanced, for both classifiers' datasets, they are divided into training and validation subsets. The training set (Train) is used to adjust model weights, while the validation set (Validation) is used to evaluate the model's performance during training and prevent overfitting. Fig. 3 shows the train and validation data distribution of the final multiclass and binary datasets. Both datasets use an 80/20 distribution for train and validation, respectively.
- **Data Augmentation**: To expand the dataset and enhance the model's generalization ability while avoiding overfitting, different transformations are applied:
  - Random Rotations: within a range of $\pm 15°$.
  - Random Flips: with a 0.5 probability of flipping in the vertical axis.
  - Random Zooms: between 0.9 and 1.1 times the original size.

  These transformations enhance the model's robustness by introducing variability in the training data.

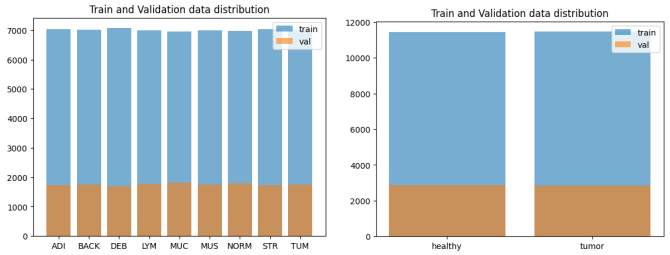

Fig. 3: Train and validation sets distribution for multiclass (left) and binary (right) tasks.

B. Classification Models

Both binary and multiclass classifiers utilize DenseNet121 architecture for feature extraction, differing only in the number of neurons in the output layer. The binary classifier has two output neurons to distinguish between healthy and tumor tissue, whereas the multiclass classifier has nine neurons corresponding to the nine tissue types in the dataset.

The models are trained using the cross-entropy loss function, which is well-suited for classification tasks. The Adam optimizer, with a learning rate of 1×10^{-5}, ensures stable convergence during training. Each model is trained for four epochs and performance is evaluated using the AUC-ROC metric to measure class separability and overall accuracy to assess general performance.

A graphical overview of the preprocessing and training steps is shown in Fig. 4, which illustrates the entire pipeline.

C. Experimental Setup

The data preprocessing and model training were implemented in Python, utilizing PyTorch and MONAI (Medical Open Network for AI). MONAI is specifically designed for deep learning applications in medical imaging, streamlining model development and deployment.

Experiments were conducted on a server at Universidad Carlos III de Madrid, equipped with an AMD Ryzen 7 3700X 8-Core processor, 32 GB of RAM, and an NVIDIA GeForce RTX 3090 GPU. Leveraging GPU acceleration significantly reduced training time compared to CPU-based training.

D. Model Explainability

Class Activation Maps (CAMs) [7] are a visualization technique used to interpret the decision-making process of convolutional neural networks. CAMs highlight the most relevant regions in an input image that contributed to a specific classification, CAMs provide valuable insights into model behavior. In this work, we employ CAMs to analyze the transition layers that connect each block of the DenseNet architecture, as well as the final ReLU activation layer, see Fig. 1. This approach allows us to better understand how the model processes features at different depths and which regions influence its predictions.

V. Evaluation and Results

The following section presents the evaluation methodology and results for the binary and the multiclass classifiers. This results are summarized in Table I in the Appendix. The performance of both models is assessed with different metrics such as Area Under the Curve (AUC), Receiver Operating Characteristic (ROC) curves and Class Activation Maps (CAMs). Moreover, confusion matrices are used to analyze the classification accuracy of the models.

Additionally, to provide a thorough evaluation, trials performed with earlier model versions are also included in this section. With this analysis, we aim to provide insights into the strengths and weaknesses of both classifiers and study possible areas that require further refinement. It is important to note that the reported results are based solely on the test split, without the use of cross-validation techniques. As such, these findings should be interpreted as an initial assessment rather than definitive conclusions. Future work should incorporate more robust validation strategies to ensure the generalizability and reliability of the models.

A. Results of the Binary Classifier

a) ***Traning Metrics****:* Fig. 5 illustrates the progression of the epoch average loss and the validation AUC during the training of the binary tissue classifier.

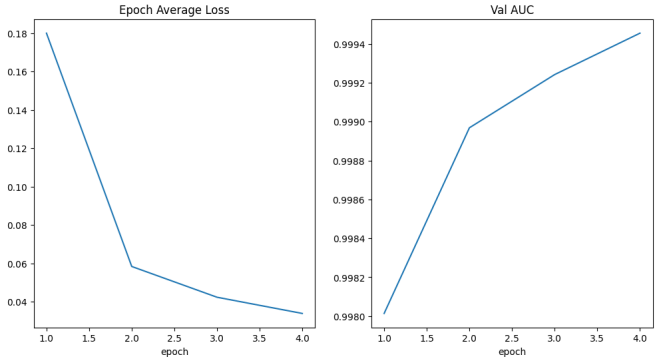

Fig. 5: Average loss and validation AUC evolution during training of the binary classifier.

Firstly, the graph on the left depicts the epoch-average loss, which starts at approximately 0.18 in the first epoch. A significant drop occurs in the second epoch, reducing the loss to around 0.06, indicating that the model is rapidly fine-tuning its ImageNet-pretrained weights to the medical classification task. This quick adaptation suggests that the pretrained features are effectively transferring to the new domain. From the second epoch onward, the loss continues to decrease more gradually, reaching approximately 0.02 by the fourth epoch. This suggests that while the model continues refining its parameters, improvements become progressively smaller, indicating convergence.

Secondly, the right graph plots the validation AUC. The curve starts at a slightly lower value, around 0.9980 in the first epoch, and increases as the training continues. The most

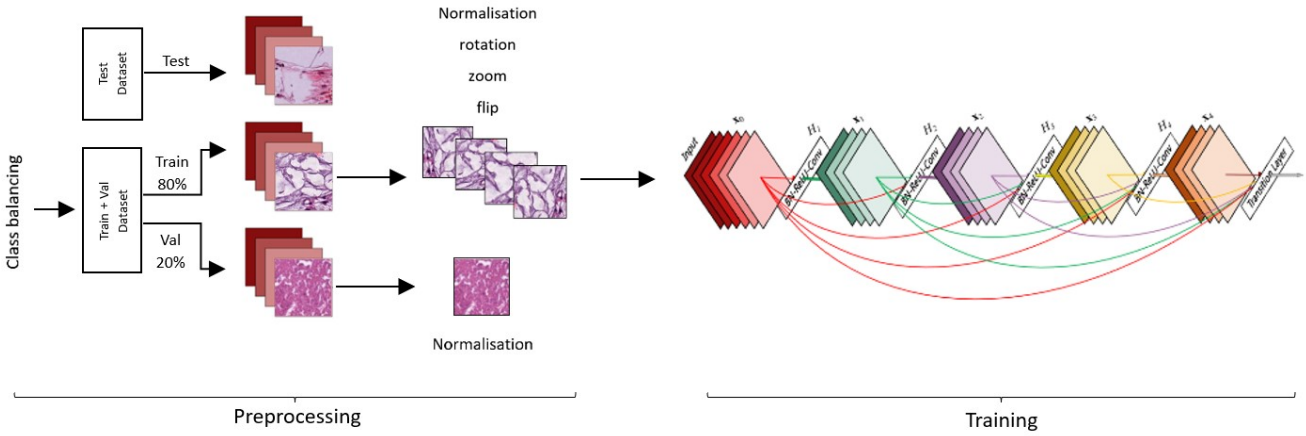

Fig. 4: Graphical abstract with the representation of preprocessing and training steps.

significant gain is observed in the second epoch, where the AUC reaches approximately 0.9990 and continues until it reaches a value of around 0.9994 in the fourth epoch. This increase indicates that the model is progressively improving its ability to distinguish between tumor and healthy tissues, achieving near-perfect performance on the validation set.

b) ***Confusion Matrix****:* Fig. 6 presents the confusion and normalized confusion matrices for the binary classifier, which differentiates between healthy and tumor tissue images.

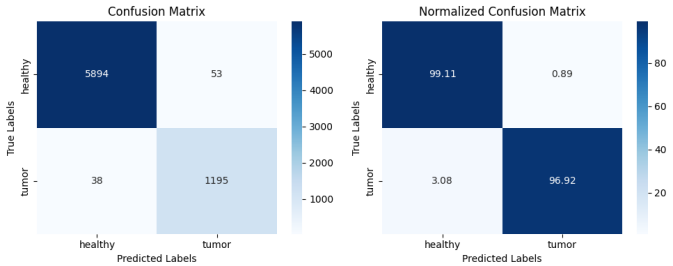

Fig. 6: Confusion and normalized confusion matrices of the binary classifier.

The confusion matrix indicates that the classifier correctly identified 5894 healthy samples and 1195 tumor samples, while it classified incorrectly 38 healthy samples as tumors and 53 tumor samples as healthy. The normalized confusion matrix further illustrates that the classifier achieved a high accuracy, correctly classifying 99.11% of healthy samples and 96.92% of tumor samples. However, 3.08% of tumors were misclassified as healthy, while only 0.89% of healthy samples were misclassified as tumors.

These results suggest that the model performs exceptionally well, especially in identifying healthy tissue, but makes a slightly worse classification for tumor samples. This should be taken into account for further improvements, since classifying tumors as healthy tissues is critical in the medical diagnosis field.

c) ***ROC and AUC Metrics****:* Fig. 7 shows the ROC curve calculated for this model. The area under the curve (AUC) is 0.9983 which indicates a significantly good performance of the model, meaning that it has an excellent ability to distinguish between the positive (tumor) and negative (healthy) classes.

The sensitivity value for a specificity of 85%, given Specificity $= 1 - FPR$, reaches close to 1 performance, $Sensitivity(85\%) = 0.9976$, as shown in Fig. 7.

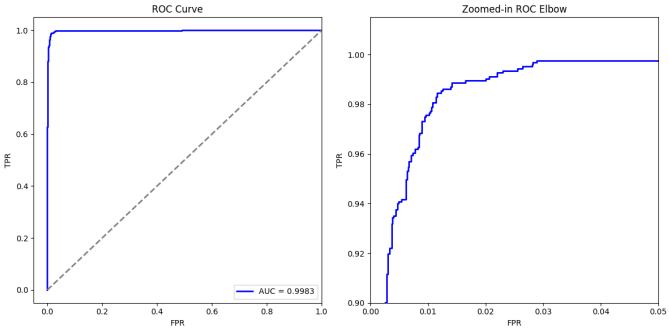

Fig. 7: ROC curve of binary classifier.

d) ***Explainability****:* Fig. 8 presents the activation maps generated using GradCAM++ [7] for the binary classification of healthy and tumor tissues.

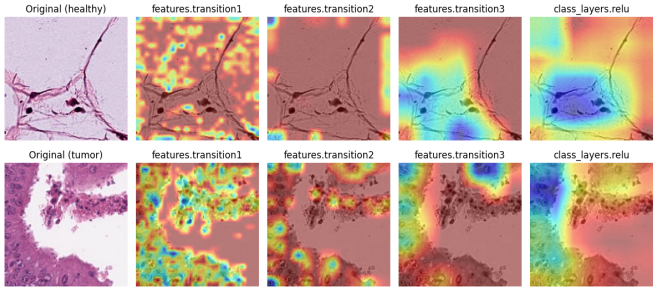

Fig. 8: Activation maps of correctly classified healthy class (top) and tumor class (bottom).

The original images in the first column show histological samples of both healthy and tumor tissues, while the other columns display the features that the model focuses on in each subsequent layer. Moreover, the last column shows the features and regions that the model finally takes into account for the classification.

For the healthy tissue of the upper image, structural regions with the highest activations occurring at connective intersections are highlighted, suggesting that the classifier focuses on specific tissue architecture when making predictions. More specifically, in the case of tumor tissue, the activation maps emphasize densely packed cellular regions, indicating that the model correctly identifies critical tumor-related features.

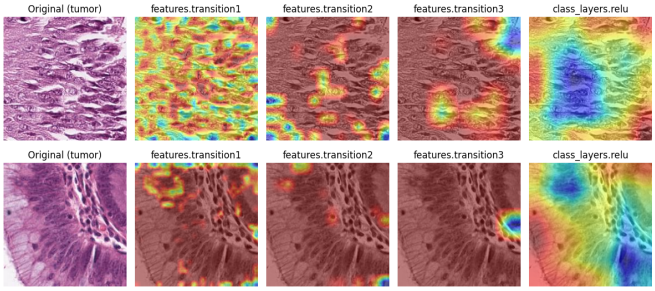

Fig. 9: Activation maps of incorrectly classified tumor tissues as healthy.

The activation maps of Fig. 9 show some examples of misclassified classes where the model has identified tumor images as healthy tissues. The images in the last column show the regions that the model has considered relevant in order to perform the classification. These activation maps could be a great tool to help a medical expert understand and visualize why the model has incorrectly classified the tissues.

B. Results of the Multiclass Classifier

a) ***Traning Metrics****:* Similarly to the explanation of Fig. 5, for the multiclass classifier, Fig. 10 shows the progression of the epoch average loss and the validation AUC.

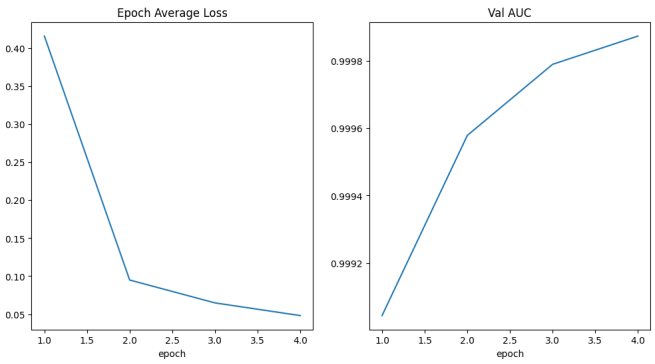

Fig. 10: Average loss and validation AUC evolution during training of the multiclass classifier.

Again, the greatest loss drop takes place in the second epoch, from a value of around 0.45 to 0.10, approximately. After that, the loss continues decreasing, reaching a value around 0.05 in the fourth epoch. On the other hand, regarding the validation AUC plot, the curve starts at a value of 0.9991 and increases up to almost 0.9999 in the last epoch.

These results show how the classifier improves its classification in each epoch, distinguishing between the nine tissue classes with very high performance.

b) ***Confusion Matrix****:* Fig. 11 show the classification performance across the different tissue classes, with the diagonal values indicating correctly classified instances.

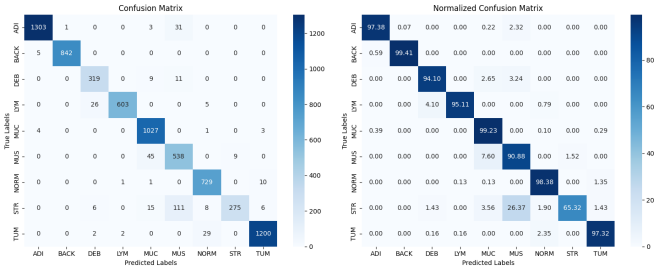

Fig. 11: Confusion and normalized confusion matrices of the multiclass classifier.

The majority of classes achieved high accuracy, as shown in the normalized confusion matrix. The ADI class was correctly classified in 97.38% of cases, with a minor misclassification into MUS (2.32%). The BACK class performed exceptionally well with an accuracy of 99.41%, showing negligible misclassification errors. Similarly, the MUC and TUM classes exhibited strong classification rates of 99.23% and 97.32%, respectively, with only slight confusion into other classes. The LYM class achieved 95.11% accuracy but was occasionally misclassified as DEB (4.10%). The DEB class, while obtaining 94.10% accuracy, showed misclassification into MUS (3.24%) and MUC (2.65%). Moreover, the MUS class had slightly lower accuracy of 90.88%, as compared to other classes, but showed some confusion with MUC (7.60%) and NORM (1.52%). The NORM class performed well with 98.38% accuracy, with minor confusion into TUM, LYM and MUC. However, the most significant misclassification was observed in the STR class, which only achieved 65.32% accuracy and was often misclassified as MUS (26.37%) and MUC (3.56%).

These results suggest that, while the overall classification model performs well, further refinement should be needed so as to improve the distinction between the STR, MUS, and MUC classes, which present the highest misclassification rates.

c) ***ROC and AUC Metrics****:* In Fig. 12, the left plot shows the ROC Curve, which evaluates the true positive rate against the false positive rate for different classification thresholds.

The curves for all tissue types are very close to the upper-left corner, indicating excellent classification performance. The AUC values are close to 1 for all classes, with the BACK class achieving a perfect score of 1, demonstrating the model's high capability in distinguishing between different tissue types.

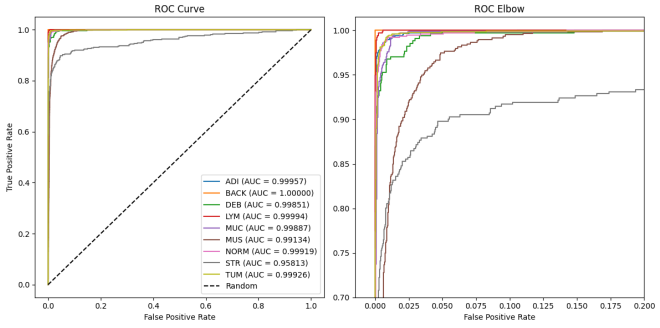

Fig. 12: ROC curves for the multiclass classifier.

The right plot, ROC Elbow, zooms in on the lower false positive rate region, highlighting how quickly the true positive rate reaches high values. The steep rise of most curves in this region further confirms the strong performance of the model. As can be seen, the weakest performance is in STR class, with an AUC of 0.95813, which is still an excellent result.

d) ***Explainability****:* Fig. 13 presents the activation maps obtained through CAM for different tissue types using a multiclass classifier. In the figure, each row corresponds to a distinct tissue class. The first column displays the original images, while the subsequent columns illustrate the activation responses at different stages of feature extraction within the neural network. To provide a concise analysis, we focus on four tissue types as examples.

To begin with, for adipose tissue (ADI), the activation maps highlight structural components, with early feature maps capturing general shapes and later layers emphasizing the cell boundaries and overall morphology. The final activation map shows a broad region of importance, reflecting the model's focus on tissue patterns.

Moreover, Lymphocyte (LYM) activations show a progressive focus on densely packed nuclei, with early layers detecting dispersed circular structures and later layers emphasizing clusters. The final class activation map highlights the most relevant cellular regions, demonstrating the model's ability to distinguish lymphocyte-rich areas.

Additionally, for mucosal tissue (MUC), early activation maps display dispersed attention across cellular structures, capturing prominent features such as glandular formations. As the layers deepen, the activation refines its focus on key regions, with the final activation highlighting specific areas indicative of mucosal architecture.

Finally, tumor tissue (TUM) shows intense activations early in the network, responding to irregular cellular formations. Intermediate layers refine these activations, focusing on regions of high nuclear density and disorganization. The final class activation highlights the most relevant tumor-associated features.

These visualizations show the classifier's ability to distinguish between different tissue types by emphasizing relevant morphological patterns at different stages of classification.

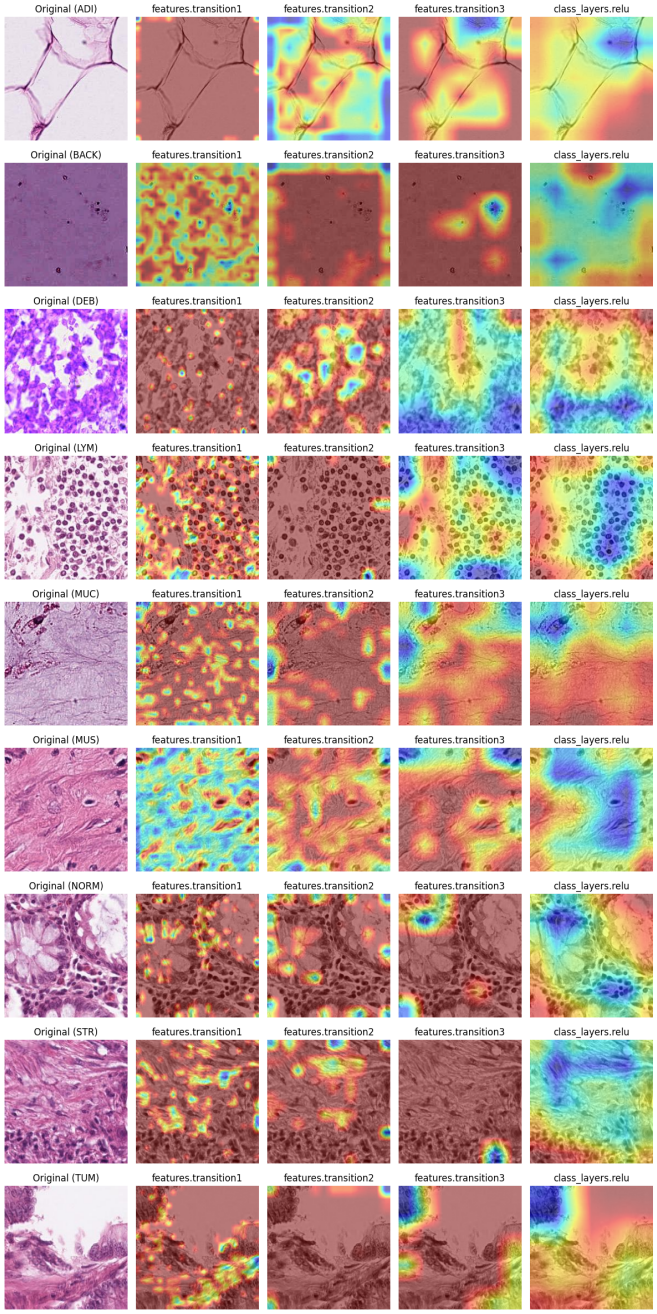

Fig. 13: Activation maps of each correctly classified class.

Fig. 14 illustrates some examples where STR images were misclassified as MUS by the model.

The activation patterns suggest that the model focuses on elongated fibrous structures, which are characteristic of both stromal and muscle tissues. The final class activation maps show a broad distribution of attention over regions that contain elongated structures, which might explain the misclassification. This suggests that the model struggles to distinguish between the overlapping morphological characteristics of STR and MUS, leading to confusion in classification. These misclassifications highlight the need for further

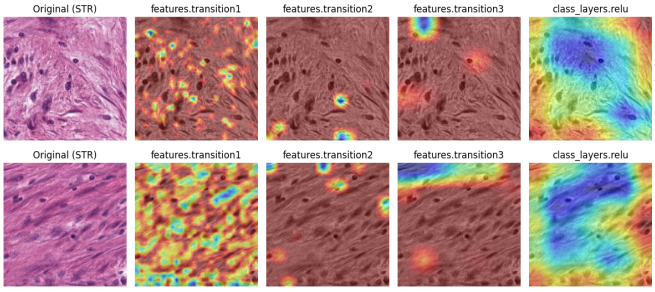

Fig. 14: Real STR images misclassified as MUS by the model.

refinement in feature representation to improve the model's ability to differentiate between similar tissue types. Moreover, these speculations should be confirmed by a specialist who can identify these cellular structures and can use the activation maps as a tool to see and understand the weaknesses of the multiclass classifier in order to refine it.

C. Additional Tests Conducted

To further evaluate the robustness of the classifiers, two additional experiments were performed, each with different training conditions.

For the binary classification task, an unbalanced model was trained to distinguish between tumor and healthy tissue, where the healthy class included images from eight different tissue types without balancing their representation. Despite this lack of balance, the results obtained were nearly identical to those of the final binary classifier, suggesting that the classifier was able to generalize well even under these conditions.

For the multiclass classification task, an earlier model was trained without balancing the classes, was not initialized with a pre-trained network, and used smaller batch sizes of 30. These differences contributed to lower classification performance compared to the final multiclass classifier, as can be observed from the results commented in this section.

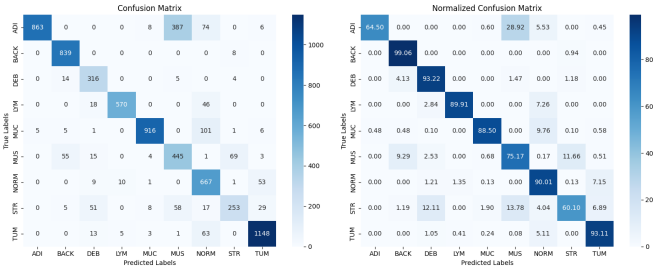

Fig. 15: Confusion and normalised confusion matrices for the earlier version of the multiclass classifier.

On the one hand, as seen in Fig. 15, the old multiclass classifier model obtained a worse classification of the different tissue classes, as compared to the final implemented model. More specifically, the greatest misclassifications were observed for the ADI (64.50%), MUS (75.17%) and STR (60.10%) classes. Regarding the ADI class, the model incorrectly identified ADI tissues as MUS in 28.92% of cases. The MUS class was classified as BACK (9.29%) and STR (11.66%). Finally, the STR class obtained the worst results with a misclassification of 12.11% as DEB, 13.78% as MUS and 6.89% as TUM classes. It is also important to note that the TUM class, although it was correctly classified in most cases, with an accuracy of 93.11%, was identified as NORM in 5.11% of the cases. This is quite relevant since, in the medical diagnosis field, correctly identifying tumor tissues is of the utmost importance, and the model should be especially careful with this class.

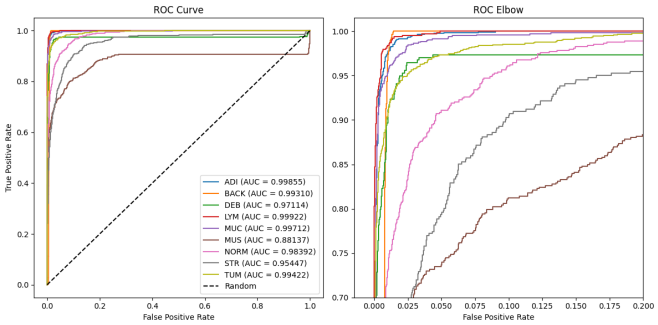

Fig. 16: ROC curves for the earlier version of the multiclass classifier.

On the other hand, Fig. 16 shows the ROC curves obtained from the earlier version of the multiclass classifier. The ROC curve of this model also shows a high classification capability for most classes, although some curves exhibit greater deviations from the perfect classifier. Although the model performs great with classes like ADI (AUC=0.99855), BACK (AUC=0.99310), LYM (AUC=0.99922) and MUC (AUC=0.99712), as can be observed in the ROC elbow on the right, the model has more difficulty in distinguishing classes like MUS (AUC=0.88137), STR (AUC=0.95447) and NORM (AUC=0.98392), which the final version of the model clearly improved (Fig. 12).

Overall, the final model outperforms the older model, achieving higher AUC values and steeper early rises in the curves, indicating a better classifier.

VI. Clinical Analysis

The proposed binary and multiclass classifiers have demonstrated promising performance in distinguishing different histological tissue types. However, before considering their direct application in clinical practice, several aspects must be thoroughly analyzed.

A key factor is the quality and diversity of the dataset used for training. To enhance clinical applicability and minimize biases, the dataset should reflect real-world variability in histological samples. This includes ensuring balanced and well-justified training, validation, and test set sizes, as well as incorporating a wider range of staining techniques, imaging protocols, and histopathological criteria from multiple institutions. Expanding the dataset further would improve model training, strengthening its generalization across different diagnostic settings.

Another important consideration is the transparency of the decision-making process. Understanding how the AI reaches its conclusions is essential for clinical adoption. Techniques such as Class Activation Mapping (CAM), as the ones we used in this work, could provide insight into the key image features influencing predictions, making the system more interpretable for medical professionals. Demonstrating that the algorithm follows diagnostic standards widely accepted in the field would also enhance trust and acceptance.

To assess clinical viability, the classifiers' performance must be compared against expert pathologists. While standard evaluation metrics such as accuracy, sensitivity, and specificity are useful, a more comprehensive assessment should measure the system's impact on diagnostic workflows, consistency, and workload reduction. Conducting comparative studies in real diagnostic settings would offer valuable insights into their practical utility.

A particularly important issue that requires further refinement is the misclassification of tumor tissues as healthy or as another type of tissue. These types of errors have severe consequences in the medical field, as they directly compromise patient health and safety. A failure to correctly identify malignant tissues could delay critical treatment, potentially leading to severe progression of the disease or even fatal outcomes. Therefore, improving sensitivity in detecting pathological cases should be a priority in future iterations of the model.

While the proposed models demonstrate strong potential, further validation, interpretability enhancements, expert comparisons, and regulatory approvals are required before they can be integrated into clinical practice. Addressing these challenges will be crucial for their effective application in medical diagnosis.

VII. Future Work

To further improve our models, we plan to enhance interpretability through occlusion sensitivity analysis and refine predictions using a sliding window inferer. Hyperparameter tuning and cross-validation will be employed to ensure robust generalization.

Additionally, addressing model robustness is crucial, with special attention needed to resolve misclassifications between similar tissue types, particularly between MUS (smooth muscle) and STR (stroma). Notably, the model exhibits sensitivity to image rotations, as illustrated in Fig. 17 of the Appendix, where applying rotations and/or flips to the same input image induces changes in model predictions. Increasing the number of training epochs and incorporating robustness techniques could help mitigate this issue. Furthermore, a benchmark comparison with state-of-the-art models will be conducted to validate performance. These improvements will enhance the model's clinical applicability and reliability.

References

[1] I. Castiglioni, L. Rundo, M. Codari, *et al.*, "Ai applications to medical images: From machine learning to deep learning," *Physica medica*, vol. 83, pp. 9–24, 2021.

[2] X. Liu, L. Song, S. Liu, and Y. Zhang, "A review of deep-learning-based medical image segmentation methods," *Sustainability*, vol. 13, no. 3, p. 1224, 2021.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[4] J. N. Kather, J. Krisam, P. Charoentong, *et al.*, "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study," *PLoS medicine*, vol. 16, no. 1, e1002730, 2019.

[5] J. N. Kather, N. Halama, and A. Marx, *100,000 histological images of human colorectal cancer and healthy tissue*, version v0.1, Zenodo, Apr. 2018. DOI: 10.5281/zenodo.1214456. [Online]. Available: https://doi.org/10.5281/zenodo.1214456.

[6] S. A. Hicks, I. Strümke, V. Thambawita, *et al.*, "On evaluation metrics for medical applications of artificial intelligence," *Scientific reports*, vol. 12, no. 1, p. 5979, 2022.

[7] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, Mar. 2018. DOI: 10.1109/wacv.2018.00097. [Online]. Available: http://dx.doi.org/10.1109/WACV.2018.00097.

[8] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, *Densely connected convolutional networks*, 2018. arXiv: 1608.06993 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1608.06993.

[9] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, 2015. arXiv: 1512.03385 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1512.03385.

APPENDIX

TABLE I: Summary of the test results for both classifiers.

Binary Classifier						
Class	**Precision**	**Recall**	**F1-score**	**AUC**	**Sensitivity**	**Training time**
healthy	0.9936	0.9911	0.9923	-	-	4 min 37 s
tumor	0.9575	0.9692	0.9633	0.9983	0.9976	
multiclass Classifier						
Class	**Precision**	**Recall**	**F1-score**	**AUC**	**Sensitivity**	**Training time**
ADI	0.9931	0.9738	0.9834	0.99957	0.99925	14 min 53 s
BACK	0.9988	0.9941	0.9964	1.00000	0.99646	
DEB	0.9037	0.9410	0.9220	0.99851	0.99705	
LYM	0.9950	0.9511	0.9726	0.99994	0.99842	
MUC	0.9336	0.9923	0.9621	0.99887	0.99903	
MUS	0.7786	0.9088	0.8387	0.99134	0.99831	
NORM	0.9443	0.9838	0.9636	0.99919	0.99865	
STR	0.9683	0.6532	0.7801	0.95813	0.92399	
TUM	0.9844	0.9732	0.9788	0.99926	0.99838	

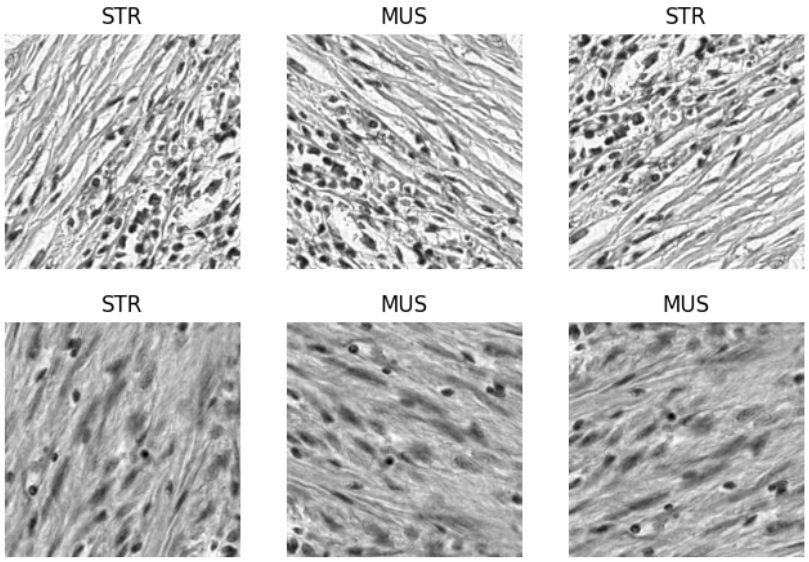

Fig. 17: Real STR images missclassified as MUS by the model when rotated or flipped.