scientific data

OPEN

DATA DESCRIPTOR

LungHist700: A dataset of histological images for deep learning in pulmonary pathology

Jorge Diosdado[1] ✉, Pere Gilabert[1], Santi Seguí[1] & Henar Borrego[2]

Accurate detection and classification of lung malignancies are crucial for early diagnosis, treatment planning, and patient prognosis. Conventional histopathological analysis is time-consuming, limiting its clinical applicability. To address this, we present a dataset of 691 high-resolution (1200 × 1600 pixels) histopathological lung images, covering adenocarcinomas, squamous cell carcinomas, and normal tissues from 45 patients. These images are subdivided into three differentiation levels for both pathological types: well, moderately, and poorly differentiated, resulting in seven classes for classification. The dataset includes images at 20x and 40x magnification, reflecting real clinical diversity. We evaluated image classification using deep neural network and multiple instance learning approaches. Each method was used to classify images at 20x and 40x magnification into three superclasses. We achieved accuracies between 81% and 92%, depending on the method and resolution, demonstrating the dataset's utility.

Background & Summary

Cancer is the second leading cause of death globally. In 2022, more than 20 million new cancer cases were reported, and approximately 9.7 million people succumbed to the disease worldwide. Lung cancer, with more than 2.5 million new cases diagnosed[1], was the most lethal, accounting for 1.8 million deaths. This staggering figure represents a fifth of all cancer deaths globally, significantly more than the second deadliest cancer, colon and rectum cancer, which caused almost 904,000 deaths in the same year, 2022[2].

The high mortality rate of lung cancer is mainly due to late detection. Early diagnosis of lung cancer is key to survival. However, by the time symptoms become apparent, the disease has often spread, resulting in a low survival rate[3]. The 5-year survival rate for early-stage lung cancer can exceed 90%, while for patients diagnosed at a late stage, it can be less than 10%0[4]. Smoking, identified as the leading risk factor by the American Cancer Society, is projected to account for 81% of lung cancer cases in 2023[5].

Carcinomas, malignancies that develop from epithelial cells, are the most common type of malignancy in the lungs. Carcinomas located in the lungs that originate there are referred to as primary lung carcinomas, distinguishing them from those that have spread to the lungs via metastasis. Primary lung carcinomas can be divided into two major histopathological types: small cell carcinoma and non-small cell carcinoma, with non-small cell carcinomas being the most frequent[6].

Non-small cell carcinoma can be classified into two main subtypes: adenocarcinomas and squamous cell carcinomas.

- **Adenocarcinomas**: These tumors exhibit microscopic glandular-related tissue cytology, tissue architecture, and/or gland-related products.
- **Squamous Cell Carcinomas**: These tumors are characterized by observable traits of squamous differentiation, such as intercellular bridges, keratinization, and the formation of squamous pearls[6].

Additionally, there are other less common types of non-small cell carcinoma, such as large cell carcinoma, adenosquamous carcinoma, and sarcomatoid carcinoma, each with its own unique histological features and clinical behaviors that may influence treatment strategies and prognosis.

[1]Dept. de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain. [2]Hospital Clínico Universitario de Valladolid, Valladolid, Spain. ✉e-mail: diosdado100591@hotmail.com

Distinguishing the histological types of lung carcinomas is crucial in the era of personalized medicine, as each tumor type can be associated with different genetic alterations within the tumor itself. These genetic changes, in turn, are related to targeted therapies aimed at those specific mutations, improving the medium and long-term prognosis[7].

Histopathological images, microscopic images of tissue samples, play a crucial role in medical diagnosis and research. They offer valuable insights into the appearance and structure of cells and tissues, enabling pathologists to accurately identify and classify diseases[8]. However, manual analysis of these images is time-consuming and prone to human error[9]. Therefore, histopathological image datasets, collections of labeled histopathological images, are essential for developing and training image analysis algorithms. These datasets provide researchers with a large and diverse set of images, facilitating the creation of artificial intelligence (AI) models that can accurately classify and diagnose diseases, thereby assisting human experts in their tasks.

The field of AI is expanding rapidly, with new applications emerging daily, particularly in the medical sector[10]. One promising application is in diagnostics, where AI can enhance both diagnostic accuracy and efficiency. AI can improve the early detection and diagnosis of lung cancer, potentially leading to better patient outcomes[11].

To develop AI algorithms using lung histopathology images, several popular datasets are frequently utilized. Three of the most important ones are TCGA-LUAD[12] for adenocarcinomas, TCGA-LUSC[13] for squamous cell carcinoma, and LC25000[14] which also includes slides from benign patients. The TCGA-LUAD and TCGA-LUSC datasets contain whole slide images (WSI) of lungs, specifically 541 slides from 478 LUAD patients and 512 slides from 478 LUSC patients.

The LC25000 dataset consists of 750 images of size 768 × 768, classified into three different categories: lung benign, lung adenocarcinoma, and lung squamous cell carcinoma, with 250 unique images in each category. Additionally, the dataset contains 500 images of the colon. All these images were then artificially augmented to create a dataset of 25,000 images. However, the absence of traceability from the original images to the augmented ones poses a challenge in accurately dividing the dataset into training, validation, and test sets. This lack of traceability can lead to potential data leakage during the training and validation stages, undermining the validity of technical conclusions drawn from studies using this dataset[15–19].

This paper introduces a novel dataset, LungHist700, comprising 691 images of size 1200 × 1600 pixels from both normal lung tissue and primary lung carcinomas. The carcinomas are categorized into two types: adenocarcinomas and squamous cell carcinomas. Each of these types is further subclassified based on the degree of carcinoma differentiation into three levels: well differentiated, moderately differentiated, and poorly differentiated.

Methods

Data was collected from 45 patients at Hospital Clínico de Valladolid in 2023 as part of a regular diagnostic process. The dataset consists of images of hematoxylin and eosin-stained samples extracted from pathology glass slides using a Leica DM 2000 microscope and a Leica ICC50 W microscope camera at two distinct magnifications: 20x and 40x. The field of view was meticulously selected by a pathologist to encompass representative tissue of the category. In most cases, this tissue is discernible in all four quadrants of the image.

All individuals included in the study were surgical patients, so all images are from patients with malignancies. Images classified as showing normal lung depict areas where the tumor has not spread.

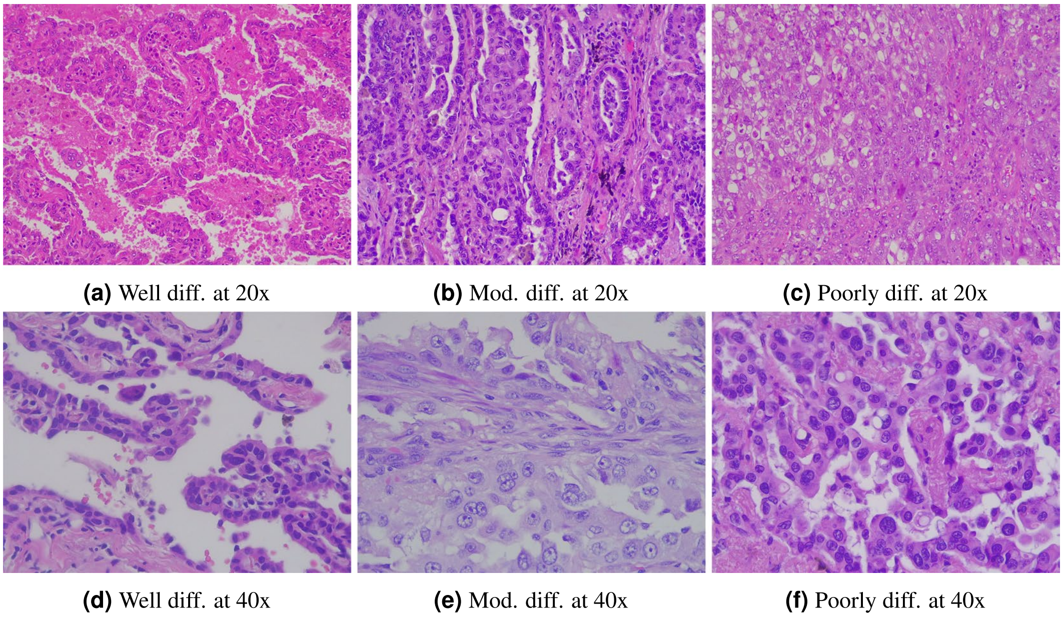
For each patient, two concurrent evaluations were conducted to determine the type of tumor (adenocarcinoma or squamous cell carcinoma) and the level of differentiation (well differentiated, moderately differentiated, or poorly differentiated). The first evaluation was a morphological analysis of the tissue based on the hematoxylin and eosin-stained samples, which determined the classification of well and moderately differentiated samples. The second evaluation involved immunohistochemical tests of the tissue to determine the type of tumor (adenocarcinoma or squamous). These tests, combined with contextual information, contributed to the accurate classification of poorly differentiated categories. The tests performed were TTF1, CK7, Napsin A, P40, and CK5/6. Using the results from all the tests, a specialist pathologist classified the images into the seven classes of the dataset.

For adenocarcinomas, the differentiation grading system recommended by the College of American Pathologists[20,21] was employed. According to their guidelines, there are three differentiation levels:

1. Well-differentiated: Tumors primarily exhibiting a lepidic pattern, with no high-grade components or less than 20% high-grade features (such as solid, micropapillary, or complex glandular patterns).
2. Moderately differentiated adenocarcinoma: Tumors mainly showing acinar or papillary patterns, with less than 20% high-grade features.
3. Poorly differentiated adenocarcinoma: Tumors that have 20% or more high-grade features.

Pulmonary squamous cell carcinoma has also traditionally been divided into well differentiated, moderately differentiated, and poorly differentiated, similar to squamous cell carcinomas of other organ systems. The degree of differentiation is generally dependent on a combination of features, such as the presence or absence of keratinization and intercellular bridges, as well as cellular pleomorphism and mitotic activity[22]. Following these guidelines, squamous cell carcinoma has been divided into the following three categories:

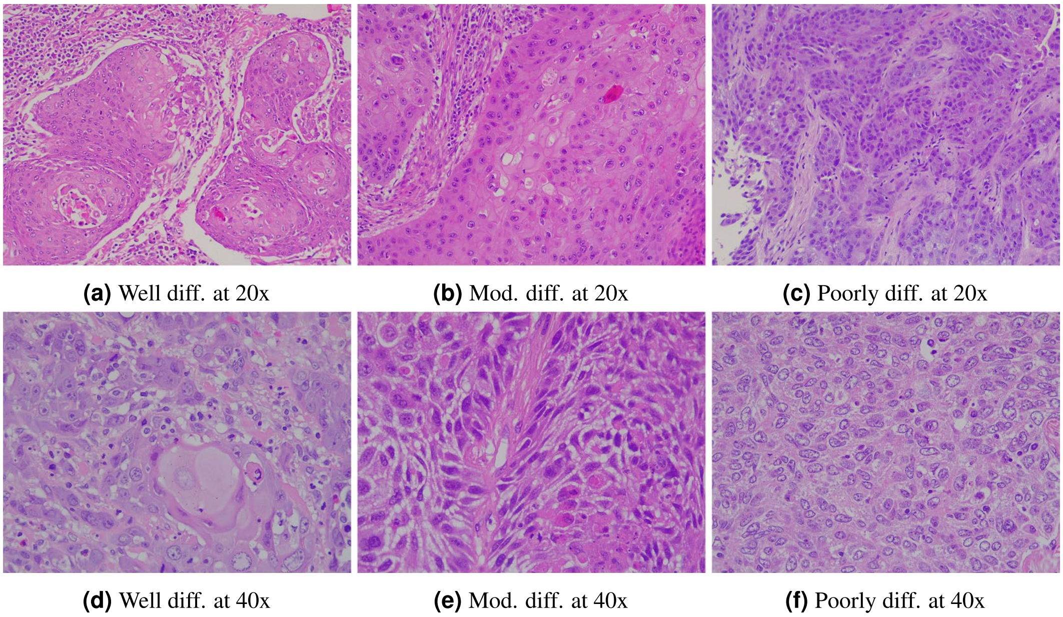
1. Well differentiated: These tumors exhibit keratinization, such as keratin pearls and intercellular bridges. They typically grow in sheets or nests, with polygonal cells that have round to oval nuclei, vesicular features, and eosinophilic cytoplasm. Additionally, mitotic figures and focal areas of hemorrhage or necrosis may be present.



(a) Well diff. at 20x **(b)** Mod. diff. at 20x **(c)** Poorly diff. at 20x

(d) Well diff. at 40x **(e)** Mod. diff. at 40x **(f)** Poorly diff. at 40x

Fig. 1 Images displaying adenocarcinoma at varying levels of differentiation and resolution.



(a) Well diff. at 20x **(b)** Mod. diff. at 20x **(c)** Poorly diff. at 20x

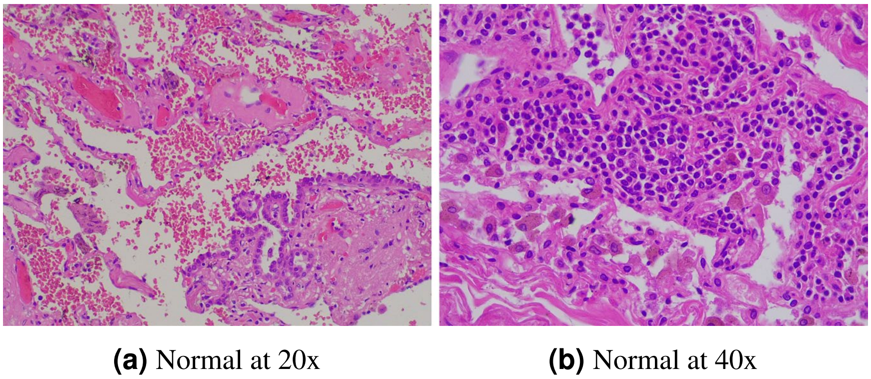
(d) Well diff. at 40x **(e)** Mod. diff. at 40x **(f)** Poorly diff. at 40x

Fig. 2 Images displaying squamous cell carcinoma at varying levels of differentiation and resolution.

2. Moderately differentiated: These tumors show increased cytologic atypia and mitotic activity. Although keratinization and intercellular bridges are still present, they are less prominent compared to well-differentiated tumors. Moreover, areas of hemorrhage or necrosis are more common.
3. Poorly differentiated: These tumors grow in sheets and are often unrecognizable as squamous type without immunohistochemistry. They display significant cellular pleomorphism, high mitotic activity, and extensive areas of necrosis.

Figure 1 shows adenocarcinoma samples, Fig. 2 displays squamous cell carcinoma samples at varying levels of differentiation and resolution. Figure 3 presents images of normal lung tissue at two different resolution.

Ethics approval. The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Ethical Committee of the Hospital Clínico Universitario de Valladolid (CEIm Área de Salud Valladolid Este) under project PI 23–3167. The committee waived participant consent given data anonymization and approved open publication of the data.



(a) Normal at 20x

(b) Normal at 40x

Fig. 3 Normal lung images at different resolution.

Description	Id.	20x	40x	Subclass total	Superclass total
Well differentiated adenocarcinoma	aca_bd	57	46	**103**	**280**
Moderately differentiated adenocarcinoma	aca_md	44	46	**90**	
Poorly differentiated adenocarcinoma	aca_pd	45	42	**87**	
Normal lung	nor	85	66	**151**	**151**
Well differentiated squamous cell carcinoma	scc_bd	50	49	**99**	**260**
Moderately differentiated squamous cell carcinoma	scc_md	30	36	**66**	
Poorly differentiated squamous cell carcinoma	scc_pd	48	47	**95**	
Total		**359**	**332**	**691**	**691**

Table 1. The dataset comprises three classes: adenocarcinoma (aca), squamous cell carcinoma (scc), and normal (nor). Images showing malignant tissue are further categorized based on their differentiation level.

Data Records

The dataset is available at figshare[23]. It consists of 691 images from 45 patients, with each image having a resolution of 1200 × 1600 pixels and stored in *.jpg* format. These images are captured at either 20x or 40x magnification levels and are categorized into seven classes (see Table 1). An accompanying*.csv* file links each image to the associated patient ID. All patients have been anonymized, and the file includes an identifier to match images from the same patient.

Technical Validation

In this section, we present two baseline methods for classifying the dataset into the three major superclasses. First, a classic approach was employed where images were resized, and a deep neural network (DNN) was trained. The second method involves a multiple instance learning (MIL) strategy, where patches of the images were extracted, and the same DNN was used to obtain multiple embeddings, one for each patch. An attention[24] layer was then applied to relate and aggregate these embeddings for image classification.

All the experiments used the same training configuration: the networks were implemented using Keras and executed on an NVIDIA RTX 3090 with CUDA 11.0. The DNN model used in both methods was a ResNet50 network pretrained on ImageNet. The Adam optimizer was employed with an initial learning rate of 1e-5, which was reduced by a factor of 0.1 if the model began to overfit. Categorical cross-entropy was used as the loss function in both experiments. The Albumentations library[25] was utilized to generate augmentations on the fly during training.

Images were classified into their superclasses: "aca" (adenocarcinoma), "scc" (squamous cell carcinoma), and "nor" (normal). The data was divided into three sets: 80% for training, 10% for validation, and the remaining 10% for testing. A patient-wise strategy was employed, ensuring that images from the same patient were placed in the same set to ensure fair evaluation and prevent data leakage.

DNN Baseline. To train the ResNet50 model, images were resized to 300 × 400 pixels to better fit this architecture. The published dataset, however, contains images at their original resolution (1200 × 1600 pixels). Figure 4 illustrates the learning curves on the training and validation splits, as well as the classification confusion matrix of the experiment on the test set for the 20x resolution. The model achieved an accuracy of 90%, a ROC-AUC of 98%, a precision of 92%, and a recall of 87%.

The experiment was then repeated with the same configuration but using images at 40x resolution. The model achieved an accuracy of 82%, a ROC-AUC of 94%, a precision of 82%, and a recall of 84%.

To assess the validity and explainability of the results, we used Grad-CAM[26] on the last convolutional layer of the ResNet50 model. The threshold was set to 0.25 to visualize the Grad-CAM activations. Figure 5 shows the explanation masks generated by the algorithm on some test images, each representing a distinct

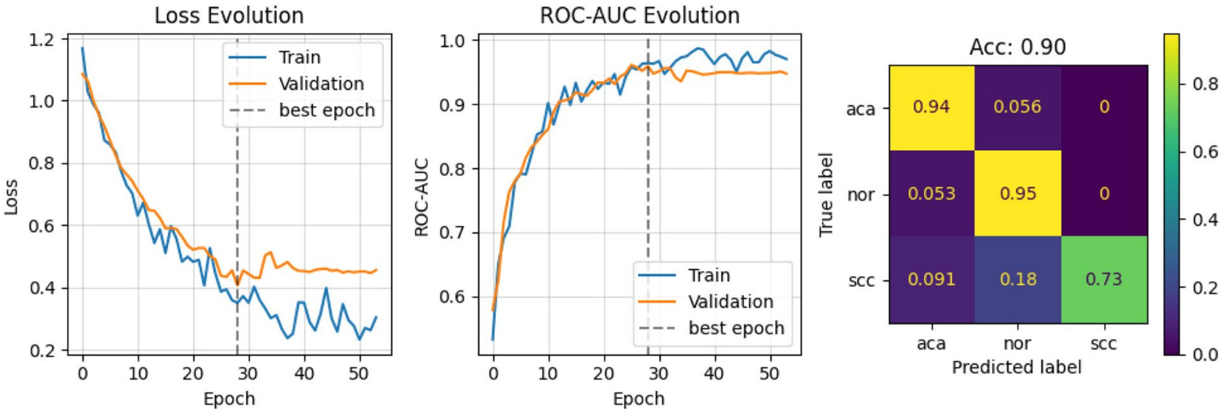
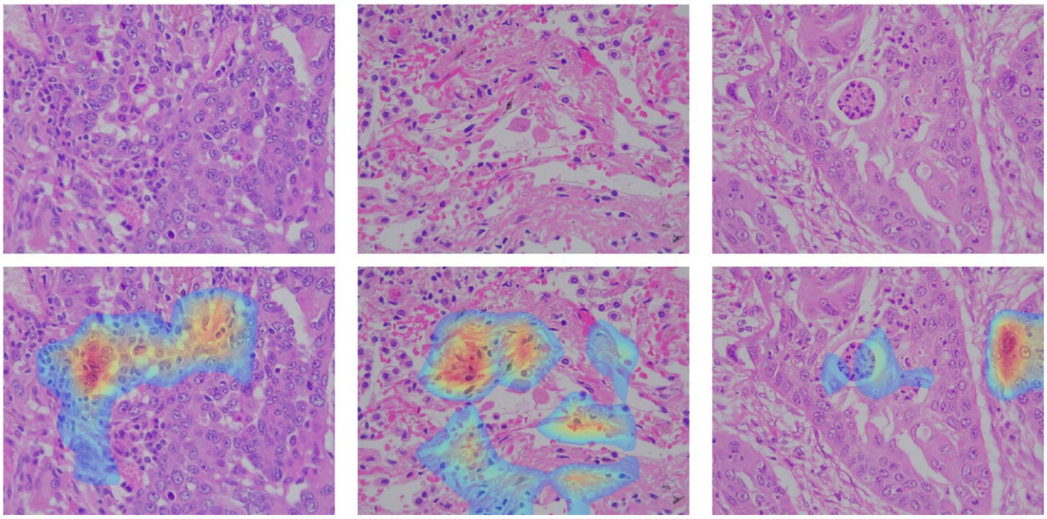


Fig. 4 Classification results of the proposed baseline for 20x resolution. Early stopping was triggered at epoch 28, based on the validation set. After that, the best weights were loaded. The confusion matrix shows the correctly classified percentage of samples and the classification errors on the test set. The results are normalized by rows (True label).



(a) Adenocarcinoma **(b)** Normal lung **(c)** Squamous cell carcinoma

Fig. 5 Masks generated by the Grad-CAM algorithm on some test images.

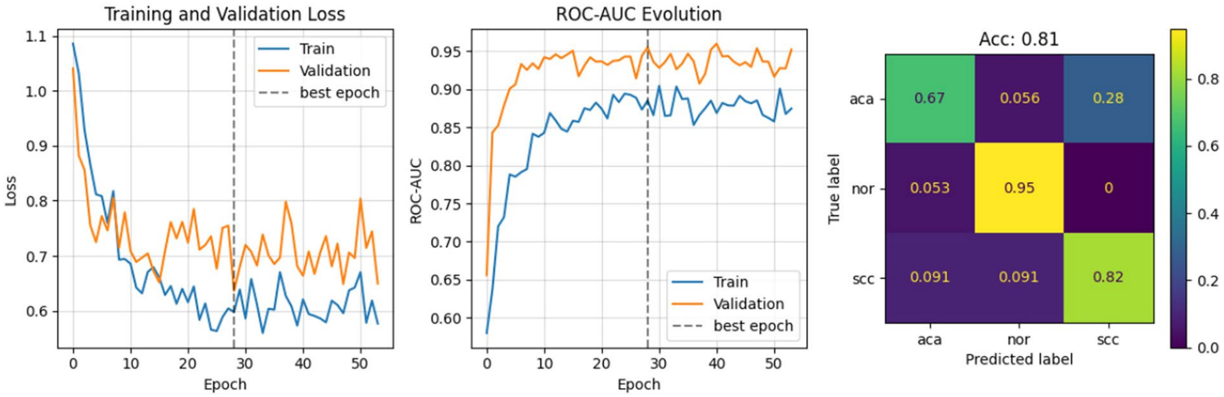


Fig. 6 Classification performance of the MIL algorithm (ResNet50 + Multi-Head Attention layer) for 20x resolution. Early stopping was triggered at epoch 28.

histopathological class: adenocarcinoma, normal tissue, and squamous cell carcinoma. The masks illustrate how the model highlights specific areas relevant to image classification. The results were cross-checked with the medical team to validate the model's output.

MIL Baseline. A second strategy based on ResNet50 was also tested. We trained a MIL algorithm that consisted of a ResNet50 followed by a Multi-Head Attention layer. During training, we extracted 20 random patches of size 224 × 224 and used the ResNet architecture to obtain embeddings for each patch. An attention layer with four heads was then applied, followed by average pooling to obtain a single embedding for classification. All the training parameters remained the same, though the batch size was reduced to three to fit within the GPU's memory constraints. The results of the MIL algorithm for images at 20x resolution are shown in Fig. 6. This baseline model achieved an accuracy of 81%, a ROC-AUC of 89%, a precision of 80%, and a recall of 81% on the test set.

Code availability

Code to reproduce the DNN baseline is available at https://github.com/jorgediosdado/LungHist700.

Received: 25 March 2024; Accepted: 30 September 2024;
Published online: 05 October 2024

References

1. Bray, F. *et al*. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* **74**, 229–263, https://doi.org/10.3322/caac.21834 (2024).
2. Cancer. https://www.who.int/news-room/fact-sheets/detail/cancer available online (Accessed in October 2023).
3. Sullivan, F. M. *et al*. Earlier diagnosis of lung cancer in a randomised trial of an autoantibody blood test followed by imaging. *European Respiratory Journal* **57**, https://doi.org/10.1183/13993003.00670-2020 (2021).
4. Ning, J. *et al*. Early diagnosis of lung cancer: which is the optimal choice. *Aging (Albany NY)* **13**, 6214, https://doi.org/10.18632/aging.202504 (2021).
5. Siegel, R. L., Miller, K. D., Wagle, N. S. & Jemal, A. Cancer statistics, 2023. *Ca Cancer J Clin* **73**, 17–48, https://doi.org/10.3322/caac.21763 (2023).
6. Kumar, V., Abbas, A., Aster, J. C. & Deyrup, A. T. *Robbins & Kumar Basic Pathology* (Elsevier Health Sciences, 2022).
7. Rodak, O., Peris-Díaz, M. D., Olbromski, M., Podhorska-Okołó, M. & Dzięgiel, P. Current landscape of non-small cell lung cancer: Epidemiology, histological classification, targeted therapies, and immunotherapy. *Cancers* **13**, https://doi.org/10.3390/cancers13184705 (2021).
8. Mukhopadhyay, S. *et al*. Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: A multicenter blinded randomized noninferiority study of 1992 cases (pivotal study. *The American Journal of Surgical Pathology* **42**, 39–52, https://doi.org/10.1097/PAS.0000000000000948 (2018).
9. Peck, M., Moffatt, D., Latham, B. & Badrick, T. Review of diagnostic error in anatomical pathology and the role and value of second opinions in error prevention. *Journal of Clinical Pathology* **71**, jclinpath–2018, https://doi.org/10.1136/jclinpath-2018-205226 (2018).
10. Esteva, A. *et al*. A guide to deep learning in healthcare. *Nature medicine* **25**, 24–29, https://doi.org/10.1038/s41591-018-0316-z (2019).
11. Khodabakhshi, Z. *et al*. Non-small cell lung carcinoma histopathological subtype phenotyping using high-dimensional multinomial multiclass ct radiomics signature. *Computers in biology and medicine* **136**, 104752, https://doi.org/10.1016/j.compbiomed.2021.104752 (2021).
12. Collisson, E. A. *et al*. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550, https://doi.org/10.1038/nature13385 (2014).
13. Kirk, S. *et al*. The cancer genome atlas lung squamous cell carcinoma collection (tcga-lusc) (version 4) [data set]. *The Cancer Imaging Archive* https://doi.org/10.7937/K9/TCIA.2016.TYGKKFMQ (2016).
14. Borkowski, A. A. *et al*. Lung and colon cancer histopathological image dataset (lc25000), https://doi.org/10.48550/arXiv.1912.12142 (2019).
15. Mehmood, S. *et al*. Malignancy detection in lung and colon histopathology images using transfer learning with class selective image processing. *IEEE Access* **10**, 25657–25668, https://doi.org/10.1109/ACCESS.2022.3150924 (2022).
16. Mangal, S., Chaurasia, A. & Khajanchi, A. Convolution neural networks for diagnosing colon and lung cancer histopathological images, https://doi.org/10.48550/arXiv.2009.03878 (2020).
17. Masud, M., Sikder, N., Nahid, A.-A., Bairagi, A. K. & AlZain, M. A. A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework. *Sensors* **21**, 748, https://doi.org/10.3390/s21030748 (2021).
18. Al-Jabbar, M., Alshahrani, M., Senan, E. M. & Ahmed, I. A. Histopathological analysis for detecting lung and colon cancer malignancies using hybrid systems with fused features. *Bioengineering* **10**, 383, https://doi.org/10.3390/bioengineering10030383 (2023).
19. Hatuwal, B. K. & Thapa, H. C. Lung cancer detection using convolutional neural network on histopathological images. *International Journal of Computer Trends and Technology* **68**, 21–24, https://doi.org/10.14445/22312803/IJCTT-V68I10P104 (2020).
20. Moreira, A. L. *et al*. A grading system for invasive pulmonary adenocarcinoma: a proposal from the international association for the study of lung cancer pathology committee. *Journal of Thoracic Oncology* **15**, 1599–1610, https://doi.org/10.1016/j.jtho.2020.06.001 (2020).
21. Schneider, F. *et al*. Protocol for the examination of resection specimens from patients with primary non-small cell carcinoma, small cell carcinoma, or carcinoid tumor of the lung. *Arch Pathol Lab Med* **133**, 1552–1559, https://doi.org/10.5858/133.10.1552 (2009).
22. Weissferdt, A. *Diagnostic Thoracic Pathology* (Springer, 2020).
23. Diosdado, J., Gilabert, P., Santi, S. & Borrego, H. Lunghist700: A dataset of histological images for deep learning in pulmonary pathology. *figshare* https://doi.org/10.6084/m9.figshare.25459174 (2024).
24. Vaswani, A. *et al*. Attention is all you need. *Advances in neural information processing systems* **30** (2017).
25. Buslaev, A. *et al*. Albumentations: Fast and flexible image augmentations. *Information* **11**, https://doi.org/10.3390/info11020125 (2020).
26. Selvaraju, R. R. *et al*. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, https://doi.org/10.1007/s11263-019-01228-7 (2017).

Acknowledgements

This work has been partially funded by 2021 SGR 01104 and FPU20/01090 grant.

Author contributions

J.D. and H.B. conceived the experiments, collected and processed the data. J.D. and P.G. conducted the experiments and wrote the manuscript. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.