Data-Centric Evaluation of Image Segmentation Under Dataset Composition Shift

Graciella S. Favoreto[1], **Marcel Koenigkam Santos**[2], **Agma J. M. Traina**[1]

[1]Institute of Mathematics and Computer Science - University of São Paulo - USP
São Carlos - SP, Brazil

[2]Ribeirão Preto Medical School (FMRP), University of São Paulo - USP
Ribeirão Preto, SP, Brazil

graciellafavoreto@usp.br, marcelk46@fmrp.usp.br, agma@icmc.usp.br

Abstract. *Lung segmentation in chest X-ray images is a key step in computer-aided diagnosis, but segmentation quality often degrades under domain shift, particularly in unseen conditions. This work investigates the impact of dataset composition on U-Net-based models trained and evaluated across two public datasets. A basic U-Net and encoder-based variants (ResNet-34, ResNet-50, and VGG-16) are compared under different protocols. Results indicate that segmentation performance is highly sensitive to dataset composition and evaluation protocol, especially in pathological samples. These findings suggest that pathological variability is a major factor in degradation and highlight the importance of data-centric evaluation for reliable benchmarking in medical image analysis.*

1. Introduction

Medical imaging datasets are inherently heterogeneous, reflecting domain shifts across different hospitals, scanner vendors, imaging protocols, and patient populations. Such variability poses significant challenges for machine learning models, particularly when deployed in real-world clinical scenarios, motivating the development of transfer learning and domain adaptation techniques to address these challenges [Zhang et al. 2020]. Current studies often neglect the clinically realistic scenario in which a network trained in one setting is applied to data from another domain [Full et al. 2021]. Among medical imaging modalities, chest X-rays are among the most widely used due to their lower cost, shorter acquisition time, and reduced radiation exposure compared with other imaging techniques [Abueed et al. 2026].

Although the impact of domain shift has been extensively investigated in medical image classification, segmentation tasks have received considerably less attention [Yan et al. 2019]. Studies have shown that deep learning models may rely on irrelevant image regions, thereby limiting their generalization in real-world clinical scenarios. Therefore, lung segmentation is important for constraining the analysis to the main region of interest and reducing the influence of irrelevant information [Rahman et al. 2021].

In the context of COVID-19 diagnosis, deep learning methods have demonstrated substantially superior performance compared with traditional segmentation approaches [Chen et al. 2023], and accurate lung segmentation in chest X-ray images is a critical component for both detection and diagnosis tasks [Peng et al. 2022]. However, lung segmentation remains challenging beyond domain shift due to several factors, including non-pathological variations (as lung shape and size vary with age, gender, and heart size),

pathological opacities caused by severe lung diseases, and the presence of foreign objects such as pacemakers, infusion lines, and medical catheters [Liu et al. 2022].

Deep learning approaches have emerged as promising solutions for automated medical image segmentation. In particular, U-Net-based architectures have received considerable attention because of their ability to capture multiscale spatial context while preserving fine structural details within segmented regions [Saifullah et al. 2025]. Particularly in classification studies, a common issue observed in medical imaging is the substantial decline in model performance when evaluated on out-of-distribution (OOD) data. Furthermore, many studies focus on a single imaging modality, limiting the understanding of cross-domain generalization behavior [Guo et al. 2026].

In this work, we present an experimental study investigating the impact of dataset composition on the generalization capability of lung semantic segmentation models under domain shift. To this end, we design a set of controlled experimental protocols that isolate the effects of class distribution and pathological diversity. The main contributions of this study are: (i) an analysis of the effects of class imbalance, domain shift, and pathological variability on semantic segmentation performance; (ii) evidence that pathological diversity, rather than domain shift alone, is the primary factor associated with cross-domain performance degradation; and (iii) a comparative evaluation across multiple public datasets demonstrating how test dataset composition influences segmentation metrics.

2. Related Work

The U-Net architecture [Ronneberger et al. 2015] has become one of the most widely adopted models for medical image segmentation due to its ability to capture both local and global contextual information through its encoder-decoder structure. A common strategy in the literature is to enhance the U-Net by replacing its encoder with more powerful backbones, such as ResNet and EfficientNet, to improve feature extraction [Wang et al. 2020]. For instance, hybrid architectures such as ResUNet50 have demonstrated strong performance in medical imaging tasks, as demonstrated by a study that employed ResNet50 as the encoder of a U-Net for brain tumor segmentation, achieving Dice scores above 0.95 and Jaccard indices exceeding 0.91 [Saifullah et al. 2025]. Similarly, another study proposed a lung segmentation model that uses EfficientNet-B4 as the encoder and combines residual blocks with LeakyReLU in the decoder, achieving Jaccard indices of 95.8% and 95.5% on the JSRT and Montgomery County datasets, respectively [Liu et al. 2022]. Despite these advances, most works focus on architectural improvements under fixed training and testing conditions, with limited analysis of how these models behave under distribution shifts across datasets.

The performance of a U-Net model trained on a specific source domain may degrade significantly when applied to a distinct target domain with different acquisition settings or equipment. Moreover, obtaining large-scale annotated datasets for each new domain is costly, labor-intensive, and often impractical, especially given the challenges related to medical image acquisition, organization, labeling, and quality assurance [Yan et al. 2019, Yoon et al. 2024].

Some studies have explored domain shift in segmentation models. For example, improved cross-vendor cardiac cine MRI segmentation across Philips, Siemens, and GE

scanners [Yan et al. 2019], while BigAug employed a deep stacked data augmentation approach designed to improve the generalization of medical image segmentation models to unseen domains [Zhang et al. 2020]. These approaches were evaluated primarily on MRI and ultrasound imaging modalities. In contrast, the present study investigates domain shift in chest X-ray lung segmentation, with a particular focus on the impact of dataset composition and pathological variability under cross-dataset evaluation scenarios.

3. Material and Methods

Figure 1 illustrates the experimental pipeline adopted in this study. The proposed framework investigates the robustness of lung segmentation models under cross-domain and pathological distribution shifts, in order to build robust and useful datasets for experiments on decision making. Unlike conventional supervised settings, all models were trained exclusively using normal chest X-ray images from a source dataset and subsequently evaluated on a different target dataset containing both normal and pathological cases. This design enables the analysis of how unseen pathological patterns affect segmentation generalization under distinct domain conditions.

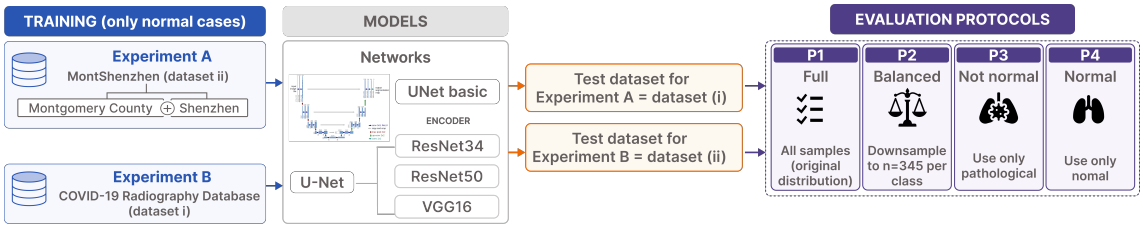

Figure 1. Experimental pipeline adopted in this work. Blue denotes the training set composed exclusively of normal images, gray indicates segmentation architectures, orange denotes the target datasets used during evaluation, and purple corresponds to the protocols applied across the experiments.

3.1. Datasets

Three publicly available chest X-ray datasets were used: (i) COVID-19 Radiography Database, comprising 3,616 COVID-19, 6,012 Lung Opacity, 1,345 Viral Pneumonia, and 10,192 Normal cases from a public Kaggle repository [Chowdhury et al. 2020, Rahman et al. 2021]; (ii) Montgomery County dataset, with 58 tuberculosis (TB) and 80 normal cases with annotated lung masks; and (iii) Shenzhen dataset, containing 662 images (326 normal, 336 TB) [Candemir et al. 2013, Jaeger et al. 2013], with 566 manually segmented lung masks from a public Kaggle source [Stirenko et al. 2018], samples without corresponding masks were excluded. The Montgomery and Shenzhen datasets were merged into a single dataset, hereafter referred to as *MontShenzhen*, resulting in 345 TB and only 359 normal images with valid lung masks. To ensure a fair comparison, each model was trained with 345 normal samples, corresponding to the smallest normal class size among the evaluated datasets. The normal images used in each experiment were selected from the corresponding dataset.

3.2. Experimental Setup

Two cross-dataset evaluation scenarios were designed to analyze the impact of pathological variability on segmentation robustness. In both scenarios, the models were trained

exclusively on normal chest X-ray images, without exposure to pathological samples during training. This setup allows the evaluation to focus on the models' ability to generalize to unseen pathological distributions rather than memorizing pathology-specific anatomical patterns. In **Experiment A**, the models were trained on normal images from the MontShenzhen dataset and evaluated on the COVID-19 Radiography dataset. This target dataset contains multiple heterogeneous pathological categories, including COVID-19, Lung Opacity, and Viral Pneumonia, as well as normal cases. Therefore, this experiment evaluates robustness under a highly diverse pathological distribution shift. In **Experiment B**, models were trained on normal images from COVID19 dataset and evaluated on a combined Tuberculosis (TB) dataset (Montgomery + Shenzhen), representing a more homogeneous pathological distribution. In contrast to Experiment A, this scenario enables the analysis of generalization under a simpler and more consistent domain shift.

Four evaluation protocols were adopted to analyze the influence of dataset composition: (P1) evaluates the full dataset using the original class distribution, with 21,165 samples from Dataset (i) and 704 from Dataset (ii); (P2) performs class-balanced evaluation with $N = 345$ samples per class; (P3) evaluates only pathological samples, totaling 10,973 samples from Dataset (i) and 345 from Dataset (ii); and (P4) evaluates only normal samples, with 345 samples from each dataset. Together, these protocols isolate the effects of class imbalance, pathological diversity, and domain shift on segmentation performance. All models used the same training configuration for comparability: input images resized to 256×256, Adam optimization with binary cross-entropy loss applied to the lung region, and Dice coefficient as the primary evaluation metric. Experiments were conducted on an *Alienware Aurora R15* workstation with an Intel Core i9-13900K CPU, 32 GB RAM, and an NVIDIA RTX 4070 Ti GPU (12 GB VRAM).

4. Results and Discussion

Figure 2 summarizes the segmentation performance across all protocols and architectures. Overall, the results demonstrate that dataset composition and pathological heterogeneity strongly affect segmentation robustness under cross-domain evaluation. Experiment A showed substantially higher degradation than Experiment B, indicating that heterogeneous pathological distributions produce more challenging domain shifts.

A clear reduction in Dice and Sensitivity scores was observed in Protocol P3 of Experiment A, where only pathological samples were evaluated. In contrast, performance recovered in P4, which contains only normal samples and therefore presents a distribution closer to the training data. These findings suggest that pathological variability, rather than class imbalance alone, is the primary factor associated with segmentation degradation. Although P1 contains severe class imbalance, its results remained closer to P2 than to P3, reinforcing that pathology diversity has greater impact than sample proportion.

Although all architectures preserved the global lung structure, qualitative differences were evident. U-Net-R34 produced more anatomically coherent contours, whereas U-Net-R50 produced larger irregular regions and false positive areas. These qualitative observations are consistent with the quantitative reductions observed in Dice and Sensitivity under pathological evaluation protocols, an example is illustrated in Figure 3.

Conversely, Experiment B remained highly stable across all protocols, with Dice scores consistently above 0.95 for all architectures. Figure 2 illustrates this behavior

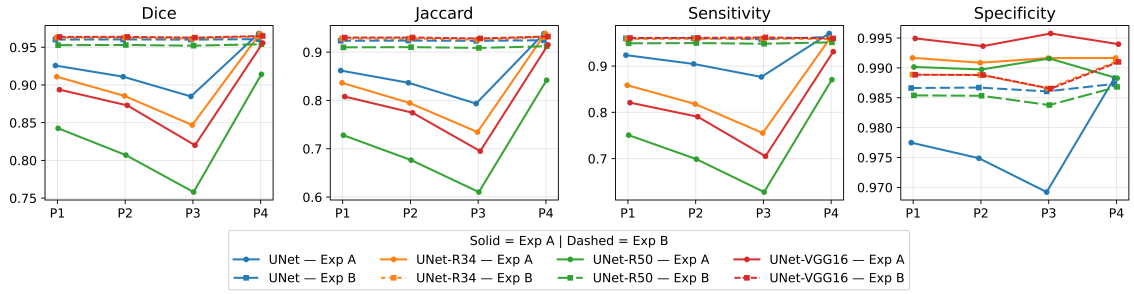

Figure 2. Segmentation performance across evaluation protocols

through the near-overlapping curves observed across protocols, indicating minimal metric variation under the tuberculosis distribution. This contrasts with Experiment A, where Dice and Sensitivity progressively decrease toward P3, particularly for architectures more affected by pathological heterogeneity. The reduced dispersion observed in Experiment B suggests that its pathological distribution introduces substantially less domain shift than the heterogeneous patterns present in the COVID-19 dataset.

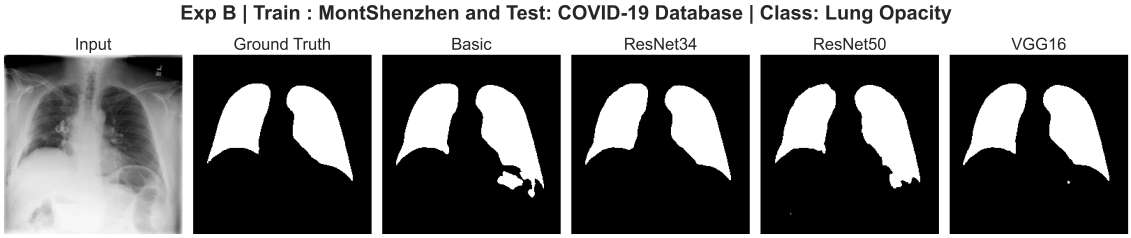

Figure 3. Example of segmentation with all architectures in worst scenario, trained in MontShenzhen and evaluated on the most affected pathological class of the COVID-19 dataset, as identified in the boxplot analysis.

Figure 4 further supports these observations through class-level Dice distributions. In Experiment A, Lung Opacity presented both the lowest median Dice values and the highest variability, indicating that this pathological pattern produces more inconsistent anatomical representations for segmentation models trained only on normal images. Conversely, Normal samples showed compact distributions and higher stability. Experiment B produced considerably tighter distributions for both classes, reinforcing the lower variability of the tuberculosis dataset.

These results emphasize the importance of data-centric evaluation protocols for medical image segmentation. Models evaluated only on globally aggregated metrics may appear robust despite exhibiting substantial degradation on underrepresented or heterogeneous pathological groups. Therefore, the combination of balanced evaluation, pathology-specific analysis, and cross-dataset validation provides a more reliable framework for assessing segmentation generalization under realistic clinical distributions.

5. Conclusion and Future Work

This study evaluated the robustness of U-Net-based lung segmentation models under cross-dataset domain shift using complementary evaluation protocols. The results demonstrate that protocol-based evaluation exposes generalization limitations more effectively

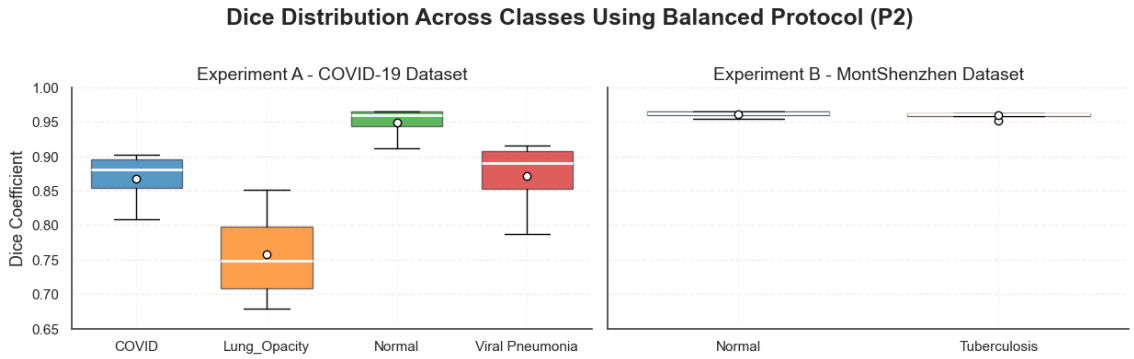

Figure 4. Class-level Dice score distributions under the balanced protocol (P2)

than conventional full dataset analysis, particularly in the presence of heterogeneous pathological patterns. Models trained exclusively on normal images achieved the best performance on normal samples of another dataset, while pathological cases produced noticeable degradation, especially in the Lung Opacity class. In contrast, the tuberculosis dataset produced substantially more stable results. Among the evaluated architectures, U-Net basic and U-Net-R34 achieved the best balance.

Although this study was designed as an experimental analysis of domain shift, training exclusively on normal samples represents an important limitation. Future work should investigate the opposite scenario by training models on more challenging pathological classes, such as Lung Opacity, to assess whether exposure to greater pathological variability improves generalization. Additionally, future analyses will incorporate statistical significance tests and explainability methods to better determine whether performance improvements correspond to meaningful and clinically relevant segmentation gains.

Acknowledgment

This work was supported by the São Paulo Research Foundation (FAPESP - grants #2024/13328-9 and #2025/15353-3) and the Coordination for Higher Education Personnel Improvement (CAPES - Finance Code 001, 88887.008174/2024-00).

References

Abueed, O. et al. (2026). Automatic semantic segmentation in chest x-ray images using deep learning approaches: a literature review. *Neural Computing and Applications*, 38(4):70.

Candemir, S. et al. (2013). Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE transactions on medical imaging*, 33(2):577–590.

Chen, Y., , et al. (2023). Ldanet: Automatic lung parenchyma segmentation from ct images. *Computers in Biology and Medicine*, 155:106659.

Chowdhury, M. E. H. et al. (2020). Can ai help in screening viral and covid-19 pneumonia? *IEEE Access*, 8:132665–132676.

Full, P. M. et al. (2021). Studying robustness of semantic segmentation under domain shift in cardiac mri. In Puyol Anton, E., Pop, M., Sermesant, M., Campello, V., Lalande, A.,

Lekadir, K., Suinesiaputra, A., Camara, O., and Young, A., editors, *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*, pages 238–249, Cham. Springer International Publishing.

Guo, B. et al. (2026). The impact of scanner domain shift on deep learning performance in medical imaging: an experimental study. *International Journal of Computer Assisted Radiology and Surgery*.

Jaeger, S. et al. (2013). Automatic tuberculosis screening using chest radiographs. *IEEE transactions on medical imaging*, 33(2):233–245.

Liu, W. et al. (2022). Automatic lung segmentation in chest x-ray images using improved u-net. *Scientific Reports*, 12:8649.

Peng, T. et al. (2022). H-segnet: hybrid segmentation network for lung segmentation in chest radiographs using mask region-based convolutional neural network and adaptive closed polyline searching method. *Physics in Medicine Biology*, 67(7):075006.

Rahman, T. et al. (2021). Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in Biology and Medicine*, 132:104319.

Ronneberger, O. et al. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer.

Saifullah, S. et al. (2025). Automatic brain tumor segmentation: advancing u-net with resnet50 encoder for precise medical image analysis. *IEEE Access*, 13:43473–43489.

Stirenko, S. et al. (2018). Chest x-ray analysis of tuberculosis by deep learning with segmentation and augmentation. In *2018 IEEE 38th International Conference on Electronics and Nanotechnology (ELNANO)*, pages 422–428.

Wang, X. et al. (2020). Automatic segmentation of pneumothorax in chest radiographs based on a two-stage deep learning method. *IEEE Transactions on Cognitive and Developmental Systems*, 14(1):205–218.

Yan, W. et al. (2019). The domain shift problem of medical image segmentation and vendor-adaptation by unet-gan. In Shen, D., Liu, T., Peters, T. M., Staib, L. H., Essert, C., Zhou, S., Yap, P.-T., and Khan, A., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 623–631, Cham. Springer International Publishing.

Yoon, J. S. et al. (2024). Domain generalization for medical image analysis: A review. *Proceedings of the IEEE*, 112(10):1583–1609.

Zhang, L. et al. (2020). Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE Transactions on Medical Imaging*, 39(7):2531–2540.