Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 21 No. 7 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v21i07.54323

PAPER

Construction of a New Data Set of Pleural Fluid Cytological Images for Research

Frida López-Córdova[1], Hugo Vega-Huerta[1](✉), Gisella Luisa Elena Maquen-Niño[2], Jaime Cáceres-Pizarro[3], Ivan Adrianzén-Olano[4], Oscar Benito-Pacheco[1]

[1]Universidad Nacional Mayor de San Marcos, Lima, Perú

[2]Universidad Nacional Pedro Ruíz Gallo, Lambayeque, Perú

[3]Hospital Nacional Cayetano Heredia, Lima, Perú

[4]Universidad Nacional Toribio Rodríguez de Mendoza, Chachapoyas, Perú

hvegah@unmsm.edu.pe

ABSTRACT

The limited availability of standardized datasets has hindered the implementation of artificial intelligence (AI) models in serous fluid cytology, particularly in pleural fluid analysis. In this paper, we present the construction of a dataset of pleural fluid cytology images. The objective is to generate a dataset of pleural fluid cytologic images validated by two pathologists and classified into five categories for cell diagnosis, which will be used to train AI models. As a methodology, the images represent pleural fluid cytology samples that have been prepared through medical procedures and transferred to slides, providing valuable information when evaluated under the microscope by medical specialists through cytological examination. We documented the entire process for building the pleural fluid cytological image dataset, from image capture, labeling, preprocessing, standardization, and uploading to public platforms. As a result, we obtained a pleural fluid cytology dataset based on the International System (TIS) criteria for reporting serous fluids, classifying samples into AUS, MAL, ND, NFM, and SFM. This dataset is intended to support medical research, deep learning applications in medical image analysis, and improved diagnostic methodologies.

KEYWORDS

dataset, construction of a dataset, image preprocessing, cytology, cytological images, pleural fluid, cytological diagnosis

1 INTRODUCTION

The analysis of cellular images has played a transcendental role in medicine. Since its discovery in 1965, the observation of cellular features through the microscope has led to a better understanding of their structure and functions [1]. Scientists have used microscopes to evaluate the efficacy of various compounds in drug development [2]. In cytology, observation of cell morphology has been used to discern malignancy in cancer cells [3]. Pleural fluid cytology is a laboratory test

López-Córdova, F., Vega-Huerta, H., Maquen-Niño, G.L.E., Cáceres-Pizarro, J., Adrianzén-Olano, I., Benito-Pacheco, O. (2025). Construction of a New Data Set of Pleural Fluid Cytological Images for Research. *International Journal of Online and Biomedical Engineering (iJOE)*, 21(7), pp. 138–151. https://doi.org/10.3991/ijoe.v21i07.54323

Article submitted 2025-01-08. Revision uploaded 2025-03-07. Final acceptance 2025-03-22.

that helps detect the presence of cancerous and other cells in the fluid surrounding the lungs. It is crucial for identifying malignant cells in patients with pleural effusion, in addition to being a low-cost, fast-turnaround, and minimally invasive diagnostic tool. Cytological analysis not only provides diagnostic information but also offers valuable insights into the staging of malignant neoplasms, guiding treatment and subsequent follow-up. However, as a visual diagnostic process, it relies on the skill and experience of the pathologist, which can lead to variability in diagnoses and potential delays [4]. In recent years, the use of deep learning models has shown great potential to improve accuracy and speed in medical image analysis [5].

Recent advancements in artificial intelligence (AI) and deep learning have demonstrated the potential to enhance accuracy and efficiency in medical image analysis [3]. AI-based approaches have been successfully applied to various fields of pathology, including hematology and histopathology, where they have improved diagnostic precision. However, the application of AI in pleural fluid cytology remains scarce, primarily due to the lack of publicly available and well-annotated datasets.

Unlike other medical imaging domains, such as radiology and dermatology, where substantial public datasets exist (e.g., ChestX-ray14 for thoracic diseases or HAM10000 for skin lesions), cytology datasets are significantly limited [6]. This lack of structured data has hindered the development of robust AI models for cytological image analysis.

A common task useful in many practical situations is to determine the category of a given cell or set of cells: a task known as cell sorting [7].

Researchers working with machine learning techniques in medical image analysis often lack formal medical training, making independent data acquisition and annotation even more complex. Additionally, medical institutions that own these datasets face legal and ethical restrictions that prevent public sharing, resulting in inconsistent and non-comparable research outcomes. Studies on automated cell classification, such as those focused on Pap smears and peripheral blood smears, have benefited from structured datasets, yet similar resources for pleural fluid cytology remain largely unavailable. The absence of publicly accessible pleural fluid cytology datasets limits the development of AI-driven diagnostic tools and the reproducibility of research in this field.

Data scarcity poses challenges in the application of machine learning methods, especially in medical imaging. Researchers in this field often lack medical training; most are technology professionals, making it difficult to independently obtain and annotate data [8].

Given this gap, our work provides a valuable contribution by developing a dataset of 2,640 pleural fluid cytology images, following rigorous clinical protocols. The dataset is validated by expert pathologists and classified under the International System for Reporting Serous Fluid Cytology (TIS) criteria, which includes categories such as Atypia of Undetermined Significance (AUS), Malignant (MAL), Non-Diagnostic (ND), Non-Fluid Malignant (NFM), and Suspicious for Malignancy (SFM). Unlike existing cytological datasets that are often limited to single-class annotations, our dataset offers a comprehensive, multi-category classification designed to support AI-driven cytological research.

In addition, the dynamics and challenges of labeling a pleural fluid cytology dataset under the TIS criteria are presented. The dataset offers an affordable and accessible solution to improve cancer cell detection and will be available to the research community.

2 LITERATURE REVIEW

Pleural fluid cytology is a discipline that studies individual cells to diagnose diseases and determine the nature of pathological processes [9], [10], [11]. In the context of pleural fluid, cytological examination is essential for the evaluation of pathologies such as infections, inflammatory diseases, and malignant neoplasms [12]. This procedure allows identifying morphological changes in cells and assessing the presence of malignancy, being a key tool in the diagnosis of secondary pleural cancers.

Medical imaging is essential in modern clinical practice, providing crucial information for cytological analysis. High-resolution image acquisition techniques have made it possible to digitize cytological samples, facilitating storage, remote interpretation, and research [13]. These tools have also been integrated with deep learning systems to automate the detection of abnormal patterns in cells, reducing human error and increasing diagnostic accuracy [3].

Cytological examination of pleural fluid is a minimally invasive procedure that involves the collection and microscopic analysis of the cells present in the pleural fluid [14]. This test is highly useful in identifying malignancies such as mesotheliomas and metastatic cancers [15]. The sensitivity of this procedure may vary, but its specificity is high when careful analysis is performed.

Deep learning has revolutionized medical image analysis, enabling significant advances in classification, segmentation, and pathology detection. Convolutional neural networks (CNNs) have proven particularly effective in cellular pattern recognition, automating processes that traditionally required human experts [16]. In the context of pleural fluid, these methods have the potential to improve diagnostic accuracy by identifying malignant cells with greater sensitivity and specificity [17].

[18] from the Kaggle repository has 169 cytological images of pleural fluid distributed in two diagnostic categories: NFM 160 and MAL 533.

A high-quality, representative dataset is essential for training deep learning models. In the case of pleural fluid cytology, current datasets are limited in terms of case diversity and image quality, making it difficult to develop robust models [19]. Building new datasets can facilitate more advanced investigations and foster the creation of highly accurate diagnostic support systems. Table 1 shows a comparison of the data sets of pleural fluid cytological images.

Table. 1. Comparative analysis of 3 datasets

Dataset	Number of Images	Number of Output Categories	Number of Images Per Output Category
Pleural2640 (Our New Dataset)	2640	5	AUS 358, MAL 645, ND 463, NFM 647, SFM 647
Almoniem	3731	4	
Body fluid cytology	693	2	NFM 16, MAL 533

3 RELATED WORKS

There is research that reinforces the importance of using curated and processed data sets to optimize the performance of machine learning and deep learning models

applied to medicine [20], [21], [22]. This is even more necessary in machine learning models that use images [23], [24], [25], [26], [27].

[3] presents a broad overview of the use of deep learning algorithms in medical image analysis, highlighting their potential to automate diagnostic tasks. In cytology, recent research has applied convolutional neural networks to detect cancer cells in biological fluids, showing promising results in terms of sensitivity and specificity.

[19] emphasizes the importance of well-structured datasets for training deep learning models. While databases such as ImageNet exist for general images, cytology-specific databases are scarce. This underscores the need for initiatives such as the present research, which seeks to create a specialized dataset for pleural fluid analysis.

[28] discusses the challenges associated with the cytological diagnosis of pleural fluid, such as variability in sample quality and observer experience. These problems can be mitigated by deep learning tools, but the lack of representative datasets remains a significant barrier.

[29] presents a real and validated dataset of serous fluid cytological images based on the TIS. The main objective is to develop a public resource for AI research applied to serous effusion diagnosis. The dataset (ALMONIEM) includes 3,731 images classified into four diagnostic categories (NFM, AUS, SFM, and MAL), obtained using standardized staining and preparation techniques.

In the field of machine learning, the quality and representativeness of datasets are critical for the development of accurate and generalizable models. An example of this relevance is [30], where an online course evaluation framework based on sentiment analysis and machine learning is proposed. In this study, an extensive dataset of online course reviews was collected by web scraping, which allowed training machine learning models with more than 90% accuracy in classifying student opinions. This study highlights how building a well-structured dataset enables the optimization of AI models.

Several studies have demonstrated the relevance of using quality datasets to improve the performance of deep learning models in medical diagnostics. In this regard, [31] emphasize the impact of dataset selection on the effectiveness of transfer learning models applied to chest X-ray images. The authors emphasize that the variability and representativeness of the dataset are determining factors in achieving high accuracy in image classification, allowing pretrained models to efficiently adapt their features to new diagnostic tasks [32].

The application of convolutional neural networks (CNN) in the prediction of skin diseases has proven to be a promising tool for the automated analysis of dermatological images, where the relevance of the dataset used is highlighted, since the accuracy and efficiency of deep learning models depend largely on the quality, diversity, and volume of the available images. The correct selection and curation of the dataset allow improving the generalization capacity of the model, reducing biases, and increasing diagnostic reliability [33].

The above background highlights the relevance of combining technological advances in deep learning with the construction of specialized datasets to address diagnostic challenges in cytology. These investigations serve as fundamental building blocks for the creation of a pleural fluid cytologic imaging dataset, thus contributing to the development of more accurate and efficient tools for clinical practice and research.

4 MATERIAL AND METHODS

High-quality data is essential to train algorithms, so it must be accurately labeled and include sufficient morphological diversity. The creation of a dataset is one of the first phases in machine learning work, and dealing with cytological images of pleural fluid, it is a necessary and relevant task since there is a deficiency in freely available datasets to train AI models. For this purpose, it is necessary to follow a structured methodology to ensure the quality, representativeness, and diversity of the images [34].

4.1 Authorization for the construction of the data set

This project was carried out in accordance with the authorization of the Institutional Research Ethics Committee of the Cayetano Heredia National Hospital (HNCH) through the Department of Anatomic Pathology, Cytology Unit, and the guidelines of the Postgraduate Unit of the Faculty of Systems Engineering and Computer Science of the Universidad Nacional Mayor de San Marcos (UNMSM).

4.2 Microscope specification

A microscope is an instrument that allows the observation of objects too small to be seen by the human eye. The term *microscope* derives from two concepts: "micro," meaning "small," and "scopio," meaning "to observe." Thus, it refers to the observation of small objects. The microscope is an optical instrument that enhances the ability to observe at such levels of magnification that it even enables the analysis of particles. The images it produces aid in the investigation of object composition. Therefore, the study and analysis of small objects is called *microscopy*.

As is common in most research, the field of view of an optical microscope can be increased to more than 60× [35].

Microscopy refers to the observation of very small objects under high magnification. The devices used for this purpose are called microscopes. In medicine, microscopy is particularly used for analyzing tissues, cells, blood components, and microorganisms. Cytological images are analyzed using an optical microscope with an integrated Olympus digital camera, equipped with a 10× eyepiece and a 60× objective, resulting in a total magnification of 600×. Images are captured at different magnifications (10×, 60×), allowing for a detailed analysis of cells. The high magnification level ensures that diagnostically relevant cellular details are clearly visible. To save the images without loss of quality, the PNG format was used, as it preserves visual fidelity without excessive compression.

The optical microscope used to capture all images has a built-in camera and dedicated software installed on the computer in the anatomic pathology department, enabling image capture and storage. Figure 1 shows different shots of the optical microscope used in this study.

Fig. 1. HNCH optical microscope

5 CONSTRUCTION OF THE DATA SET

The dataset consisted of 2640 high-quality images in PNG format, each with a resolution of 4140×3096 pixels representing the resolution of the image, the size of the images varying between 11 and 12 MB. The following steps were taken to build the dataset:

5.1 Image capture

The slide is placed in the microscope with a 20× magnification objective lens, and the area with suspicious cells for the diagnostic category being sought is located. Then the objective lens is changed to 60 magnifications, and the best image to be captured is focused with the microscope's own software installed in the computer of the anatomic pathology department. The microscope has a camera that is operated through the computer (the camera has a cable that connects directly to the computer). With the objective of magnification, the microscope's field of view is pointed at the

cells that are the subject of the search. Once the image is fixed in the visual field to be captured, the image is visualized through the computer to be captured. The final focus is processed on the computer screen; the image is captured in color and stored in the folder corresponding to the diagnostic category.

5.2 Image labeling

The images were reviewed by two pathologists to verify that they had been labeled in their respective classes. The 2 pathologists have 20 years and eight years of experience respectively. The images are organized into categories labeled according to the presence or absence of malignancy. Pathologists with expertise in cytology identified the images, labeling different types of cells found according to the TIS (AUS atypical cells, MAL malignant cells, ND non-diagnostic, NFM malignancy negative cells, and SFM malignancy suspicious cells), as visualized in Figure 2.

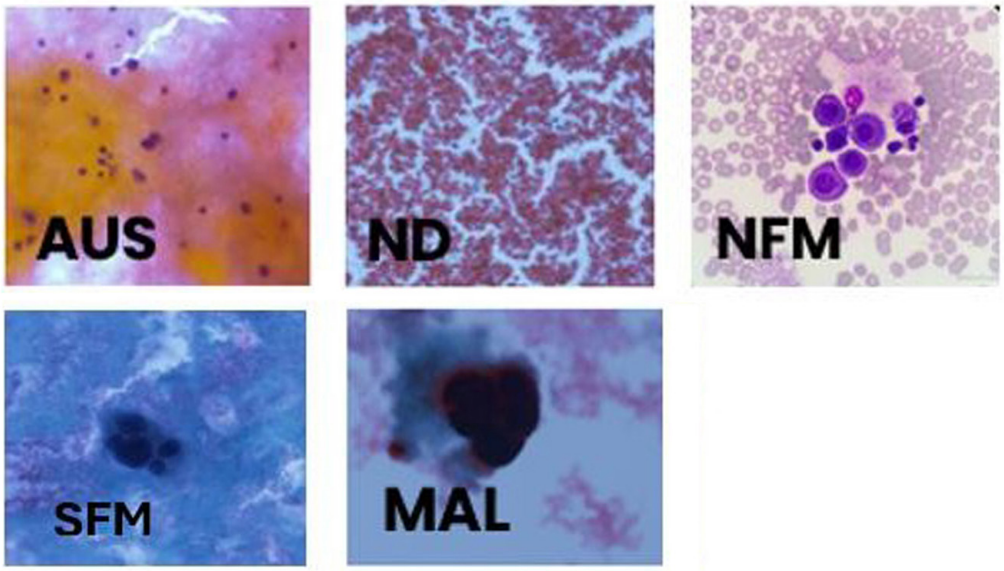

Fig. 2. Types of cells

Note: AUS Atypia of undetermined significance, ND Not diagnosed, NFM Negative for malignancy, SFM Suspicious for malignancy, and MAL Malignant.

5.3 Compliance with ethical and legal standards

The corresponding procedures were carried out before the Institutional Research Ethics Committee of the HNCH, obtaining authorization registered under code 102-2024, dated 12/12/2024. Likewise, the anonymity of the images of the patients was guaranteed, ensuring that the privacy and confidentiality regulations and the protection of sensitive data were respected.

5.4 Image preprocessing

At this point, a Python code was developed to execute the different image processing tasks, such as 1) naming and enumerating the images in their

respective folders, 2) configuring the path of the five classes, 3) resizing the images to a size of 250×250 pixels to facilitate training, and 4) converting to grayscales as shown in Figure 3.

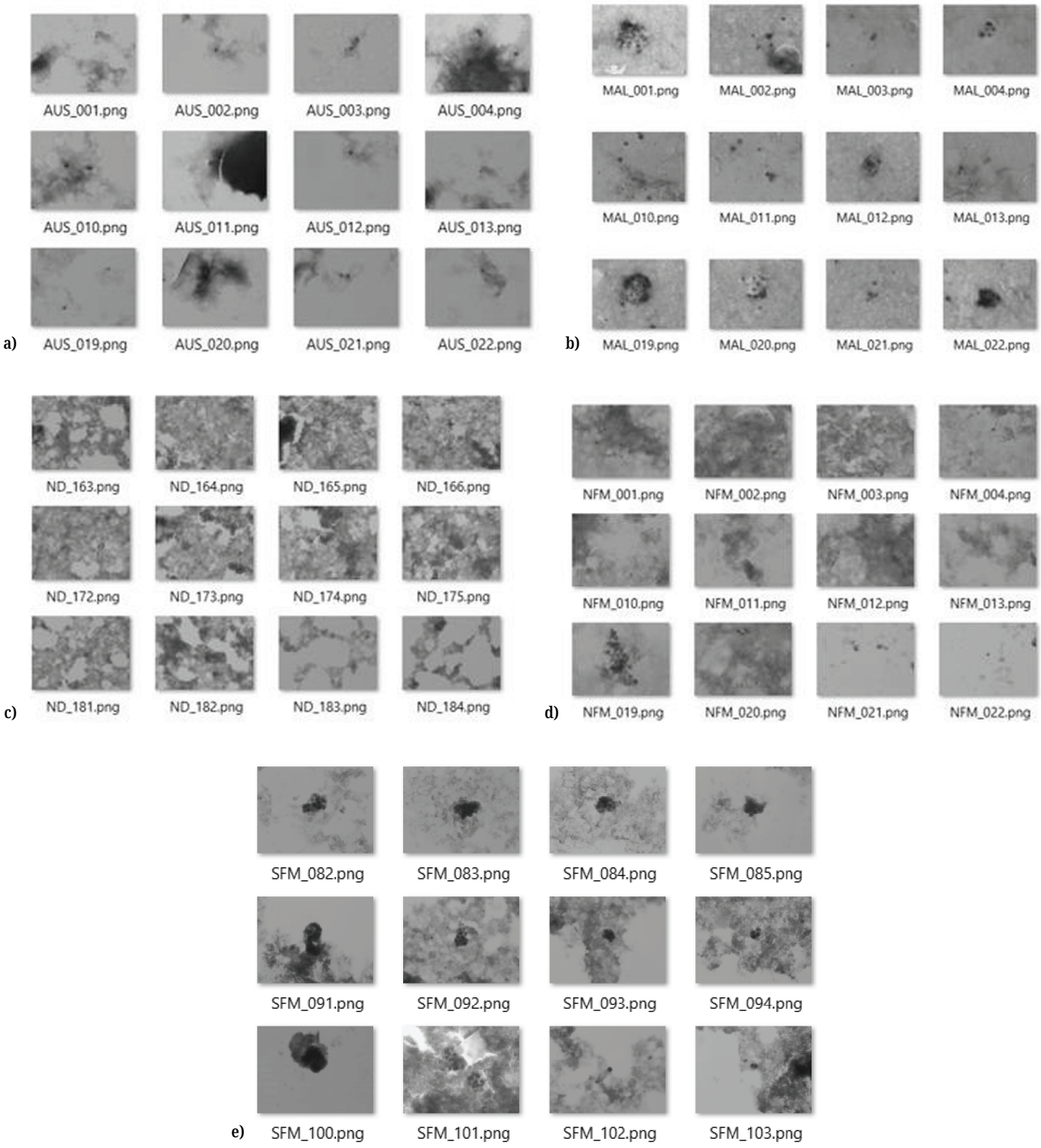

Fig. 3. Images named and numbered by category: a) AUS, b) MAL, c) ND, d) NFM, e) SFM

5.5 Documentation, publication, and access

At the request of a researcher to the corresponding author, the research team will provide the link to the dataset stored in a research repository with DOI. This dataset will be accessible to the research community under the terms of use and distribution license of the journal where this study paper will be published.

6 RESEARCH CONTRIBUTION

This new dataset will be useful for clinical diagnosis with the use of intelligent models in the health area. It also seeks to promote the development of future research in cytology, opening up the possibility of the development of personalized medicine as a form of advanced tool for the early detection and diagnosis of complex diseases. Among the main contributions that the new dataset would bring include:

- Promote the implementation of AI-based intelligent models for the detection of neoplastic cells and other pathologies in pleural fluid. The developed models will support pathologists in the interpretation of samples, allowing for an automated second opinion that reinforces diagnostic accuracy.
- Research in new diseases could be used to study cellular structure and how they affect pleural fluid at the cellular level and generate findings to improve diagnostic criteria.
- Medical training and education for physicians and students new to cytology could use the images to learn to recognize different cellular patterns associated with various pathologies.
- Validation and comparison of AI algorithms, being a public dataset, researchers from different locations could compare the performance of their models under the same conditions, establishing a standard of evaluation for diagnostic algorithms for cancer and other pleural diseases.

7 DISCUSSION OF RESULTS

The construction of the pleural fluid cytologic imaging dataset described in this study represents a significant advancement in diagnostic cytology, addressing critical limitations identified in previous research. Prior studies, such as [17], have underscored the potential of CNNs for automating cancer cell detection, emphasizing the necessity of well-labeled and structured datasets. However, existing cytologic imaging datasets often lack standardization in labeling and categorization, which limits their applicability to deep learning tasks. In contrast, our dataset, comprising 2,640 carefully labeled images categorized according to the TIS, ensures both consistency and clinical relevance, providing an essential resource for training deep learning models with high sensitivity and specificity.

A comparative analysis with existing datasets reveals several key advantages of the proposed resource. Unlike general cytology datasets, which aggregate diverse serous fluid samples without distinguishing pleural-specific characteristics, our dataset focuses exclusively on pleural fluid cytology, offering a targeted approach for model training. While [29] developed a dataset for serous fluids, its scope does not specifically address the unique cytologic features of pleural samples. In contrast, our dataset introduces a structured classification into five diagnostic categories

(AUS, NFM, SFM, MAL, and ND), enabling more precise case stratification and facilitating model interpretability.

Another major challenge in cytologic diagnosis, as noted by [15], is the variability in sample quality and the reliance on observer expertise. Our dataset mitigates these concerns through rigorous expert review by pathologists with over 20 years of experience, ensuring high reliability and diagnostic accuracy. Additionally, ethical and legal compliance measures reinforce the dataset's credibility and potential for widespread use in clinical applications.

Methodologically, our dataset advances previous efforts by incorporating standardized imaging preprocessing using Python-based tools, improving data uniformity, and minimizing artifacts that could affect model performance. This approach directly addresses the concerns raised by [18], who highlighted the necessity of structured datasets tailored for deep learning applications in cytology.

Despite its strengths, this study acknowledges certain limitations. Image quality may be influenced by capture conditions, including the imaging device, preparation techniques, and the skill of the specialist. Furthermore, annotation requires the expertise of pathologists, making it a costly and potentially variable process. Additionally, the challenge of assembling a sufficiently large and diverse dataset remains a critical factor in the field of pleural fluid cytology.

8 CONCLUSION

During the process of building the dataset, the following phases were carried out: image capture, image labeling by pathologists, compliance with ethical regulations, preprocessing, and publication in a research repository with a DOI.

Image capture was performed using a microscope with 20× and 60× magnification lenses, identifying and focusing on suspicious cells. The camera integrated into the microscope transferred the image to the computer, where the final focus was adjusted, the image was captured in color, and it was stored in diagnostic folders.

Two expert pathologists, with 20 and eight years of experience, reviewed and labeled cytologic images into categories according to malignancy (AUS, MAL, ND, NFM, SFM).

Subsequently, authorization was obtained from the HNCH Ethics Committee (code 102-2024, 12/12/2024), and the anonymity of the images was guaranteed, ensuring privacy, confidentiality, and compliance with sensitive data protection regulations.

The images were then preprocessed using Python code to organize and enumerate them. The five class paths were configured, the images were resized to 250×250 pixels, and they were converted to grayscale.

As a final result, a dataset of 2,640 images was obtained, distributed into five categories according to the TIS for reporting serous fluid cytology and risk of malignancy: AUS, MAL, ND, NFM, and SFM.

The construction of this new dataset of pleural fluid cytology images will enable future researchers to develop new models capable of detecting disease. These models could assist pathologists in more accurately identifying neoplastic cells, reducing human error, and improving diagnostic accuracy. Additionally, it could serve as an educational resource for students and health professionals, helping them improve their diagnostic skills in a controlled environment. Finally, the proposed methodology will provide researchers with a reference standard for creating cytological image datasets, fostering the development and advancement of machine learning modeling research.

For future work, we plan to investigate other pleural fluid-related pathologies, standardize cytologic images to create collaborative datasets, and develop datasets for body fluids such as peritoneal or pericardial fluids, applying the proposed methodology.

9 ACKNOWLEDGMENTS

We thank Dr. Catherinne Amaro and Dr. Jaime Cáceres Pizarro of the Department of Anatomic Pathology, Hospital Nacional Cayetano Heredia, for the excellent set of classified medical images and for the explanations and suggestions on the medical aspects of this study.

10 FINANCIAL SUPPORT

This study was supported by Universidad Nacional Mayor de San Marcos.

11 REFERENCES

[1] P. Mazzarello, "A unifying concept: The history of cell theory," *Nat. Cell Biol.*, vol. 1, pp. E13–E15, 1999. https://doi.org/10.1038/8964

[2] J. C. Caicedo, S. Singh, and A. E. Carpenter, "Applications in image-based profiling of perturbations," *Current Opinion in Biotechnology*, vol. 39, pp. 134–142, 2017. https://doi.org/10.1016/j.copbio.2016.04.003

[3] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017. https://doi.org/10.1016/j.media.2017.07.005

[4] B. Baykal, O. F. Sarioglu, and O. Aydin, "The role of cytology in the diagnosis of malignant pleural effusions," *J. Cytol.*, vol. 37, no. 1, pp. 1–7, 2020. https://doi.org/10.4103/JOC.JOC_123_19

[5] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Medicine*, vol. 17, p. 195, 2019. https://doi.org/10.1186/s12916-019-1426-2

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," Accessed: vol. 4, 2025. http://lmb.informatik.uni-freiburg.de/

[7] M. Shifat-E-Rabbi, X. Yin, C. E. Fitzgerald, and G. K. Rohde, "Cell image classification: A comparative overview," *arXiv preprint arXiv:1906.03316*, 2019. http://arxiv.org/abs/1906.03316

[8] J. C. States, L. A. Donehower, and D. Pinkel, "Cytogenetic and molecular cytogenetic studies of malignant effusions," *Cancer Cytopathology*, vol. 121, no. 1, pp. 47–57, 2013. https://doi.org/10.1002/cncy.21209

[9] A. H. Jafarian, A. Tasbandi, and N. Mohamadian Roshan, "Evaluation of photoshop based image analysis in cytologic diagnosis of pleural fluid in comparison with conventional modalities," *Diagnostic Cytopathology*, vol. 46, no. 7, pp. 578–583, 2018. https://doi.org/10.1002/dc.23952

[10] L. Pairman, L. E. L. Beckert, M. Dagger, and M. J. Maze, "Evaluation of pleural fluid cytology for the diagnosis of malignant pleural effusion: A retrospective cohort study," *Internal Medicine Journal*, vol. 52, no. 7, pp. 1154–1159, 2022. https://doi.org/10.1111/imj.15725

[11] L. G. Koss, *Diagnostic Cytology and Its Histopathologic Bases*, 5th ed. Philadelphia, PA: Lippincott Williams & Wilkins, 2020.

[12] R. W. Light, *Pleural Diseases*. Philadelphia, PA: Lippincott Williams & Wilkins, 2013.
[13] H. Müller, A. Kosem, and R. Schaer, "Automated cytology: Current and future diagnostic applications," *Acta Cytol.*, vol. 62, no. 3, pp. 177–188, 2018. https://doi.org/10.1159/000490504
[14] Y. Xu, A. Y. Hu, S. M. Wang, Q. Wang, Y. C. Pan, and S. H. Zhang, "A retrospective analysis of pleural effusion specimens based on the newly proposed international system for reporting serous fluid cytopathology," *Diagnostic Cytopathology*, vol. 49, no. 9, pp. 997–1007, 2021. https://doi.org/10.1002/dc.24804
[15] J. M. Porcel, R. W. Light, and A. Esquerda, "Diagnostic approach to pleural effusion in adults," *Am Fam Physician*, vol. 90, no. 2, pp. 99–104, 2014.
[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015. https://doi.org/10.1038/nature14539
[17] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, 2017. https://doi.org/10.1038/nature21056
[18] P. Sanyal, "Body cavity fluid cytology images," *Kaggle*, 2022. Accessed: Mar. 5, 2025. [Online]. Available: https://www.kaggle.com/datasets/cmacus/body-cavity-fluid-cytology-images
[19] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, pp. 211–252, 2015. https://doi.org/10.1007/s11263-015-0816-y
[20] J. Yauri, M. Lagos, H. Vega-Huerta, P. De-La-Cruz-VdV, G. L. E. Maquen-Niño, and E. Condor-Tinoco, "Detection of epileptic seizures based-on channel fusion and transformer network in EEG recordings," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 14, no. 5, 2023. https://doi.org/10.14569/IJACSA.2023.01405110
[21] H. Vega, E. Sanez, P. De La Cruz, S. Moquillaza, and J. Pretell, "Intelligent system to predict university students dropout," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 18, no. 7, pp. 27–43, 2022. https://doi.org/10.3991/ijoe.v18i07.30195
[22] J. V. Cubas and G. L. E. Maquen-Niño, "Machine learning model in the detection of phishing websites," *RISTI – Revista Iberica de Sistemas e Tecnologias de Informacao*, vol. 2022, no. E52, pp. 161–173, 2022.
[23] H. Vega-Huerta, K. R. Pantoja-Pimentel, S. Y. Quintanilla Jaimes, G. L. E. Maquen-Niño, P. De-La-Cruz-VdV, and L. Guerra-Grados, "Classification of Alzheimer's disease based on deep learning using medical images," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 20, no. 10, pp. 101–114, 2024. https://doi.org/10.3991/ijoe.v20i10.49089
[24] G. L. E. Maquen-Nino *et al.*, "Brain tumor classification deep learning model using neural networks," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 9, pp. 81–92, 2023. https://doi.org/10.3991/ijoe.v19i09.38819
[25] G. L. E. Maquen-Niño, J. G. Nuñez-Fernandez, F. Y. Taquila-Calderon, I. Adrianzén-Olano, P. De-La-Cruz-VdV, and G. Carrión-Barco, "Classification model using transfer learning for the detection of pneumonia in chest X-ray images," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 20, no. 5, pp. 150–161, 2024. https://doi.org/10.3991/ijoe.v20i05.45277
[26] H. Vega-Huerta *et al.*, "Classification model of skin cancer using convolutional neural network," *Ingénierie des systèmes d information*, vol. 30, no. 2, pp. 387–394, 2025. https://doi.org/10.18280/isi.300210
[27] H. Vega-Huerta *et al.*, "Convolutional neural networks on assembling classification models to detect melanoma skin cancer," *International Journal of Online and Biomedical Engineering*, vol. 18, no. 14, pp. 59–76, 2022. https://doi.org/10.3991/ijoe.v18i14.34435
[28] J. M. Porcel, "Biomarkers in the diagnosis of pleural diseases: A 2018 update," *Therapeutic Advances in Respiratory Disease*, 2018. https://doi.org/10.1177/1753466618808660

[29] E. Abd-Almoniem, N. Abd-Alsabour, S. Elsheikh, R. R. Mostafa, and Y. F. Elesawy, "A novel validated real-world dataset for the diagnosis of multiclass serious effusion crytology according to the international system and ground-truth validation data," *Acta Cytol.*, vol. 68, no. 2, pp. 160–170, 2024. https://doi.org/10.1159/000538465

[30] J. Zeng, K. Luo, Y. Lu, and M. Wang, "An evaluation framework for online courses based on sentiment analysis using machine learning," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 18, no. 18, pp. 4–22, 2023. https://doi.org/10.3991/ijet.v18i18.42521

[31] M. M. M. Alghamdi, M. Y. H. Dahab, and N. H. A. Alazwary, "Enhancing deep learning techniques for the diagnosis of the novel coronavirus (COVID-19) using X-ray images," *Cogent Eng.*, vol. 10, no. 1, 2023. https://doi.org/10.1080/23311916.2023.2181917

[32] L. R. Ali, S. A. Jebur, M. M. Jahefer, and B. N. Shaker, "Employing transfer learning for diagnosing COVID-19 disease," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 18, no. 15, pp. 31–42, 2022. https://doi.org/10.3991/ijoe.v18i15.35761

[33] O. Iparraguirre-Villanueva and M. Cabanillas-Carbonell, "Application of convolutional neural networks in skin disease prediction: Accuracy and efficiency in dermatological image analysis," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 21, no. 2, pp. 18–37, 2025. https://doi.org/10.3991/ijoe.v21i02.52871

[34] H. A. Phoulady and P. R. Mouton, "A new cervical cytology dataset for nucleus detection and image classification (Cervix93) and methods for cervical nucleus detection," *arXiv preprint arXiv:1811.09651*, 2018. [Online]. Available: http://arxiv.org/abs/1811.09651

[35] M. Aubreville, C. A. Bertram, T. A. Donovan, C. Marzahl, A. Maier, and R. Klopfleisch, "A completely annotated whole slide image dataset of canine breast cancer to aid human breast cancer research," *Scientific Data*, vol. 7, no. 417, 2020. https://doi.org/10.1038/s41597-020-00756-z

12 AUTHORS

Frida López-Córdova is a Professor at the Universidad Nacional Mayor de San Marcos in Lima, Peru. She is a PhD researcher in Systems Engineering and Computer Science. She has published articles on her research and is a member of the YACHAY Research Group at UNMSM (E-mail: frida.lopez@unmsm.edu.pe).

Hugo Vega-Huerta is a Principal professor at Universidad Nacional Mayor de San Marcos in Lima, Perú. He is a researcher specialized in Artificial Intelligence. He was the Academic Vice Dean at the Faculty of Systems Engineering at UNMSM and the Director of the Software Engineering Program at URP. He is responsible for the YACHAY Research Group at UNMSM (E-mail: hvegah@unmsm.edu.pe).

Gisella Luisa Elena Maquen-Niño is a professor at Pedro Ruiz Gallo National University, Lambayeque, Peru, and aresearcher specializing in Artificial Intelligence, Machine Learning and currently conducting research in image processing. Postgraduate studies in Machine Learning, Deep Learning, and its Applications in Industry at the San Pablo Catholic University of Arequipa- Perú, Doctor's degree in Computer Science and Engineering from the Señor de Sipan University – Peru and master's degree in Information Technology and Educational Informatics (E-mail: gmaquenn@unprg.edu.pe).

Jaime Cáceres-Pizarro is a specialist in Pathology and Laboratory Medicine at the Cayetano Heredia National Hospital. He is a Professor in the Pathology Section of the UPCH. He has published research articles in his specialty (E-mail: jaime.caceres@upch.pe).

Ivan Adrianzén-Olano is a Professor at Toribio Rodríguez de Mendoza de Amazonas National University, Amazonas, Peru, researcher specialized in

Artificial Intelligence and Machine Learning. Concluded studies of master's in systems Engineering with mention in Information Systems at the Antenor Orrego Private University, Master in Educational Sciences with Teaching and University Management from Pedro Ruiz Gallo National University (E-mail: ivan.adrianzen@untrm.edu.pe).

Oscar Benito-Pacheco is a Professor at Universidad Nacional Mayor de San Marcos in Lima, Perú. He is a researcher specialized in Software Engineering and Artificial Intelligence. He developed his research work in the master's degree on requirements engineering for agile methods. He is a member of the Yachay research Group at UNMSM (E-mail: obenitop@unmsm.edu.pe).