Unsupervised Deep Learning for Clustering Tumor Subcompartments Histopathological Images in Non-Small Cell Lung Cancer

Matheus de Freitas Oliveira Baffa[a,1], Nadine Sarah Schaadt[b,2], Friedrich Feuerhake[b,3], and Thomas M. Deserno[c,4]

[a]Department of Computing and Mathematics, University of São Paulo. Ribeirão Preto, Brazil.
[b]Institute for Pathology, Hannover Medical School. Hannover, Germany.
[c]Peter L. Reichertz Institute for Medical Informatics, Technical University of Braunschweig. Braunschweig, Germany.

ABSTRACT

Lung cancer ranks as the second leading type of cancer globally. The predominant forms are non-small cell carcinoma and small cell carcinoma. Lung cancer is diagnosed based on biopsies, surgical resection specimens, or cytology. Standard work-up of histopathological lung cancer samples includes immunohistochemistry (IHC) staining, which allows the visualization of specific proteins expressed on cellular structures in the sample. Computational methods play a key role in evaluating immune cells and detecting immune checkpoint markers in distinct tissue sections. These analyses are essential for designing targeted immuno-oncology treatments. Current pathological analysis of these samples is both time-intensive and challenging, often hinging on the expertise of a few highly skilled pathologists. An automated solution using computer vision, has the potential to assist pathologists in achieving a more accurate and consistent diagnosis. Our paper introduces a novel approach that leverages deep unsupervised learning techniques to autonomously label regions within IHC-stained samples. We developed a robust clustering model by actively extracting radiomic features from small patches within whole slide images. To achieve this, we utilized Self-Organizing Maps, a type of artificial neural network trained using unsupervised learning to produce a low-dimensional, discretized representation of the input space of the training samples. This method allowed us to effectively cluster the extracted features, enhancing our ability to analyze and interpret the complex data presented in the whole slide images. Our findings indicate that unsupervised clustering is a promising approach to meet the increasing demand for high-quality annotations in the emerging field of computational pathology.

Keywords: Computational Pathology, Immunohistochemistry Staining, Lung Cancer, Unsupervised Deep Learning, Self-Organizing Maps

1. INTRODUCTION

Lung cancer is characterized by uncontrolled growth of neoplastic cells derived from pulmonary tissues such as bronchial epithelium. Besides association with smoking, which accounts for approximately 85% of all cases,[1] and some less common environmental factors, its etiology is largely unknown. According to the World Cancer Research Fund, in 2020, there were 2,206,771 diagnosed cases of lung cancer worldwide, with a higher prevalence in men.[2] Non-small cell lung cancer (NSCLC) represents the majority of lung cancer cases, encompassing approximately 80% to 85% of all new diagnoses, while small cell lung cancer (SCLC) accounts for the remaining 10% to 15% of cases.[3]

1 Matheus de Freitas Oliveira Baffa, mbaffa@usp.br

2 Nadine Sarah Schaadt, schaadt.nadine@mh-hannover.de

3 Friedrich Feuerhake, feuerhake.friedrich@mh-hannover.de

4 Thomas Deserno, thomas.deserno@plri.de

Medical Imaging 2024: Imaging Informatics for Healthcare, Research, and Applications,
edited by Hiroyuki Yoshida, Shandong Wu, Proc. of SPIE Vol. 12931, 129310Z
© 2024 SPIE · 1605-7422 · doi: 10.1117/12.3009180

Lung cancer is diagnosed based on biopsies, surgical resection specimens, or cytology. Standard work-up of histopathological lung cancer samples includes immunohistochemistry (IHC) staining, which allows the visualization of specific proteins expressed on cellular structures in the sample. Expert pathologists then analyze these IHC-stained slides, concentrating on cellular patterns and potential indications of malignant cells.[4] This thorough assessment is critical to confirm the diagnosis, determine the subtype, and evaluate prognostic and predictive markers that can guide the therapy. Currently, relevant markers include the immune checkpoint protein PD-L1. In addition to markers expressed in tumor cells, current biomarker strategies increasingly consider patterns of immune cell infiltration that reflect the patient's response to the tumor and may have additional potential for biomarker discovery.

Integrating advanced image processing and analysis techniques has the potential to make the pathology workflow more efficient, increase accuracy, and discover relevant image features beyond the limits of human perception. However, there are significant challenges in employing supervised learning methods for this purpose. One major hurdle is the need for large amounts of annotated data, which demands considerable time and expertise from pathologists. Moreover, variations in slide preparation and staining can make the curation of consistent training data difficult.

Unsupervised learning, which seeks patterns in data without pre-existing labels, offers a unique advantage in medical imaging scenarios where extensive labeled datasets are challenging to obtain. In the context of non-small cell lung cancer, one relevant task is the evaluation of tumor-, stroma-, and immune cells. Computational methods are now at the forefront of these evaluations. For instance, before lung cancer biopsies are subjected to Next-Generation Sequencing (NGS) analyses, as part of "molecular pathology", there is a need to precisely quantify the tumor cell content. Furthermore, the identification and scoring of immune cells, especially in the context of the expression of immune checkpoint markers in specific tissue regions or cell populations, has been increasingly recognized.

This paper explores the use of unsupervised learning to detect and cluster different tissue types in IHC-stained images. By employing Self-Organizing Maps (SOM) and deep feature extraction techniques, we adopted methods originally developed for analysis of radiology images ("radiomics") for our aim to enhance the analysis of IHC-stained lung cancer samples.

2. RELATED WORKS

In the field of computational pathology, various methodologies have been explored by researchers. Coundray *et al.*[5] classified NSCLC and predicted mutations based on histological images. Their process begins with segmenting the digital histological images, which are obtained from samples stained with (H&E), into smaller sections or 'tiles'. Non-relevant background tiles are then filtered out. The remaining relevant images serve as training data for an Inception V3 Convolutional Neural Network (CNN). This network is employed to create a classification model capable of identifying coherent structures within the samples. The method was able to detect cancerous structures accurately.

Similarly, Graham *et al.*[6] focused on detecting lung cancer in histology images using patch-level summary statistics. Their approach begins with the extraction of patches from H&E stained samples. These patches are then utilized to train a ResNet-32 CNN. The method is designed to classify identified structures within the samples as non-diagnostic, lung adenocarcinoma, or lung squamous cell carcinoma.

In 2021, Carrillo-Perez *et al.*[7] introduced a method that combines RNA-Seq probability data with histology images for the detection of non-small cell lung cancer. This approach differs from others in the literature by integrating omics and histology data, aiming to provide a more comprehensive analysis of the material and thereby achieving a refined classification accuracy. The process involves dividing the sample into tiles, which are then used for training a ResNet-18 CNN. Additionally, they train a support vector machine (SVM) to identify patterns within the RNA-Seq data. The probabilities generated by each classifier are subsequently fused, leading to the final prediction for the sample. The authors reported that this combined method demonstrates improved efficiency when compared to classifications using whole slide images (WSI) or RNA-Seq data alone.

Li *et al.*[8] focused on the identification of specific subtypes, including adenosquamous carcinoma (ASC), lung squamous cell carcinoma (LUSC), and SCLC. Utilizing a dataset of 121 WSIs, they applied the Relief algorithm to extract relevant features from these images. These extracted features were then used to train an SVM for the classification of the mentioned subtypes. Their results indicated varying levels of classification accuracy: they achieved a 73.91% accuracy rate in distinguishing between LUSC and ASC, an 83.91% accuracy rate for LUSC and SCLC classification, and a 73.67% accuracy rate in differentiating between ASC and SCLC.

Baranwal *et al.*[9] conducted an evaluation of various Deep CNNs architectures, including VGG19, ResNet-50, Inception-ResNet, and DenseNet-121, to classify histopathology images of lung cancer. The initial phase of their research involved preprocessing the dataset, which included steps such as data preparation, normalization, cleaning, and formatting. After that, the histopathology images were processed through the aforementioned CNN architectures. Among these, DenseNet-121 demonstrated superior performance. It achieved an overall accuracy of 99.08%, alongside a sensitivity rate of 99.08%.

In contrast with existing research in the field, our study identifies a notable gap in the exploration of different staining techniques, particularly in the context of unsupervised methods. Immunohistochemistry staining, which we focus on, offers distinct features that could provide new insights into the problem. Our work, therefore, makes significant contributions in several key areas:

- **Development of an Unsupervised Methodology:** Our paper introduces an unsupervised method designed to cluster various structures within the samples. This approach represents a shift from the more commonly used supervised methods in histological image analysis.

- **Interpretability through Categorization:** In our approach, each cluster within the samples is categorized by a pathologist. This process enhances the understanding and interpretability of the various structures present in the samples.

3. METHODS

The methodology of our study was structured around a four-step process, encompassing patch generation (Fig. 1 – Step 1), radiomic feature extraction (Fig. 1 – Step 2), feature normalization (Fig. 1 – Step 3), and clustering using self-organizing maps (SOM) (Fig. 1 – Step 4). Initially, patches of size 512x512 pixels were generated from WSI. This stage included the identification and removal of background elements from the images. Subsequently, radiomic features were extracted for each patch and systematically catalogued in a CSV file, along with their corresponding sample identification, to maintain an accurate record of the origin of each patch. This step was critical for ensuring reliable sample tracking during the analysis. The study design involved using different samples for training and evaluating the clusters. The extracted features were then normalized using the MinMax method. The final stage involved training a SOM algorithm to create clusters. This was executed with varying configurations, resulting in maps of 2x2, 3x3, and 4x4 dimensions, corresponding to 4, 9, and 16 clusters, respectively. The subsequent sections provide a detailed description of each of these steps.

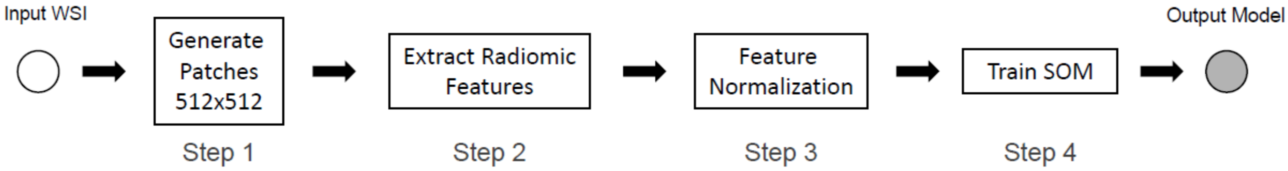


Figure 1. Method overview from whole slide image (WSI) using self-organizing Maps (SOM) to output.

3.1 Data Acquisition and Preprocessing

The dataset used in this study was obtained from the Hannover Medical School and encompassed 12 WSIs from a previously published study.[10] These slides presented a wide variety of tissue types, including both cancerous and adjacent lung and pleural regions, as well as multiple sub-types (i.e., squamous and adenocarcinoma) NSCLC.

To prepare the tissue samples for imaging, each specimen was sliced into sections measuring 3 μm in thickness. Following this, the sections underwent an automated IHC staining protocol (Ventana Benchmark XT, Roche

Diagnostics, Tucson, AZ). Through this staining procedure, T-cells (CD3, company) were distinctly colored brown via the use of 3,3 - Diaminobenzidine (DAB), while B-cells (CD20) were rendered in red using Neufuchsin. Other cell structures, such as cell nuclei, were highlighted using a hematoxylin counterstain that is not specific for a particular cell lineage.

The WSIs were captured using a scanner (Aperio AT2, Leica Microsystems, Wetzlar, Germany) at a resolution of 0.253 μm/pixel, equivalent to a 40x magnification, ensuring the detailed representation of cellular structures and interactions within the presented tissue samples.

Following the acquisition of the WSIs, we extracted patches of dimensions 512×512 pixels (Fig. 2). Utilizing the OpenSlide for Python framework, we slid over the entire WSI, cropped and isolated individual patches. To ensure quality, each patch underwent a background check. Background regions in WSIs often manifest as near-uniform grayscale, typically with average pixel intensities exceeding 200. By converting patches to grayscale and applying a threshold of 200 for average intensity value, we were able to effectively identify and exclude such background patches from subsequent analyses.

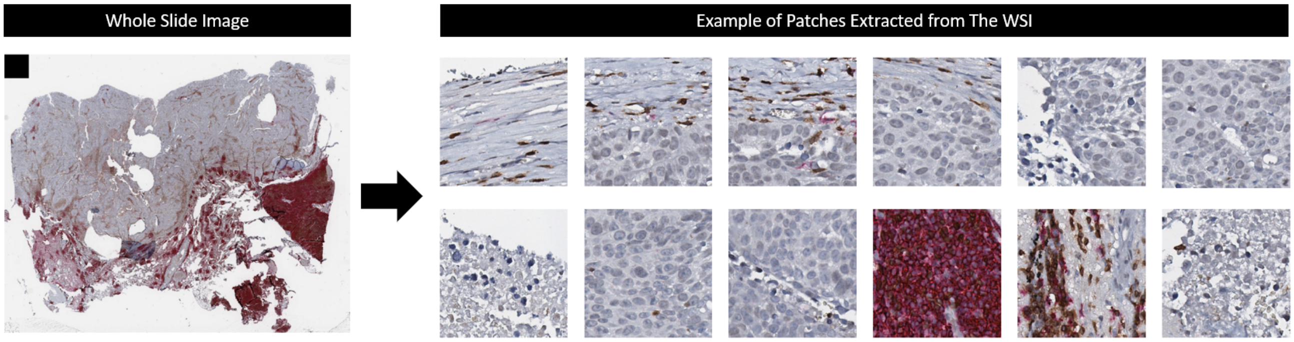


Figure 2. Example of patches extracted from WSI.

3.2 Feature Extraction and Normalization

Radiomics stands for the extraction of a large number of image features from radiographic images, turning them into mineable data.[11] The idea behind radiomics is to utilize the comprehensive quantitative details embedded in medical images, which might often be imperceptible to the human eye. The data-driven nature of radiomics allows it to capture subtle intratumoral heterogeneity which can be important for diagnosis, prognosis, and prediction of treatment response.[12] With the surge in computational power and the increased sophistication of machine learning algorithms, radiomics is positioned at the forefront of leveraging medical imaging for precision medicine.

Our feature extraction process followed the principles of radiomics. Using the PyRadiomics library,[13] we extracted a detailed set of features from our patches. This included nine shape 2D features to capture the structural intricacies, 18 first order statistics detailing basic pixel value properties, 22 features from the gray level co-occurrence matrix (GLCM) representing textural properties, 16 from the gray level run length matrix (GLRLM) identifying colinear voxels with consistent gray levels, 16 from the neighbouring gray tone difference matrix (NGTDM) highlighting gray level differences, and 14 from the gray level dependence matrix (GLDM) measuring gray level dependencies. In total, 95 features were extracted from each patch and were systematically stored in a patient-based CSV file.

To ensure the features were on a consistent scale, we applied MinMax normalization to our dataset. This normalization technique scales each feature to a specified range, between zero and one. This standardized the feature range and improved the convergence speed of algorithms that rely on gradient computation.

3.3 Clustering Methodology

Given the high dimensionality of our dataset, characterized by 95 radiomic features extracted from each patch, we chose the SOM method for its proven capability in effectively handling high-dimensional data spaces. We configured the SOM with 95 neurons in the input layer, to directly correspond with the number of features. The chosen map sizes were 2x2 and 3x3 configurations, organizing the patches into distinct clusters of four and nine, respectively. It is imperative to note that global clustering can sometimes lead to scenarios where not all clusters are represented in the final images. We analyzed both configurations to gain a comprehensive understanding of the clustering patterns, reinforcing the suitability of the SOM approach for this research.

4. EXPERIMENTS

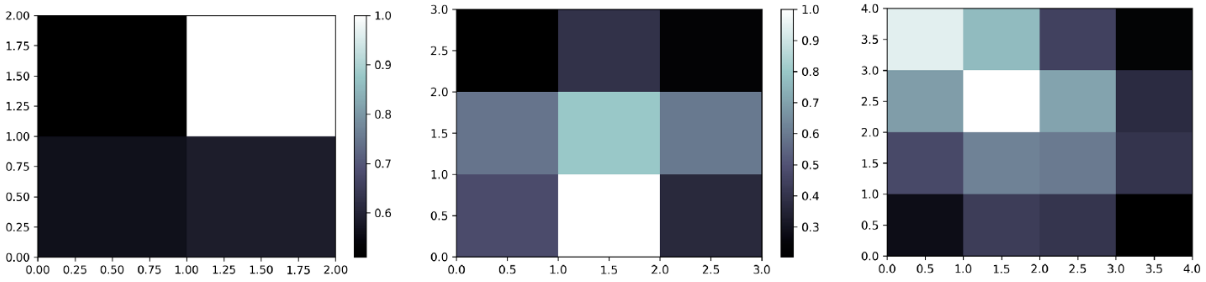
Our experiments were conducted on a server equipped with dual Intel Xeon Silver processors, 192 GB of DDR4 RAM, and two NVIDIA RTX A4000 graphics cards, running on the Linux Ubuntu 22.04 operating system. The programming work was carried out using Python 3.10. For the implementation and testing of the SOM, we utilized the MiniSom framework[14] within the Python environment. Additionally, OpenSlide was employed for loading the WSI, OpenCV version 4 for patch generation and background removal, and PyRadiomics 3.1.0 for feature extraction.

We performed grouped k-fold cross-validation, $k = 6$. This cross-validation scheme was chosen to ensure that the validation data from each fold always consisted of WSIs from distinct patients. This setup was essential, given the individual variability between patients' samples. Therefore, for each fold, 10 patients were used for training, while the samples from the remaining two patients were used for validation. To prevent any class imbalance, the number of patches in the training set was balanced using random under-sampling. After training each fold, the clustering model was then applied to the validation data to generate the resultant cluster images.

The quantization error was computed for each clustering model. For each data point, the quantization error was the distance between that data point and the weight vector of its best matching unit (BMU) in the SOM. A smaller quantization error generally indicates that the map represents the data more accurately. For biological interpretation of the clusters, a pathologist analyzed how many patches of a certain cluster belonged to a tumor, healthy, or immune infiltration regions.

5. RESULTS

As previously mentioned, the configuration for the SOM in our study was established with dimensions of 2x2, 3x3, and 4x4, consequently generating 4, 9, and 16 clusters, respectively. Figure 5 illustrates examples of the maps generated by the algorithm using our training data. These maps are subsequently utilized to cluster the data based on the extracted radiomic features. Each cluster in the SOM represents a group of patches with similar characteristics. The algorithm organizes these patches spatially on the map, where patches with more similar feature profiles are placed closer together. This spatial arrangement allows for the intuitive visualization and interpretation of the data's inherent patterns and relationships. Specifically, by clustering the radiomic features, we can discern underlying structures and relationships that are not immediately apparent from the images alone. This approach facilitates a deeper understanding of the complex data characteristics and enables us to explore potential biomarkers or correlations between the radiomic feature clusters and clinical outcomes.



(a) (b) (c)

Figure 3. Clustering 2x2, 3x3, and 4x4 maps generated on our training data.

In the clustering experiments with four classes, the quantization error across the six folds showcased a consistency, ranging between 0.5748 and 0.6364. This variability, though relatively tight, indicated subtle variations in clustering precision across different training and validation sets. Specifically, the third fold exhibited the highest quantization error at 0.6364, while the first fold showed the least error at 0.5748.

The experiments using nine classes, a discernible reduction in quantization error was observed, suggesting a potentially improved representation capability of the clustering model in this setting. The quantization error in this configuration spanned from 0.4447 to 0.5236. The third fold emerged as the best-performing set, recording the lowest quantization error at 0.4447, while the fifth fold presented the highest at 0.5236.

6. DISCUSSION

The clinical validation of our clustering approach presented a similar aspect as the computational insights from the quantization error metrics. In the four-cluster model, there was a noted overlap of cellular phenotypes, with both benign and malignant cell types manifesting within the same cluster. Such co-occurrences introduce potential ambiguities in diagnostic procedures, underscoring the necessity for methodological refinement.

In the nine-cluster model, the results demonstrated significant promise. The increased number of clusters enabled a better differentiation of distinct tissue types. Through this model, a pathologist was able to adeptly identify a range of tissue types, including tumor tissues and adjacent non-neoplastic tissues such as the lung, lung inflammation, lymphoid tissues, pleural tissue, and connective and fatty tissues. For example, the potential biological significance for cluster number two was healthy lung tissue. Clusters 3, 4, and 5 were predominantly characterized as tumor cells. Cluster 6 predominantly represented the surrounding stroma, while Cluster 7 included both the surrounding and intratumoral stroma. Cluster 8 was characterized by immune cells predominantly infiltrating the stroma, and cluster 9 was quite indicative of immune cells within pre-existing newly-formed lymphoid tissue. Figure 4 presents the clustering results for a WSI, comparing maps derived from both 4-cluster and 9-cluster configurations alongside the original IHC image.

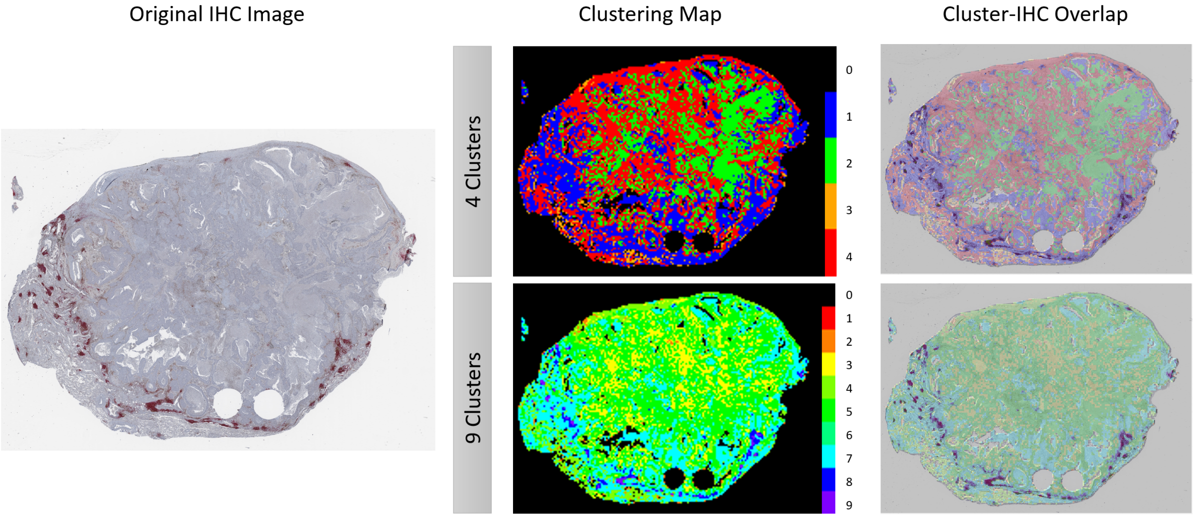


Figure 4. Comparative visualization of immunohistochemistry (IHC)-stained lung tissue. From left to right: original IHC image, clustering maps for four and nine clusters, and overlays of the IHC image with respective clustering results.

Given the high demand for high-quality image annotations in the field of machine learning - based biomarker analysis in translational biomarker research, our method holds substantial promise to accelerate current analysis workflows. The unsupervised categorization into distinct clusters can support the process of image annotation. In addion, there is potential for novel insights into tumor heterogeieity

Besides, the use of IHC staining in our method presents advantages over traditional H&E analysis, marking a significant novelty in our approach. IHC staining allows for the specific identification and visualization of certain proteins and antigens within tissue sections, providing a more targeted and detailed understanding of the tumor microenvironment. This specificity is particularly beneficial in the context of lung cancer, where the expression of certain proteins can be indicative of tumor type, prognosis, and potential therapeutic targets. Unlike H&E staining, which primarily reveals morphological features, IHC staining can highlight molecular characteristics, offering deeper insights into the biological behavior of cancer cells. Additionally, IHC is invaluable in distinguishing between different types of lung cancer subtypes, which can be challenging with H&E staining alone.

The use of an unsupervised methodology in our research presents significant advantages, particularly in the context of analyzing complex histopathology data. One of the primary benefits is the ability to discover hidden patterns and relationships within the data without the need for pre-labeled training sets. This aspect is especially crucial in the field of oncology, where the variability and subtlety of cancerous tissues can be large, and predefined categories may not capture all relevant features. Unsupervised learning algorithms, such as the SOM used in our study, are adept at identifying inherent structures and clusters within the data, leading to potentially novel insights about tumor characteristics and behavior.

Furthermore, additional studies are essential to enhance the significance of the clusters identified in our research. The findings, especially the superior performance of the 16-cluster configuration using SOM, highlight the need for further exploration into combining clusters with overlapping or redundant characteristics. The increased number of clusters, while effective in categorizing radiomic features, introduces the challenge of potential overlap and redundancy. Future research should, therefore, aim to develop methodologies for merging clusters that represent similar aspects of lung cancer histopathology.

7. CONCLUSION

In this study, unsupervised learning techniques were applied to segment NSCLC and lung tissue samples. The nine-cluster configuration, using SOM, showed the potential to effectively differentiate key histopathological features within WSIs. The clusters were associated with neoplastic regions, stromal areas, and specific patterns of immune cell infiltration. This approach suggests the potential of unsupervised learning to overcome the lack of high-quality annotations and to enhance tissue segmentation and classification in digital pathology.

Acknowledgment

M.F.O.B. is supported in part by Coordination for the Improvement of Higher Education Personnel (CAPES) under grant 88887.498626/2020-00 and 88887.695355/2022-00.

REFERENCES

[1] World Health Organization, "Lung cancer." https://www.who.int/news-room/fact-sheets/detail/lung-cancer (2023). Accessed: 2023-08-22.

[2] World Cancer Research Fund, "Lung cancer statistics." https://www.wcrf.org/cancer-trends/lung-cancer-statistics/ (2022). Accessed: 2023-08-22.

[3] American Cancer Society, "Key statistics for lung cancer." https://www.cancer.org/cancer/types/lung-cancer/about/key-statistics.html#:~:text=Overall%2C%20the%20chance%20that%20a,t%2C%20the%20risk%20is%20lower. (2023). Accessed: 2023-08-25.

[4] de Margerie-Mellon, C., De Bazelaire, C., and De Kerviler, E., "Image-guided biopsy in primary lung cancer: Why, when and how," Diagn. Interv. Imaging **97**, 965–972 (2016).

[5] Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, 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).

[6] Graham, S., Shaban, M., Qaiser, T., Koohbanani, N. A., Khurram, S. A., and Rajpoot, N., "Classification of lung cancer histology images using patch-level summary statistics," in [Medical Imaging 2018: Digital Pathology], **10581**, 327–334, SPIE (2018).

[7] Carrillo-Perez, F., Morales, J. C., Castillo-Secilla, D., Molina-Castro, Y., Guillén, A., Rojas, I., and Herrera, L. J., "Non-small-cell lung cancer classification via rna-seq and histology imaging probability fusion," BMC bioinformatics **22**(1), 1–19 (2021).

[8] Li, M., Ma, X., Chen, C., Yuan, Y., Zhang, S., Yan, Z., Chen, C., Chen, F., Bai, Y., Zhou, P., et al., "Research on the auxiliary classification and diagnosis of lung cancer subtypes based on histopathological images," Ieee Access **9**, 53687–53707 (2021).

[9] Baranwal, N., Doravari, P., and Kachhoria, R., "Classification of histopathology images of lung cancer using convolutional neural network (cnn)," in [Disruptive Developments in Biomedical Applications], 75–89, CRC Press (2022).

[10] Oswald, E., Bug, D., Grote, A., Lashuk, K., Bouteldja, N., Lenhard, D., Löhr, A., Behnke, A., Knauff, V., Edinger, A., et al., "Immune cell infiltration pattern in non-small cell lung cancer pdx models is a model immanent feature and correlates with a distinct molecular and phenotypic make-up," Journal for immunotherapy of cancer **10**(4) (2022).

[11] Rizzo, S., Botta, F., Raimondi, S., Origgi, D., Fanciullo, C., Morganti, A. G., and Bellomi, M., "Radiomics: the facts and the challenges of image analysis," Europ Radiol Exp **2**, 1–8 (2018).

[12] Tomaszewski, M. R. and Gillies, R. J., "The biological meaning of radiomic features," Radiology **298**, 505–516 (2021).

[13] Griethuysen, V. et al., "Computational radiomics system to decode the radiographic phenotype," Cancer Res **77**, e104–e107 (2017).

[14] Vettigli, G., "Minisom: minimalistic and numpy-based implementation of the self organizing map," (2018).