

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/379536095>

A full pipeline to analyze lung histopathology images

Conference Paper · April 2024
DOI: 10.1117/12.3006708

CITATIONS
0

READS
39


9 authors, including:



Lluís Borrás Ferris
HES-SO Valais-Wallis

5 PUBLICATIONS 22 CITATIONS


SEE PROFILE



Niccolò Marini
University of Applied Sciences and Arts Western Switzerland

34 PUBLICATIONS 290 CITATIONS


SEE PROFILE



Alessandro Caputo
Università degli Studi di Salerno

102 PUBLICATIONS 765 CITATIONS

SEE PROFILE



Francesco Ciompi
Radboud University Medical Centre (Radboudumc)

110 PUBLICATIONS 17,454 CITATIONS

SEE PROFILE

A full pipeline to analyze lung histopathology images

Lluís Borràs Ferrís^a, Simon Püttmann^a, Niccolò Marini^a, Simona Vatrano^d, Filippo Frassetto^d,
Alessandro Caputo^d, Francesco Ciompi^e, Manfred Atzori^{ab}, and Henning Müller^{ac}

^aUniversity of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland

^bDepartment of Neuroscience, University of Padua, Padua, Italy

^cThe Sense Innovation and Research Center, Sion and Lausanne, Switzerland

^dPathology Unit, Gravina Hospital Caltagirone ASP, Catania, Italy

^eRadboud University Medical Center, Nijmegen, Netherlands

ABSTRACT

Histopathology images involve the analysis of tissue samples to diagnose several diseases, such as cancer. The analysis of tissue samples is a time-consuming procedure, manually made by medical experts, namely pathologists. Computational pathology aims to develop automatic methods to analyze Whole Slide Images (WSI), which are digitized histopathology images, showing accurate performance in terms of image analysis. Although the amount of available WSIs is increasing, the capacity of medical experts to manually analyze samples is not expanding proportionally. This paper presents a full automatic pipeline to classify lung cancer WSIs, considering four classes: Small Cell Lung Cancer (SCLC), non-small cell lung cancer divided into LUng ADenocarcinoma (LUAD) and LUng Squamous cell Carcinoma (LUSC), and normal tissue. The pipeline includes a self-supervised algorithm for pre-training the model and Multiple Instance Learning (MIL) for WSI classification. The model is trained with 2,226 WSIs and it obtains an AUC of 0.8558 ± 0.0051 and a weighted f1-score of 0.6537 ± 0.0237 for the 4-class classification on the test set. The capability of the model to generalize was evaluated by testing it on the public The Cancer Genome Atlas (TCGA) dataset on LUAD and LUSC classification. In this task, the model obtained an AUC of 0.9433 ± 0.0198 and a weighted f1-score of 0.7726 ± 0.0438 .

Keywords: Self-supervised, Weakly-supervised, MIL, Lung Cancer, Histopathology, CAD.

1. INTRODUCTION

The integration of digital processes into clinical histopathology is opening the door to Computer-Assisted Diagnostic (CAD) tools based on Deep Learning (DL) models capable of autonomously learning from clinical data. Nevertheless, several challenges still need to be addressed in the field.

Histopathology is the gold standard for the diagnosis of most cancer types.¹ Histopathology involves the analysis of tissue samples. Tissue samples include several tissue structures, ranging from microscopic entities (such as single-cell nuclei) to macroscopic components (such as tumor solid mass). The typical workflow in a histopathology laboratory starts with the collection of a sample (e.g. through a biopsy) and follows different procedures to obtain the final histologic sample that a pathologist will examine. Currently, in most hospitals, pathologists manually inspect slides using microscopes without the aid of automatic artificial intelligence algorithms.²

A Whole Slide Image (WSI) is a digitized slide that is scanned at high-resolution and stored in a multi-scale (pyramidal) format as shown in Fig. 1. Digital pathology is becoming increasingly integrated into some hospitals, with an additional step in the workflow involving the digitization of glass slides through the use of automated digital pathology scanners that offer magnification equivalent to a microscope. The acquisition of a WSI typically occurs up to x40 magnification level, resulting in images of over 100,000 pixels in each dimension, with a pixel size of $0.25 \mu\text{m}$.³

The availability of WSIs paved the way for the development of computational pathology domain. Computational pathology aims to develop automatic algorithms to analyze WSIs unleashing the power of digital pathology.

Further author information: (Send correspondence to L.B.F.: e-mail: lluis.borrasferris@hevs.ch.)

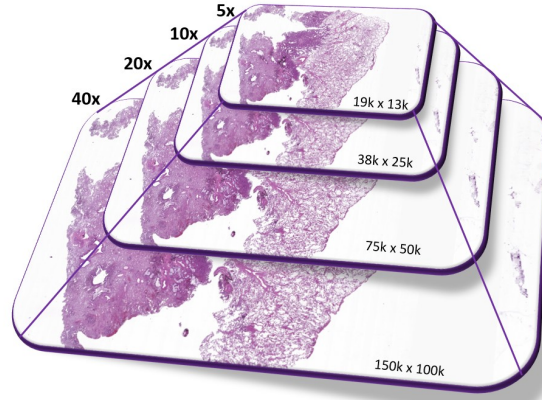


Figure 1. Example of digitized Whole Slide Image (WSI) scanned at 40x (0.25 $\mu\text{m}/\text{pixel}$) as high-resolution level and stored in a multi-scale (pyramidal) at four different magnifications.

Most of these algorithms are currently based on Machine Learning (ML), specifically on DL algorithms to improve the accuracy and efficiency of cancer diagnosis.⁴

Thanks to technological advancements and the improvement of the healthcare systems, the number of biopsies collected worldwide has been increasing over the years. However, the number of pathologists is not increasing equally with a consequent workload on the sector, as indicated by the ratio of pathologists per number of inhabitants. For example, in Europe a ratio of one pathologist per 32,018 inhabitants. Switzerland, the USA, and Canada have one pathologist per 35,355, 20,658, and 25,325 inhabitants, respectively.⁵

This work presents a full pipeline to classify different WSIs by using self-supervised pre-training and Multiple Instance Learning (MIL) for training, employing 2,226 WSIs from two different hospitals. The objective is the development of a CAD tool to help pathologists in the analysis of WSIs and reduce their workload with an increasing number of biopsies.

The paper is focused on the analysis of lung cancer WSIs. Lung cancer exhibits the highest mortality rate worldwide among all cancer types with an 18.0 Age-Standardized mortality Rate (ASR) per 100,000 inhabitants in 2020.^{6,7} The initial step in diagnosing lung cancer involves performing chest radiography on patients who exhibit symptoms associated with a possible tumor. In cases where radiography indicates positive results, a biopsy is performed on the area with abnormal lung findings.⁸ Consequently, histopathology is critical in determining a patient's prognosis and in identifying appropriate surgical and/or treatment interventions⁹

Lung cancer is classified into two primary groups: Non-Small-Cell Lung Cancer (NSCLC) and Small-Cell Lung Cancer (SCLC). NSCLC is further categorized into three subtypes: LUng ADenocarcinoma (LUAD), LUng Squamous cell Carcinoma (LUSC), and large-cell lung carcinoma. Among these, LUAD is the most common subtype and accounts for 50% of all NSCLC cases. Of all lung cancer cases, 80% are classified as NSCLC, while the remaining 20% are classified as SCLC.^{10,11} Accurate identification of the distinct categories and subcategories is essential for patient's prognosis. Treatment options can vary significantly depending on the cancer subtype.¹²

1.1 Contributions

In this work, an algorithm is proposed to perform a 4-class classification task among the 3 most prevalent lung cancer types SCLC, LUAD, and LUSC, and healthy tissue. The model is pre-trained in a self-supervised algorithm and trained using a weakly-supervised learning strategy by training a MIL model. This work presents the following contributions.

- An innovative pipeline is proposed that combines self-supervised pre-training and weakly-supervised training using MIL for the classification task of lung cancer between four different classes, the most prevalent three cancer types (SCLC, LUAD, and LUSC) and normal tissue.

- Evaluation of the performance using self-supervised learning to pre-train the model, compared with the model pre-train on ImageNet in the downstream classification task.
- The models are tested on The Cancer Genome Atlas (TCGA) public dataset to analyze their generalization capabilities.

2. STATE OF THE ART

Recent advances in computational pathology and DL techniques have demonstrated the potential to improve tumor histopathology evaluations. Due to their large size, WSIs are commonly divided into patches, as current Graphics Processing Units (GPU) cannot handle the entire WSI at its original size. The lack of large, annotated, datasets and data heterogeneity are still open challenges in computational pathology. For example, in a typical fully-supervised training strategy, pixel-wise annotations are needed, which is a time-consuming task for pathologists and very expensive. Several methods are proposed to overcome these problems such as weak-supervised learning as shown in Tab. 1.

Table 1. Review of state-of-the-art methods in the classification of histopathological lung cancer Whole Slide Images (WSI) highlighting the different training strategies for learning and the number of WSI used on the specific datasets. LUAD: LUng ADenocarcinoma, LUSC: LUng Squamous cell Carcinoma, NL: NormaL tissue, SCLC: Small-Cell Lung Carcinoma, PTB: Pulmonary TuBerculosis, OP: Organizing Pneumonia, AUC: Area Under de ROC Curve, TCGA: The Cancer Genome Atlas, ICGC: International Cancer Genome Consortium KMC: Kyushu Medical Centre, MH: Mita Hospital, TCIA: The Cancer Imaging Archive, DPGFLCD: Department of Pathology of the Georges François Leclerc Cancer Center in Dijon, UHC: University Hospital of Caen, TMUH: Taipei Medical University Hospital, WFH: Taipei Municipal Wanfang Hospital, SHH: Taipei Medical University Shuang-Ho Hospital, SYSU: First Affiliated Hospital of Sun Yat-sen University, SZPH: Shenzhen People’s Hospital dataset.

Paper	Training strategy	Dataset	Subtypes	Results	Preprocessing
13	Fully-supervised	TCGA	567 LUAD, 609 LUSC or 459 NL	AUCs of 0.993 tumour vs. NL, 0.950 LUAD vs. LUSC, 0.968 3-class	Patch
14	Fully-supervised	TCGA ICGC	427 LUAD 457 LUSC 87 LUAD 38 LUSC	AUC LUAD vs. LUSC 0.927 ± 0.004 , AUC LUAD vs. LUSC 0.842 ± 0.011	Patch
15	Weakly-supervised	4,054 KMC, 500 MH, 680 TCGA 500 TCIA	Lung carcinoma and non-neoplastic carcinoma	AUCs 0.975 KMC, 0.974 MH 0.988 TCGA, and 0.981 TCIA	Patch
16	Fully-supervised	DPGFLCD, UHC TCGA	66 nonLUSC 66 LUSC, 45 nonLusc 20 LUSC, 30 nonLUSC 30 LUSC	Accuracy 0.85 DPGFLCD UHC: 0.81 AUC, TCGA: AUC 0.78	Patch
17	Weakly-supervised	TCGA	55 LUAD and 55 LUSC	AUC of 0.902 ± 0.016 for lung	Patch
18	Fully-supervised	741 SYSU1, 318 SYSU2 212 SZPH and 422 TCGA	LUAD, LUSC, SCLC, PTB, OP and NL	AUCs 0.970 SYSU1, 0.918 SYSU2, 0.963 SZPH and 0.978 TCGA	Patch
19	Fully-supervised	1,723 KMC, 500 MH, and 905 TCGA	LUAD, LUSC SCLC and NL	AUCs 0.94 - 0.99 in LUAD, LUSC, SCLC and neoplastic vs. non-neoplastic	Patch
20	Self-supervised Weakly-supervised	TCGA	10,678 33 cancer types 1,008 LUAD and LUSC	Self-supervised pre-trained on 10,678 WSI AUC of 0.952 ± 0.021 LUAD vs LUSC	Scaling

2.1 Fully-supervised learning

Fully-supervised learning (strong supervision) relies on manual pixel-level annotations to train the DL models. That means that for every patch in a WSI, a label must be provided to train the model. It requires a group of pathologists to provide these manual pixel-level annotations, which are expensive and a very time-consuming task. These fully-supervised models achieve the best performances and are the most widely strategy used in the state-of-the-art models in lung cancer subtype classification.

2.2 Weakly-supervised learning

The line of research that investigates how to best use image-level diagnostic labels, is known as weakly-supervised learning. Without pixel-level annotations, weak supervision models approach the training of a model only using WSI-level annotations which better replicates the real scenario, where a pathologist provides only one diagnosis per image.¹⁵ These WSI-level annotations are noisy by nature because only a small portion of the patches are representative of the label.

MIL is the state-of-the-art method to solve the noisy annotations problem among all the weakly-supervised algorithms. In the MIL problem, instead of a single instance like in typical classification problems such as the one

presented using the ImageNet dataset, there is a bag of instances that represent one unique label.²¹ The instances should not depend on each other and their order within the bag should not be considered significant. These two strong definitions imply that the model must be permutation-invariant. Therefore, the permutation-invariant bag probability is computed using a scoring function for a set of instances that is a symmetric function.²² The score function is used to compute the bag probability and a permutation-invariant function referred to as MIL pooling ensures that this score function is a symmetric function by using commonly MIL pooling the max or mean operators.

To represent the bags of patches in the MIL frameworks two different strategies are studied to aggregate the instance-level features into a bag-level representation. These strategies aim to capture the key characteristics of the lung tissue samples and differentiate between the different cancer types and normal cells.

- **Instance-level aggregation:** One approach, is to build an instance-level classifier that returns scores for each patch. Then the individual scores are aggregated by MIL pooling (such as max pooling or average pooling). This pooling operation summarizes the information within each patch, capturing essential features associated with different cancer types.¹⁵
- **Embedding-level aggregation:** Alternatively, the instances are mapped to a low-dimensional embedding. Afterward, MIL pooling is used to obtain a bag representation independent of the number of patches in each bag.²²

2.3 Transfer learning

Transfer learning approaches leverage pre-trained models that have been trained on extensive datasets like ImageNet.²³ The primary objective is to take advantage of models that have already acquired a feature representation of a sizable image dataset. Consequently, the classifier layers of the network can be retrained, or in some cases, specific layers of the model can be unfrozen to learn representative features from the images in a new dataset. This process involves training the model on the targeted dataset to perform the new classification task.²⁴ All of the works, except the last one that uses self-supervised learning, presented in Tab. 1 take advantage of this strategy to load a pre-trained network trained on ImageNet and train only the classifier and in the specific lung cancer classification task.²³

2.4 Self-supervised learning

In the past years, unsupervised representation algorithms have gained prominence in the field of computer vision. Instead of relying on pre-training models with weights from for example ImageNet, these approaches aim to pre-train the models using the dataset's images, constructing tokenized dictionaries for unsupervised learning. For example, in Natural Language Processing (NLP), tokenization is the process of breaking down the text into smaller inputs, such as words, called tokens. A tokenized dictionary would contain these individual tokens as its entries, allowing for efficient lookup and analysis of specific words within the dictionary. In the computer vision domain building these dictionaries is an open challenge since the data exists in a high-dimensional space.

Ref. 25 addressed this challenge by introducing Momentum Contrast (MoCo), a technique that constructs dynamic, large, and consistent dictionaries using contrastive loss. In their work, they demonstrate that MoCo effectively narrows the gap between unsupervised and supervised representations in computer vision tasks such as object detection and segmentation, employing widely recognized datasets like PASCAL, VOC, and COCO.

Additionally, Ref. 26 proposed a straightforward algorithm for contrastive learning. Through their work on SimCLR, they highlighted the significance of data augmentation composition, the incorporation of a learnable nonlinear transformation between the representation and the contrastive loss, and larger batch sizes (4k - 8k batch size) together with more training steps. With these three important findings, they enhanced model effectiveness and achieved a new state-of-the-art on the ImageNet dataset.

Building upon these findings from SimCLR, the researchers at Facebook AI Research introduced MoCo v2,²⁷ which incorporated more aggressive data augmentation and a Multi-Layer Perceptron (MLP) projection head exhibited improved performance compared to the work of the Google research team on the ImageNet dataset. What's more, they show that with MoCo v2 is possible to process a large set of negative samples without

requiring large training batches and consequently powerful GPUs. In contrast, Moco v2 can run on a typical 8-GPU machine. Furthermore, Ref. 28 demonstrated in their work that leveraging MoCo v2 in a self-supervised learning framework effectively closed the gap between weakly-supervised and fully-supervised learning using histopathology images from the Camelyon16 dataset.

Ref. 20 uses a different strategy to pre-train the self-supervised model using a new approach with the potential of transformers called DINO and applying it to histopathology data.²⁹ They used DINO for pre-training with 10,678 WSIs from 33 different cancer types collected from the public TCGA dataset. Afterwards, they train a MIL model, using weak labels, for a binary classification task on 1,008 WSIs of LUAD and LUSC achieving an Area Under the ROC Curve (AUC) of 0.952 ± 0.021 .

2.5 Preprocessing strategy

WSIs are images characterized by an immense number of pixels, rendering it unfeasible to process them in their original size due to limitations posed by GPU hardware. The prevalent approach, employed by a majority of researchers working with histopathology images involves partitioning the images into smaller patches. These patches are subsequently utilized to train an ML model, enabling the model to learn a downstream task.

Alternatively, Ref. 20 instead of preprocessing the WSIs patching or resizing them, they take advantage of the potential of transformers to scale through different stages to learn representable features of these high-resolution images from lower to higher patch-level resolutions to capture information from individual cells to tissue microenvironment.

3. MATERIAL AND METHODS

3.1 Datasets

The WSIs selected from all three datasets are composed only of WSIs stained with H&E. The private datasets from Azienda Ospedaliera per l’Emergenza Cannizzaro Catania (AOEC) and Radboud University Medical Centre (RUMC) are imbalanced due to the characteristics of digital pathology workflows. With a higher number of WSIs with normal tissue (negative biopsies without malignancy) LUAD, being the most prevalent subtype among the lung cancer ones, followed by LUSC, with fewer examples, and SCLC, which is the less prevalent one.

The self-supervised model is pre-trained using all the data from AOEC with a total of 1,271 WSIs as shown in Tab. 2 resulting in 2,950,251 images after preprocessing. The MIL model is trained in two different scenarios, only using WSIs from AOEC and WSIs from both, AOEC and RUMC.

Table 2. Overview of the dataset composition. It includes lung images from digital pathology laboratories in Azienda Ospedaliera per l’Emergenza Cannizzaro Catania (AOEC) and Radboud University Medical Centre (RUMC). The training dataset is divided into train and validation using 5-fold cross-validation. Additionally, the model is tested on The Cancer Genome Atlas (TCGA) public dataset. SCLC: Small-Cell Lung Carcinoma, LUAD: Lung ADenocarcinoma, LUSC: Lung Squamous cell Carcinoma.

Source	SCLC	LUAD	LUSC	Normal	Total labels	Total images
Training dataset from two different private datasets:						
AOEC	53	601	353	237	1,244	1,225
RUMC	0	297	205	499	1,001	1,001
Total	53	898	558	736	2,245	2,226
Testing private datasets:						
AOEC	17	16	9	14	46	46
RUMC	0	29	18	45	92	92
Total	17	45	27	59	138	138
Testing public dataset:						
TCGA	0	530	506	0	1,036	1,036

The model trained on data only from AOEC is tested in two scenarios, using a set of 46 WSIs from AOEC and 138 WSIs from both hospitals. The model trained with WSIs from the two hospitals is tested only on this second set. Additionally, both models were tested on TCGA public dataset composed of 1,036 LUAD and LUSC WSIs^{30,31} from 5 different centers in the USA (Washington University, University of Pittsburgh, University of North Carolina, Lahey Hospital & Medical Center and Roswell Park).

3.2 Pipeline

Fig. 2 provides a comprehensive overview of the pipeline developed to classify histopathology images of the lung into four distinct classes, LUAD, LUSC, SCLC, and healthy WSIs. First, all the WSIs are preprocessed to extract the patches. These patches are utilized for pre-training a self-supervised model based on MoCo v2.²⁷ The objective is to pre-train a feature extractor specific to the histopathology lung data. Finally, using the feature vectors from the previous step, the MIL model employs weak labels to train the classification model.

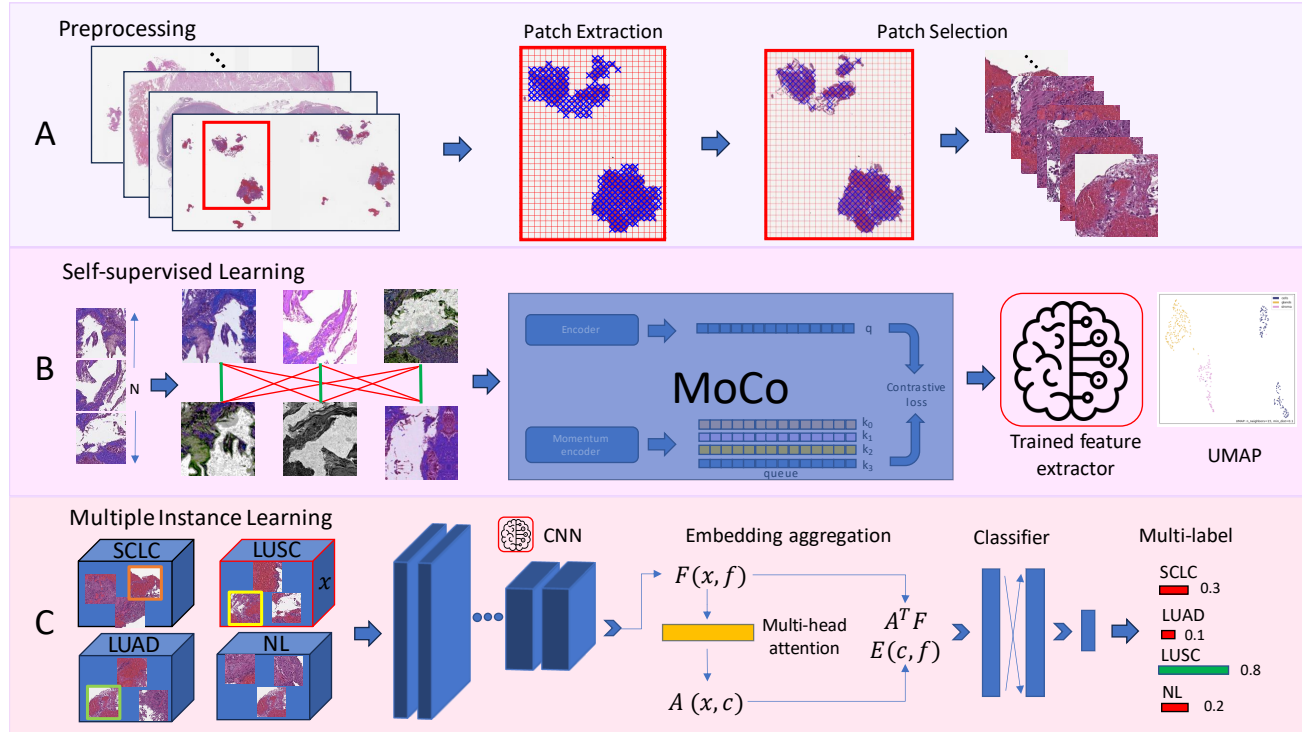


Figure 2. Full Pipeline to train the lung cancer subtype classification model. A: Preprocessing of the Whole Slide Images (WSI) for patch extraction. B: Self-supervised learning to pre-train the feature extractor capturing high-level concepts directly from the histopathological lung patches using Momentum Contrastive Learning (MoCo v2). C: Training of a Multiple Instance Learning (MIL) model for the lung cancer classification task among, LUAD, LUSC, SCLC, and Normal tissue (NL), using a multi-label strategy. UMAP: Uniform Manifold Approximation and Projection, CNN: Convolutional Neural Network, x : number of patches per WSI, f : feature vector, c : number of classes.

3.2.1 Preprocessing

The preprocessing stage is composed of two components presented in Fig. 3. The initial task involves extracting patches from each WSI. Each WSI is divided into multiple patches, ranging from a few hundred to thousands, depending on the WSI's size and the amount of tissue contained in each slide. The subsequent task involves selecting patches from the extracted set that contain representative tissue information from the respective WSI.

Patch Extraction: The initial preprocessing step is shown in Fig. 3.A, it involves dividing each Whole Slide Image (WSI) into patches to meet the GPU requirements for training the model at a specific magnification level using the PyHIST tool.³² After consultation with an expert pathologist and comparing with state-of-the-art methods for the lung classification task is chosen to downsample all the WSIs at a 10x magnification level to extract patches, with a tile size of 256x256. This process involves three main steps:

1. Mask extraction separating effectively the foreground (tissue) from the background content of the WSI.
2. Creation of a grid of non-overlapping tiles overlaid on the mask, followed by an evaluation to determine whether each tile belongs to the foreground or background.

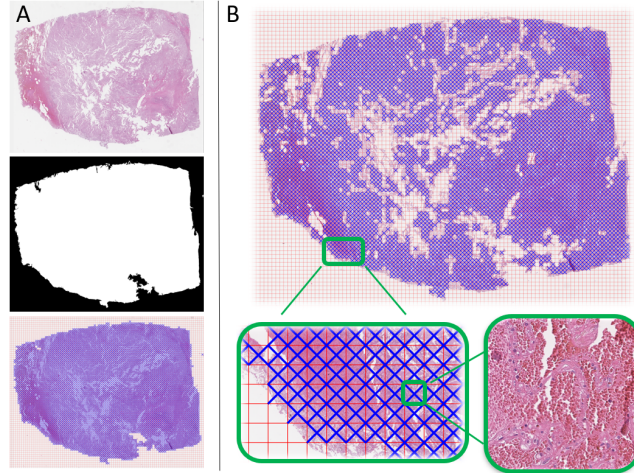


Figure 3. Preprocessing steps for one Whole Slide Image (WSI). A: Patch extraction: by generating a mask and selecting the patches that fall into this mask. B: Patch selection: To remove wholes, patches with insufficient information, and errors from the extracted patches.

3. Selection of patches by choosing tiles that fall within the extracted mask at the desired magnification level.

Patch Selection: The patch extraction algorithm achieves a low number of true negatives but produces a significant number of false positives, as shown in Fig. 3.A, in the bottom image. Upon careful inspection of these false positives, several common characteristics are identified. The WSIs may contain macro holes, and all WSIs contain numerous micro holes that PyHIST cannot accurately identify as background. These patches are mainly composed of white pixels. Some WSIs contain text (labels in some glass slides) that does not contain tissue information and can be considered potential confounders. These texts are written in black letters.

The primary objective of the patch selection is to reduce the number of false positives by removing unnecessary patches that would introduce noise in the model training, while still preserving patches with significant tissue information as illustrated in Fig. 3.B. From the histogram, if the number of bins with tissue information exceeds 50% of the total number of pixels the patch is retained. Contrarily, if the number of pixels with important content falls below 50% of the total number of pixels, the patch is discarded.

3.2.2 Self-supervised learning

Self-supervised algorithms aim to learn a stronger data representation, exploiting data its-self, without the need for annotations.²⁷ Employing a feature representation of the images extracted from a pre-trained model, rather than simply relying on weights from a pre-train model using transfer learning, can potentially produce better results in the downstream task.²⁸

By leveraging recent advancements in self-supervised learning, a feature extractor is trained specifically on histopathology images of the lung. MoCo v2 is chosen as a reference from the FAIR research group as self-supervised architecture.²⁷ The primary objective is to train an encoder using contrastive learning. The encoder learns to associate images within the dataset by performing a dictionary look-up task summarized in Fig. 2.B. ResNet34 is implemented as the backbone for the self-supervised model,³³ loaded from the PyTorch framework.³⁴

In the field of lung cancer classification, state-of-the-art methods commonly utilize a frozen feature extractor that has been pre-trained on ImageNet, however, this is identified as a potential limitation. An alternative approach using self-supervision would be to train a feature extractor specifically for histopathology lung images. In this work, the results are compared from two different feature extractors in the task of lung cancer classification using weakly-supervised learning with MIL. The goal is to compare the results obtained from training the MIL model using the features extracted from the self-supervised model (MoCo v2) and using the same model but extracting features using the pre-trained model with weights from ImageNet.

3.2.3 Multiple Instance Learning

The goal is to develop a MIL approach for the multi-label classification of lung cancer based on histopathology lung images. The dataset consists of WSIs of lung tissue samples, where each WSI represents a bag containing multiple image patches or instances. The patches can be classified into four classes: SCLC, LUAD, LUSC, or normal cells. The objective is to train a model that can accurately predict the presence of these cancer types within each WSI, which allows for the identification and classification of different cancer types in a single image. The MIL framework is suitable for this task as the exact location and quantity of cancerous cells within a WSI may vary. Using the MIL approach completely avoids the need for a pathologist to assign instance-level labels like in the fully-supervised approach and trains a weak model using only the WSI-level labels.

The score function chosen determines two different approaches to modeling the label probability as described in Section 2.3, the instance-level and embedding-level aggregation. The embedding-level aggregation is selected as proposed by²¹ as the ground truth of all the instances is not known. Therefore, the first approach will be potentially trained insufficiently, and the prediction will be probably lower than the second approach. Moreover, each WSI has a different number of patches, in the instance-level aggregation this would result in inconsistent matrix sizes for the attention score. Instead in the embedding aggregation, the matrix is always fixed to an embedded score (c,f), being c the number of classes and f the number of features coming from the feature extractor as shown in Fig. 2.C. Additionally, instead of using the typical max or averaging max pooling layer in this work is employed a trainable attention-based MIL pooling layer as described in Ref. 22.

The result of the attention-based MIL pooling layer is passed to the classifier (a linear fully-connected layer) that outputs the model prediction for each class. In this work, the idea is to be able to classify if necessary more than one class at a time, as in the real case scenario when more than one cancer can be present in the same WSI. In the multi-label problem, the final prediction is transformed into a probability using the sigmoid function and applied individually for each class, instead of the typical softmax applied in multi-class scenarios. If the probability is higher than 0.5 for a given class the model prediction is positive and negative otherwise.

3.3 Experimental set-up

3.3.1 K-fold cross-validation

For the training step of the MIL model, the WSIs coming from both hospitals were divided into train and validation following k-fold cross-validation. The training and validation sets are carefully split to avoid having images from the same patient in the different sets. The goal is to prove the robustness of the model to the selected training data.

The training data is divided into k (k=5) groups. In each training iteration, the data from k-1 groups are utilized to train the CNN, while the remaining group is used for validation. This division is performed at the patient level to ensure that images are not shared between the training and validation partitions. Subsequently, the CNN is evaluated on the test partition, and the average and standard deviation of the k models are reported.

3.3.2 Hyperparameters

Self-supervised pre-training: The self-supervised model is trained using all the patches available from the AOEC dataset (2,950,251 images). With this amount of data, a single experiment takes around 100 hours using an Nvidia A100 80GB. An initial Learning Rate (LR) of 0.03 and a MultiStepLR scheduler decrease the LR in epochs 3, 6, and 11 by a gamma factor of 0.5. As the optimizer, Adam from PyTorch was chosen, and a batch size of 256. In both cases, the last fully connected layer has the same size resulting in a 128-feature vector as output. The temperature, τ is set to 0.07 and the queue to 32,768. Strong transformations are applied for the data augmentation²⁶ using the Albumentations Python library.³⁵ With a probability of 0.5, the following transformations are chosen: random resize crop, vertical and horizontal flips, a random rotation of 90°, hue value saturation, color jitter, elastic transformations, grid distortion, blurring, optical distortion, histogram equalization, and with a probability of 0.2 converting the images to grey.

Weakly-supervised training: To train the MIL model, the 128-feature vector is loaded for all patches per WSI and only the MIL pooling attention-based and the classifier are trained. Training a model lasts around five hours for the five models using the 5-fold cross-validation strategy using an Nvidia A100 80GB. As a criterion,

The Binary Cross Entropy with Logits loss function (BCEWithLogitsLoss) with weights (0.824, 0.477, 0.809, and 0.742 for SCLC, LUAD, LUSC, and normal tissue, respectively) was selected to work with the multi-label paradigm and to address the class imbalance of the dataset as shown in Tab. 2. The final model was trained using a batch size of 512, and the Adam optimizer with an initial LR of 0.0003 was modified at training time using the CosineAnnealingLR scheduler with a T_max of 10 and eta_min of 0.00003.

3.4 Evaluation

3.4.1 Self-supervised model

The performance of the self-supervised model is evaluated from two different points of view. Qualitatively, the model’s performance before implementing it in the downstream classification task. The main idea is to interpret if the model is learning concepts from the patches. An expert pathologist selected patches from 10 different WSIs that contain cells, glands, or stroma. The feature vectors of these patches are computed from the already-trained feature extractor. Afterward, a Principal Component Analysis (PCA) is performed to reduce the number of features from 128 to 20 components. Finally, compute a dimension reduction using Uniform Manifold Approximation and Projection (UMAP)³⁶ to reduce the 20 components from the PCA to 2 dimensions for each patch. The goal is to evaluate if the model is extracting separable feature vectors and learning that patches are different. Quantitatively, the result using the self-supervised model is compared in the lung cancer classification downstream task against the pre-trained model on ImageNet.

3.4.2 Classification task

The imbalance in lung cancer subtype prevalences within the dataset highlights the dominance of LUAD and outlines the rationale behind using specific metrics for comprehensive model evaluation in a multi-label scenario where more than one class could be positive for the same WSI. To overcome these challenges the Receiver Operating Characteristic (ROC) curve is computed for each class and the average-micro ROC curve as a global metric of the model. Together with the ROC, the average micro-AUC is computed for each class. The idea is to understand the balance between the true positive rate and the false positive rate at different thresholds. The AUC provides a global metric of the performance of the model individually for each class and globally with the average micro-AUC. The f1-score is also evaluated which gives a global idea of the precision/recall metrics of the model.³⁷ For the final evaluation of the model, all the metrics are calculated on the test set with WSIs from different patients never seen in the training phase.

3.5 Qualitative evaluation

The idea is to provide a visualization tool for qualitative evaluation that overlaps the attention of the score for a given class in the WSI. In the MIL training, a multi-head attention layer is used for the MIL pooling. This layer provides an attention score for all the extracted patches in a WSI per class. Through this process, the pathologist would be able to understand better in which regions the model is predicting a given class. The results are presented in the form of heatmaps¹⁷ to interpret if the model is correctly looking at where the malignancy is present on the histopathology slide of the lung. Moreover, in clinical practice, the heatmaps are a potential tool to help the pathologist not only receive a prediction from the model but also present in the form of a report the prediction of the model together with this qualitative evaluation.

4. RESULTS

This section presents the results obtained in the 4-class classification task among lung adenocarcinoma, lung squamous cell carcinoma, and small-cell lung carcinoma as cancer subtypes and normal healthy tissue. The section is divided into three subsections to present the different studies performed on this work:

- Classification performance between using a self-supervised model as a feature extractor or a model trained using the weights directly from ImageNet. Additionally, the generalization capabilities of the trained models are analyzed by conducting a test study on data from the TCGA public dataset.
- Qualitative evaluation by computing heatmaps to show where the model is looking on the WSIs and compare it with where a pathologist finds the lung cancer pathology.

4.1 Classification performance

4.1.1 Test on private datasets

Tab. 3 presents the results of the model trained on data from AOEC and tested on AOEC, trained on AOEC and tested in data from AOEC and RUMC and the model trained on data from AOEC and RUMC and tested also in both datasets. In each section, the models are pre-trained using self-supervised learning in comparison with the same model pre-trained on ImageNet. The self-supervised model obtains an f1-score of 0.5945 ± 0.0749 when trained on data only from AOEC. On the contrary, the model pre-trained on ImageNet obtains an f1-score of 0.5175 ± 0.0627 . When trained using both datasets, AOEC and RUMC the performance remains similar

Table 3. Results of the lung cancer subtype classification using the model trained on the Azienda Ospedaliera per l’Emergenza Cannizzaro Catania (AOEC) dataset and the model trained on both AOEC and Radboud University Medical Centre (RUMC) datasets. The table shows the metrics tested on AOEC or AOEC and RUMC and compares the performance of the self-supervised pre-training model and the model pre-trained on ImageNet. SCLC: Small-Cell Lung Carcinoma, LUAD: LUNG ADenocarcinoma, LUSC: LUNG Squamous cell Carcinoma, AUC: Area Under de ROC Curve.

Pre-training	AUC SCLC	AUC LUAD	AUC LUSC	AUC Normal	micro-AUC	weighted f1-score
Train and test on AOEC:						
ImageNet	0.7766 ± 0.0369	0.7176 ± 0.0642	0.8093 ± 0.0547	0.5500 ± 0.0410	0.7264 ± 0.0305	0.5175 ± 0.0627
Self-supervised (AOEC)	0.83333 ± 0.0316	0.7744 ± 0.1192	0.8275 ± 0.0430	0.6808 ± 0.0903	0.8024 ± 0.0450	0.5945 ± 0.0749
Train on AOEC and test on AOEC and RUMC:						
ImageNet	0.7506 ± 0.065	0.6812 ± 0.0357	0.8007 ± 0.0462	0.7242 ± 0.0570	0.6603 ± 0.0469	0.5314 ± 0.0389
Self-supervised (AOEC)	0.7779 ± 0.0540	0.67985 ± 0.0352	0.7574 ± 0.0180	0.7234 ± 0.0718	0.6604 ± 0.0493	0.5068 ± 0.0342
Train and Test on AOEC and RUMC:						
ImageNet	0.8784 ± 0.096	0.7497 ± 0.0268	0.8764 ± 0.0247	0.8446 ± 0.0071	0.8596 ± 0.0143	0.6380 ± 0.0148
Self-supervised (AOEC + RUMC)	0.8825 ± 0.0712	0.7457 ± 0.0267	0.8428 ± 0.0171	0.8468 ± 0.0130	0.8558 ± 0.0051	0.6537 ± 0.0237

4.1.2 Test on the public TCGA dataset

The results of the models tested on the public TCGA dataset are illustrated in Tab. 4. The model trained only on data from AOEC and the model trained using the AOEC and RUMC datasets are presented using both, self-supervised pre-training and pre-training on ImageNet. The self-supervised version obtains an f1-score of 0.7726 ± 0.0438 and the model pre-trained on ImageNet an f1-score of 0.7737 ± 0.0259 on WSIs coming from TCGA public dataset not used during training time.

Table 4. Results of the lung cancer subtype classification of both models, trained in the Azienda Ospedaliera per l’Emergenza Cannizzaro Catania (AOEC) dataset and trained on the AOEC and Radboud University Medical Centre (RUMC) datasets and tested on The Cancer Genome Atlas (TCGA) public dataset. SCLC: Small-Cell Lung Carcinoma, LUAD: LUNG adenocarcinoma, LUSC: LUNG Squamous cell Carcinoma, AUC: Area Under de ROC Curve.

Pre-training	AUC SCLC	AUC LUAD	AUC LUSC	AUC Normal	micro-AUC	weighted f1-score
Train on AOEC:						
ImageNet	1.0 ± 0.0	0.8754 ± 0.0081	0.8639 ± 0.0191	1.0 ± 0.0	0.9215 ± 0.0323	0.7212 ± 0.073
Self-supervised (AOEC)	1.0 ± 0.0	0.8464 ± 0.0290	0.8735 ± 0.0370	1.0 ± 0.0	0.8762 ± 0.0205	0.6688 ± 0.0143
Train on AOEC and RUMC:						
ImageNet	1.0 ± 0.0	0.8861 ± 0.0178	0.8875 ± 0.0168	1.0 ± 0.0	0.9448 ± 0.0078	0.7737 ± 0.0259
Self-supervised (AOEC + RUMC)	1.0 ± 0.0	0.8818 ± 0.0163	0.8856 ± 0.0179	1.0 ± 0.0	0.9433 ± 0.0198	0.7726 ± 0.0438

4.2 Clustering proficiency

UMAPs of the self-supervision model and the model pre-train on ImageNet are calculated to visualize the differences in the capabilities to cluster different patch types. The UMAPs are computed from the features extracted from the self-supervised model and the model pre-train on ImageNet. The UMAPs are computed on 384 patches (135 cells, 158 glands, and 91 stroma) selected by an expert pathologist.

Qualitatively, as shown in Fig. 4 the self-supervised model can almost perfectly cluster the three types of patches into clear separable regions. Contrarily, the UMAP of the pre-train model on Image-Net does not separate the different clusters as precisely as the self-supervised model. The UMAP of the model pre-train on Imagenet has some overlaps on the regions between cells and stroma and glands and stroma while the self-supervised model can separate the different patches containing cells, glands, and stroma without overlapping areas.

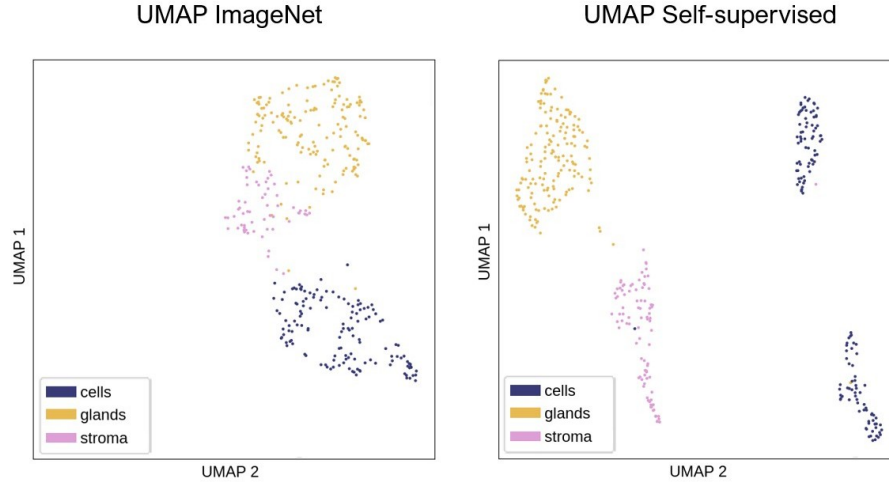


Figure 4. Uniform Manifold Approximation and Projection (UMAP) for dimension reduction computed on 384 patches (135 cells, 158 glands, and 91 stroma), selected by an expert pathologist using the pre-trained model from ImageNet in comparison with the self-supervised pre-trained model.

4.3 Qualitative evaluation: Heatmaps

For qualitative evaluation of the MIL model performance, a heatmap is computed on the model with the best performance. The model is pre-trained using self-supervised learning and trained on the AOEC and RUMC datasets. Fig. 5 shows that the model obtains a higher attention score in two of the three areas where the cancer is located as pointed out by the manual annotations from an expert pathologist.

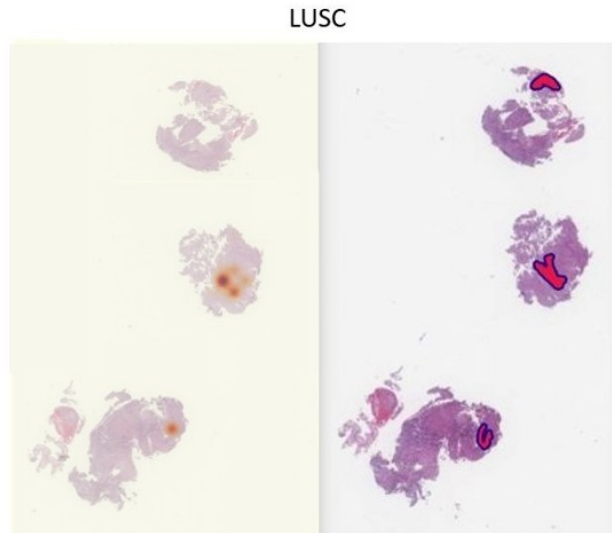


Figure 5. Heatmap computed using the final trained model (left side) compared with the annotations from an expert pathologist (right side) of a Whole Slide Image (WSI) diagnosed with Lung Squamous cell Carcinoma (LUSC).

5. DISCUSSION

The goal of this work was to develop a full pipeline to classify, using lung WSIs, between the three most prevalent cancer types, lung adenocarcinoma, lung squamous cell carcinoma, small-cell lung carcinoma, and normal tissue. Because of the size of the WSIs, composed of billions of pixels, the datasets are preprocessed splitting the WSIs into non-overlapping patches of size 256x256 at a 10x magnification level. Afterward, the model is pre-trained with self-supervised learning using Momentum Contrastive Learning (MoCo v2). Finally, the MIL model with an embedding aggregation approach is trained to perform the classification task using weakly-supervised learning. The performance of the model is evaluated in three different scenarios following the section 4.

The accuracy of the model can be influenced by the amount of heterogeneous data common in the computational pathology domain where big differences are found in the WSIs from the presence of different H&E stains to differences in the characteristics of the scanners. Two scenarios are presented: training the model only using data from AOEC and training the MIL model using data from both hospitals, AOEC and RUMC. As illustrated in Tab. 3 the performance when testing on both hospitals improved from a f1-score of 0.5068 ± 0.0342 when trained on AOEC to a f1-score of 0.6537 ± 0.0237 when trained in both datasets. The model effectively improved the performance because it was trained with more heterogeneous data combining both datasets for training.

In the validation of the self-supervision learning strategy is evaluated quantitatively the performance of the model against the same model pre-trained on ImageNet. In the first scenario, the model was trained only using the AOEC dataset. The self-supervision outperformed the model pre-trained on ImageNet with a gap of almost 0.1 on the weighted f1-score. However, when both models were trained using both datasets (AOEC and RUMC) the results were similar. The self-supervised version obtained an f1-score of 0.6537 ± 0.0237 in comparison to the model pre-trained on ImageNet with a f1-score of 0.6380 ± 0.0148 .

To evaluate the generalization capabilities of the trained models in the AOEC dataset and the AOEC and RUMC datasets, they were tested on the public TCGA dataset. The best results were in the models trained on both datasets and using the two different strategies, the self-supervised version and the pre-trained on ImageNet, resulting in a weighted f1-score of 0.7726 ± 0.0438 and 0.7737 ± 0.0259 , respectively. These results point out the good generalization capabilities of both models performing good predictions on an unseen diverse dataset from another country. Additionally, both models trained on data from AOEC and RUMC surpass both models trained only on the AOEC dataset. This also supports the fact that a model trained on more heterogeneous data can perform predictions more accurately on unseen data than its counterpart.

For the qualitative evaluation of the networks, a tool was developed called heatmaps. These heatmaps are computed from the attention scores coming from the multi-head attention of the MIL model for each class of a given WSI. On these heatmaps, only the attention scores of the predicted class are computed and compared with the manual annotations from an expert pathologist. As shown in Fig. 5 the model accurately gives high scores for the LUSC class to the patches that are in the region similar to the manual annotations of the pathologist. Of the three areas where the pathologist indicates that there is the presence of LUSC two are well localized for the model and no false positives are given.

Fully-supervised learning is usually the best approach in terms of performance for training models to classify lung histopathological data as presented in Section 2 with the work of.^{13,18,19} The major drawbacks of this approach are the need for pixel-level annotations for training which is a very time-consuming task and that this approach does not simulate the real scenario where one label (or more if more than one malignancy is present) is reported per WSI. Nevertheless, our model with weakly-supervised learning has similar results to the fully-supervised model presented by¹⁴ and surpasses the work of.¹⁶ Moreover, the self-supervised model achieves this good performance in a 4-class classification task while the two papers mentioned above only present a binary classification. Among the weakly-supervised strategies, our model surpasses the performance of and¹⁷ with an AUC of 0.902 ± 0.016 . Both works used transfer learning as a pre-train strategy in comparison with the self-supervision presented in this work.

Recent findings on the computer vision field with the work of²⁹ with DINO, and more specifically for histopathological images,²⁰ with HIPT, showing the potential of self-supervised learning to improve the prediction of ML models. They show that using Vision Transformers (ViT) is possible to obtain better feature representation of the images than using the architectures proposed by²⁷ with MoCo v2 and²⁶ with SimCLR

using CNNs. The idea behind HIPT is to use a scaling strategy in two stages using two consecutive ViT. First patching pre-training using patches of 256x256 and on top of this another ViT that performs a region pre-training of size 4,096x4,096 using the features coming from the first stage. Finally, these regions are used as feature extractors to feed a MIL model that performs the downstream classification task. The only drawback is that training HIPT,²⁰ uses a dataset with 10,678 WSIs, with a total of 433,779 regions of 4,096x4,096 pixels that take 7.7 TB of space. These specifications need powerful GPUs to be able to handle the training size.

As proven in this work, the training of MIL models potentially takes advantage of self-supervised pre-training.²⁰ using HIPT obtain an AUC of 0.952 ± 0.021 in the binary classification between LUAD and LUSC using 1,008 WSIs from the TCGA dataset. In our case for the 4-class classification presented in this work, the MIL model obtains an AUC of 0.9448 ± 0.0078 but AUCs of 0.8818 ± 0.0163 and 0.8856 ± 0.0179 in LUAD and LUSC, respectively. HIPT obtains better performance by using the scaling ViT strategy.

6. CONCLUSION

This paper presents a full pipeline to analyze lung histopathological images using both self-supervision and transfer learning approaches for the pre-training of the model and MIL to train the model using only weak labels without the need for pixel-level annotations. The goal of the model is to classify the lung WSIs between the three most prevalent lung cancer types, LUAD, LUSC, SCLC, and healthy tissue.

Both models obtain similar results when trained in data coming from two different private datasets AOEC and RUMC. However, as presented with the UMAPs, the self-supervision is capable of better clustering similar patches thanks to the refined pre-training using lung histopathological images. Moreover, the models are capable of generalizing as they obtain good results in the public TCGA dataset that is not used for training.

The models show a great performance using weakly-supervised learning with a MIL architecture which allows training the model only with WSI-level annotations avoiding the need for pathologists to obtain pixel-level annotations to train the fully-supervised counterparts which is a very time-consuming task.

Finally, thanks to the attention-layer used in the MIL approach a powerful is presented, the heatmaps to indicate in which patches the model is looking to obtain the final prediction.

ACKNOWLEDGMENTS

This project has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No. 825292 (ExaMode, <http://www.examode.eu/>).

REFERENCES

- [1] Tornillo, L. and Franco, R., “The role of histopathology in cancer diagnosis and prognosis,” *Frontiers Research Topics* (2022).
- [2] Fischer, A. H., Jacobson, K. A., Rose, J., and Zeller, R., “Hematoxylin and eosin staining of tissue and cell sections,” *CSH protocols* **2008**, pdb-prot4986 (2008).
- [3] Marini, N., Otálora, S., Podareanu, D., van Rijthoven, M., van der Laak, J., Ciompi, F., Müller, H., and Atzori, M., “Multi_scale.tools: a python library to exploit multi-scale whole slide images,” *Frontiers in Computer Science* **3**, 684521 (2021).
- [4] Marini, N., Marchesin, S., Otálora, S., Wodzinski, M., Caputo, A., Van Rijthoven, M., Aswolinskiy, W., Bokhorst, J.-M., Podareanu, D., Petters, E., et al., “Unleashing the potential of digital pathology data by training computer-aided diagnosis models without human annotations,” *NPJ digital medicine* **5**(1), 102 (2022).
- [5] Märkl, B., Füzesi, L., Huss, R., Bauer, S., and Schaller, T., “Number of pathologists in germany: comparison with european countries, usa, and canada,” *Virchows Archiv* **478**, 335–341 (2021).
- [6] Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., and Bray, F., “Cancer statistics for the year 2020: An overview,” *International journal of cancer* **149**(4), 778–789 (2021).

- [7] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F., “Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: a cancer journal for clinicians* **71**(3), 209–249 (2021).
- [8] Travis, W. D., Brambilla, E., Noguchi, M., Nicholson, A. G., Geisinger, K. R., Yatabe, Y., Beer, D. G., Powell, C. A., Riely, G. J., Van Schil, P. E., et al., “International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma,” *Journal of thoracic oncology* **6**(2), 244–285 (2011).
- [9] American Cancer Society, “Cancer statistics center.” <http://cancerstatisticscenter.cancer.org>. Accessed April 12, 2023.
- [10] Goldstraw, P., Ball, D., Jett, J. R., Le Chevalier, T., Lim, E., Nicholson, A. G., and Shepherd, F. A., “Non-small-cell lung cancer,” *The Lancet* **378**(9804), 1727–1740 (2011).
- [11] Van Meerbeeck, J. P., Fennell, D. A., and De Ruysscher, D. K., “Small-cell lung cancer,” *The Lancet* **378**(9804), 1741–1755 (2011).
- [12] Kumar, V., Abbas, A., and Aster, J. C., [*Robbins basic pathology e-book*], Elsevier Health Sciences (2017).
- [13] Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyo, D., Moreira, A. L., Razavian, N., and Tsirigos, A., “Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning,” *Nature medicine* **24**(10), 1559–1567 (2018).
- [14] Yu, K.-H., Wang, F., Berry, G. J., Re, C., Altman, R. B., Snyder, M., and Kohane, I. S., “Classifying non-small cell lung cancer types and transcriptomic subtypes using convolutional neural networks,” *Journal of the American Medical Informatics Association* **27**(5), 757–769 (2020).
- [15] Kanavati, F., Toyokawa, G., Momosaki, S., Rambeau, M., Kozuma, Y., Shoji, F., Yamazaki, K., Takeo, S., Iizuka, O., and Tsuneki, M., “Weakly-supervised learning for lung carcinoma classification using deep learning,” *Scientific reports* **10**(1), 9297 (2020).
- [16] Le Page, A. L., Ballot, E., Truntzer, C., Derangère, V., Ilie, A., Rageot, D., Bibeau, F., and Ghiringhelli, F., “Using a convolutional neural network for classification of squamous and non-squamous non-small cell lung cancer based on diagnostic histopathology images,” *Scientific Reports* **11**(1), 23912 (2021).
- [17] Lu, M. Y., Williamson, D. F., Chen, T. Y., Chen, R. J., Barbieri, M., and Mahmood, F., “Data-efficient and weakly supervised computational pathology on whole-slide images,” *Nature biomedical engineering* **5**(6), 555–570 (2021).
- [18] Yang, H., Chen, L., Cheng, Z., Yang, M., Wang, J., Lin, C., Wang, Y., Huang, L., Chen, Y., Peng, S., et al., “Deep learning-based six-type classifier for lung cancer and mimics from histopathological whole slide images: a retrospective study,” *BMC medicine* **19**, 1–14 (2021).
- [19] Kanavati, F., Toyokawa, G., Momosaki, S., Takeoka, H., Okamoto, M., Yamazaki, K., Takeo, S., Iizuka, O., and Tsuneki, M., “A deep learning model for the classification of indeterminate lung carcinoma in biopsy whole slide images,” *Scientific Reports* **11**(1), 8110 (2021).
- [20] Chen, R. J., Chen, C., Li, Y., Chen, T. Y., Trister, A. D., Krishnan, R. G., and Mahmood, F., “Scaling vision transformers to gigapixel images via hierarchical self-supervised learning,” in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 16144–16155 (2022).
- [21] Ilse, M., Tomczak, J., and Welling, M., “Attention-based deep multiple instance learning,” in [*International conference on machine learning*], 2127–2136, PMLR (2018).
- [22] Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J., “Deep sets,” *Advances in neural information processing systems* **30** (2017).
- [23] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L., “Imagenet: A large-scale hierarchical image database,” in [*2009 IEEE conference on computer vision and pattern recognition*], 248–255, Ieee (2009).
- [24] Cheplygina, V., de Bruijne, M., and Pluim, J. P., “Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis,” *Medical image analysis* **54**, 280–296 (2019).
- [25] He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R., “Momentum contrast for unsupervised visual representation learning,” in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 9729–9738 (2020).
- [26] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G., “A simple framework for contrastive learning of visual representations,” in [*International conference on machine learning*], 1597–1607, PMLR (2020).

- [27] Chen, X., Fan, H., Girshick, R., and He, K., “Improved baselines with momentum contrastive learning,” *arXiv preprint arXiv:2003.04297* (2020).
- [28] Dehaene, O., Camara, A., Moindrot, O., de Lavergne, A., and Courtiol, P., “Self-supervision closes the gap between weak and strong supervision in histology,” *arXiv preprint arXiv:2012.03583* (2020).
- [29] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A., “Emerging properties in self-supervised vision transformers,” in [*Proceedings of the IEEE/CVF international conference on computer vision*], 9650–9660 (2021).
- [30] Kirk, S., Lee, Y., Kumar, P., Filippini, J., Albertina, B., Watson, M., Rieger-Christ, K., and Lemmerman, J., “The cancer genome atlas lung squamous cell carcinoma collection (tcga-lusc) (version 4) [data set].” The Cancer Imaging Archive (2016). <https://doi.org/10.7937/K9/TCIA.2016.TYGKKFMQ>.
- [31] Albertina, B., Watson, M., Holback, C., Jarosz, R., Kirk, S., Lee, Y., Rieger-Christ, K., and Lemmerman, J., “The cancer genome atlas lung adenocarcinoma collection (tcga-luad) (version 4) [data set].” The Cancer ImagingF Archive (2016). <https://doi.org/10.7937/K9/TCIA.2016.JGNIHEP5>.
- [32] Muñoz-Aguirre, M., Ntasis, V. F., Rojas, S., and Guigó, R., “Pyhist: a histological image segmentation tool,” *PLoS computational biology* **16**(10), e1008349 (2020).
- [33] He, K., Zhang, X., Ren, S., and Sun, J., “Deep residual learning for image recognition,” in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 770–778 (2016).
- [34] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., “Pytorch: An imperative style, high-performance deep learning library,” *Advances in neural information processing systems* **32** (2019).
- [35] Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., and Kalinin, A. A., “Albumen-tations: fast and flexible image augmentations,” *Information* **11**(2), 125 (2020).
- [36] McInnes, L., Healy, J., and Melville, J., “Umap: Uniform manifold approximation and projection for dimension reduction,” *arXiv preprint arXiv:1802.03426* (2018).
- [37] Wu, X.-Z. and Zhou, Z.-H., “A unified view of multi-label performance measures,” in [*international conference on machine learning*], 3780–3788, PMLR (2017).