Introducing SPINE: A Holistic Approach to Synthetic Pulmonary Imaging Evaluation Through End-to-End Data and Model Management

Nikolaos Ntampakis, Vasileios Argyriou, Konstantinos Diamantaras, Konstantinos Goulianas, Panagiotis Sarigiannidis, *Member, IEEE*, and Ilias Siniosoglou

***Abstract*—In the evolving field of medical imaging and machine learning (ML), this paper introduces a novel framework for evaluating synthetic pulmonary imaging aiming to assess synthetic data quality and applicability. Our study concentrates on synthetic X-ray chest images, crucial for diagnosing respiratory diseases. We employ SPINE (Synthetic Pulmonary Imaging Evaluation) framework, a threefold synthetic images evaluation method including expert domain assessment, statistical data analysis and adversarial evaluation. In order to replicate and validate our methodology, we followed an End-to-End data and model management process which begins with a dataset of Normal and Pneumonia chest X-rays, generating synthetic images using Generative Adversarial Networks (GANs) and training a baseline classifier, essential in the adversarial evaluation axis, testing synthetic images against real data assessing their predictive value. The critical outcome of our approach is the post-market analysis of synthetic images. This innovative method evaluates synthetic images using clinical, statistical, and scientific criteria independently from traditional generation performance metrics. This independent evaluation provides deep insights into the clinical and research effectiveness of the synthetic data. By ensuring these images mirror real data's statistical properties and maintain clinical accuracy, our framework establishes a new standard for the ethical and reliable use of synthetic data in medical imaging and research.**

Manuscript received 29 December 2023; revised 25 May 2024 and 8 July 2024; accepted 8 July 2024. Date of publication 12 July 2024; date of current version 6 August 2024. This work was supported by the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement 101095435 (REALM). The review of this article was arranged by Editor Andrew F. Laine. *(Corresponding author: Nikolaos Ntampakis.)*

Nikolaos Ntampakis is with the Department of Information & Electronic Engineering, International Hellenic University, 57001 Sindos, Greece, and also with the MetaMind Innovations, 50100 Kozani, Greece (e-mail: nikontam1@iee.ihu.gr).

Vasileios Argyriou is with the Kingston University London, KT2 7LB London, U.K. (e-mail: vasileios.argyriou@kingston.ac.uk).

Konstantinos Diamantaras and Konstantinos Goulianas are with the Department of Information & Electronic Engineering, International Hellenic University, 57001 Sindos, Greece (e-mail: kdiamant@ihu.gr; gouliana@ihu.gr).

Panagiotis Sarigiannidis and Ilias Siniosoglou are with the MetaMind Innovations, 50100 Kozani, Greece, and also with the University of Western Macedonia, 50100 Kozani, Greece (e-mail: psarigiannidis@uowm.gr; isiniosoglou@uowm.gr).

Digital Object Identifier 10.1109/OJEMB.2024.3426910

***Index Terms*—Framework, post-market, synthetic, X-ray.**

***Impact Statement*—The SPINE framework revolutionizes synthetic pulmonary imaging evaluation by ensuring synthetic data's clinical accuracy and statistical reliability, enhancing their ethical and practical use in medical diagnostics. By integrating expert domain assessment, statistical analysis, and adversarial evaluation, SPINE provides a comprehensive post-market evaluation method for synthetic medical data accelerating medical research, profoundly impacting the healthcare industry.**

I. Introduction

THE integration of machine learning in medical imaging has catalyzed significant advancements, reshaping the landscape of diagnostics and research. Despite these advancements, the field faces notable challenges, particularly in the ethical sourcing and use of large, diverse datasets. Systematic challenges such as data biases and the alignment of research incentives are impeding progress in this area [1]. Synthetic data generation has emerged as a viable solution to these challenges, addressing critical issues related to data availability, privacy concerns and imbalances in medical datasets.

Our paper introduces SPINE framework for evaluating synthetic pulmonary imaging, which is set to become an essential tool in medical diagnostics and research. The need for this framework is emphasized by the significant attention synthetic data has garnered in medicine and healthcare, as it can improve existing AI algorithms through data augmentation. However, there remains a lack of clarity on the wider roles of synthetic data in AI systems in healthcare, including challenges in establishing clinical-quality measures and evaluation metrics for synthetic data [2].

Our initiative aims to bridge the gap between the generation of synthetic medical images and their effective utilization in practical scenarios. While prior research has primarily focused on creating synthetic images [3], our framework shifts towards a more comprehensive evaluation of these datasets, adopting a three-pronged evaluation strategy that includes domain or expert knowledge, data statistical analysis and adversarial evaluations. This multifaceted approach is designed to compare the utility of synthetic data with real-world data, particularly in the context of predictive modeling in machine learning.

The core of our research is centered on the post-market evaluation of synthetic data. This critical phase involves assessing the synthetic data after it has been deployed, ensuring it meets the necessary standards and performs effectively in real-world applications. To ensure comprehensive control over the entire process, we consciously decided not to rely on pre-existing open-source synthetic medical data. Instead, we adopted a holistic approach tailored to the specific needs of this medical data generation case. This approach acknowledges that different medical scenarios necessitate tailored adjustments due to the varying data types (images, tabular, volumetric, etc.) and the specific expert knowledge required for different medical conditions. Nevertheless, our proposed framework, applied to lung X-ray images for detecting or excluding pneumonia, is designed to be versatile and adaptable to any synthetic lung X-ray imaging scenario, demonstrating the broader applicability of our method.

Central to our validation methodology was the use of an open-source dataset from Mendeley Data, specifically "Chest X-Ray Images (Pneumonia)" [4]. This dataset is categorized into two primary categories: Normal and Pneumonia, including chest X-rays of pediatric patients aged one to five years from Guangzhou Women and Children's Medical Center. The dataset was not only pivotal in generating synthetic images via Generative Adversarial Networks (GANs) but also formed the foundation for training our baseline classifier. This classifier trained to differentiate between Normal and Pneumonia cases, is integral to the adversarial evaluation phase of our framework.

This research establishes a new standard for the ethical and reliable use of synthetic data in medical imaging and research by implementing a rigorous evaluation process that ensures synthetic images maintain clinical accuracy and mirror the statistical properties of real data. By independently evaluating synthetic images through clinical, statistical, and adversarial lenses, our framework provides a robust methodology that enhances the usability and reliability of synthetic data in clinical decision-making and machine learning. This comprehensive approach addresses the ethical considerations and practical challenges associated with synthetic data, promoting more ethical, privacy-aware, and effective data usage in medical research.

In summary, this paper introduces several contributions to the integration of machine learning in medical imaging. The key innovations include:

- Novel Framework for Synthetic Pulmonary Imaging: Introduction of a novel framework (SPINE) to evaluate synthetic pulmonary imaging, addressing the urgent need for ethical and effective use of synthetic data in medical diagnostics and AI systems.
- Three-Pronged Evaluation Strategy: This paper adopts a comprehensive and versatile evaluation approach, blending domain expertise, statistical data analysis and adversarial evaluations. While this strategy is meticulously applied to chest X-rays in the current work, its underlying logic and methodology are designed to be universally applicable across various categories of medical data. This adaptability allows for the effective comparison of synthetic and real-world data's utility in predictive medical machine learning modeling, ensuring broad applicability and potential for future extensions in different medical imaging domains.
- Post-Market Evaluation of Synthetic Data: Focus on post-generation analysis, moving beyond mere creation of synthetic images to ensure their practical utility and relevance in various medical scenarios.

The rest of the paper is organised as follows: In Section II an overview of existing literature is provided. Then, in Section III, we delve into the specific methods and criteria used for each evaluation axis of our study. Section IV, reports on our experimental results, subdivided into the system configuration, dataset description, evaluation metrics used and the results obtained. An ablation study is detailed in Section V and the paper concludes with Section VI summarizing our key findings and suggesting potential areas for future research.

II. Related Work

The evaluation of synthetic medical data is a crucial aspect of this field, yet it has received comparatively less focus in the literature. Current studies primarily focus on the generation of synthetic data, with less emphasis on comprehensive evaluation frameworks. For instance, Goncalves A. et al. [5] highlight the diverse methods for generating synthetic data, noting the difficulties in directly comparing these methods due to the use of different datasets and evaluation metrics. This underscores a significant gap in the literature: a lack of standardized guidelines or discussions on how to compare and evaluate different synthetic data generation methods to select the most appropriate one for a given application.

Most existing evaluation approaches for synthetic medical data are fragmented, primarily focusing on individual aspects like statistical accuracy, fidelity to real-world data or privacy preservation. For example, Chao Y. et al. [6] points out that while various synthetic data generation techniques, particularly GANs, have been extensively proposed, their systematic benchmarking and evaluation remain underdeveloped. This lack of comprehensive evaluation is further highlighted by the absence of a unified set of metrics to assess synthetic data, leading to inconsistencies in model comparisons and evaluations based on Khaled E. et al. [7]. Moreover, according to Lan L. et al. [8], the evaluation often neglects the diverse use cases of synthetic data, thereby failing to ascertain the conditions ideal for specific simulation models. This scenario underscores the need for a holistic framework offering a more thorough and multifaceted assessment of synthetic medical datasets.

Reflecting on the extensive research in this field, it becomes apparent that the proposed framework in our work, integrating domain expert knowledge, data statistics and adversarial evaluation, represents a novel approach. To our knowledge, no existing studies have presented a framework that unites these three axes in the evaluation of synthetic medical data, not only for pulmonary imaging but generally in the context of medical synthetic data.

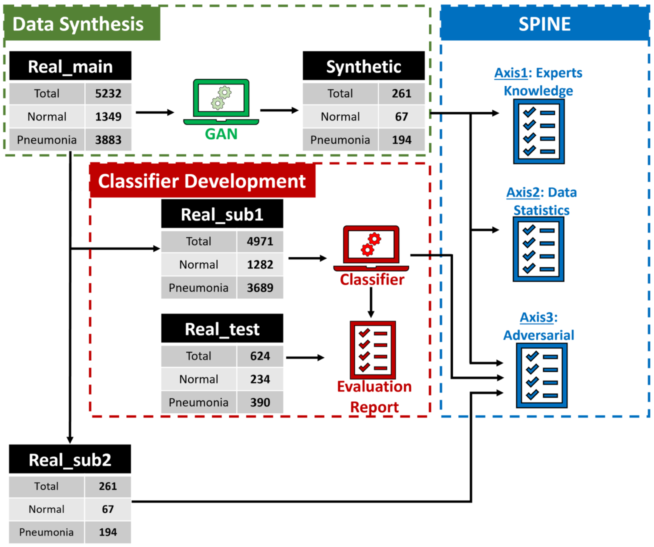

Fig. 1. SPINE development schema.

This marks our framework as a significant and innovative contribution to the field, potentially setting new benchmarks for the evaluation of synthetic medical datasets.

III. Methodology

Our development schema is structured to address the unique challenges and requirements of evaluating synthetic medical datasets, particularly in the context of chest X-rays used for distinguishing between normal patients and those with pneumonia. The core of our task, called SPINE in Fig. 1, involves the development of an evaluation framework encapsulating three axes: expert/domain knowledge evaluation, data statistics evaluation, and adversarial evaluation.

To replicate and validate our SPINE framework, we employed an End-to-End data and model management process. This process comprises two preparatory tasks, as shown in Fig. 1. The first task is Data Synthesis, where we employ a Generative Adversarial Network (GAN) to create a synthetic dataset. This process is vital to generate the synthetic data that SPINE framework will evaluate. The second task, called Classifier Development, involves developing a baseline classifier trained to differentiate between normal and pneumonia cases in chest X-rays. This step is essential for the adversarial evaluation axis of SPINE, where the classifier tests synthetic images against real data, assessing their predictive value.

Our comprehensive, step-by-step approach prioritizes control and specificity in the evaluation of synthetic medical datasets. From data synthesis and classifier development to a multifaceted evaluation framework, each step is designed to meet the unique demands of our research, contributing to the advancement of medical data analysis and synthetic data evaluation.

A. Preparatory Task 1: Data Synthesis

Generative Adversarial Networks (GANs), first introduced by Goodfellow I. et al. [9], represent a groundbreaking approach in the field of generative models. GANs consist of two neural networks, the generator and the discriminator, which are trained simultaneously through adversarial processes. The generator creates data samples, while the discriminator evaluates them against real data, fostering a continuous improvement in the quality of generated samples. Extending the capabilities of GANs, conditional Generative Adversarial Networks (cGANs) were introduced, adding a conditional aspect to the generative process. In cGANs, both the generator and the discriminator receive additional label information, allowing the generation of targeted data samples based on specific conditions or categories. This modification, as detailed by Mirza M. et al. [10], significantly enhances the control over the data generation process.

In our study, we developed a custom cGAN model aiming to generate synthetic chest X-ray images for distinguishing between normal and pneumonia cases. The generator of our cGAN model begins with two separate inputs: the dimensionality of the latent space and the number of classes for labels. Firstly, it creates a label input pathway. This starts with an input layer for the label, which is processed through an embedding layer to transform it into a dense representation. The output is then fed into a dense layer to expand its dimensions, followed by a reshape operation to form a 32 × 32 feature map. Simultaneously, a noise input pathway is created. This takes a latent space input and passes it through a dense layer, which significantly expands its dimensions. The output is then reshaped into a 32 × 32 feature map, but with a depth of 256, indicating many feature maps stacked together. These two pathways, noise and label, are then merged together using a concatenate operation. This merged tensor forms the input to a sequential model, which consists of several layers designed to upsample the input to a higher resolution image.

The upsampling is achieved through a series of transposed convolutional layers each of which doubles the dimensions of the feature map. A transposed convolutional layer, used in your generator model, serves the purpose of upsampling the input feature maps to a higher spatial resolution. It's the reverse operation of a conventional convolutional layer [11]. The mathematical operation of the transposed layers of the generator explained in the (1) below:

$$Y = X * T(F) \tag{1}$$

where Y is the output feature map from the transposed convolutional layer, X is the input feature map, T denotes the transposed convolution operation with F representing the size of the filter. The stride S controls the step of the filter across the input, affecting the upsampling, and the padding P adjusts the spatial size of the output. The stride and padding are implicit in this operation but are not explicitly represented in the equation for simplicity. A stride of (2,2) is employed to our generator when padding maintains the spatial dimensions post convolution.

These layers are interleaved with instance normalization, where it normalizes the input across each channel in each data sample independently. As activation function for the layers LeakyReLU activation [12] was used as below in (2).

$$f(x) = \begin{cases} x & \text{if } x > 0, \\ \alpha x & \text{otherwise} \end{cases} \tag{2}$$

$f(x)$ is the output of the LeakyReLU function given an input x, and α is a constant of 0.01 that defines the slope of the function for negative input values.

The model progressively upsamples the feature maps from 32 × 32 to 64 × 64, then to 128 × 128, and finally to 256 × 256. The last layer of the model is another transposed convolutional layer that outputs a single 256 × 256 feature map, with a hyperbolic tangent function(tanh) activation function. Finally, the model connects the merged input to the sequential model and returns the complete generator model. This model takes noise and a class label as input and generates an image corresponding to that class.

The discriminator's architecture, is tailored to assess the authenticity of the generated images. It also receives two inputs: an image input of shape (256, 256, 1) and a label input. The label is first embedded into a 50-dimensional vector, then expanded through a dense layer to match the spatial dimensions of the image input, and finally reshaped to a single-channel format. These two inputs are then concatenated along the channel dimension, creating a merged input that combines image and label information.

Within the model, a series of convolutional layers with LeakyReLU activations progressively downsample this merged input. The downsampling process involves halving the spatial dimensions at each convolutional layer, starting from 128 × 128 to 64 × 64, and finally to 32 × 32, with a stride of (2,2) at each step. LeakyReLU is employed to introduce non-linearity and prevent gradient vanishing. Dropout layers with a rate of 0.3 are interspersed between these convolutional layers to mitigate overfitting by randomly disabling a fraction of neurons during training.

The final stage of the model involves flattening the downsampled feature map into a vector, which is then passed through a dense layer with a sigmoid activation. This layer acts as a classifier, outputting a probability that indicates whether the input image-label pair is real or fake, aligning with the discriminator's role in a GAN to distinguish between real and generated images.

For preprocessing, images were resized to 256 × 256 and labels were appropriately processed. The generator and discriminator undergo alternating updates, using binary cross-entropy for loss calculation (3) and the Adam optimizer [13] for adjustments.

$$L = -\sum_i y_i \cdot \log(p_i) \tag{3}$$

In (3), L is the Cross-Entropy Loss, y_i denotes the true label for the i-th class, and p_i is the predicted probability for the i-th class by the model.

B. Preparatory Task 2: Classifier Development

We employed a state-of-the-art image classification framework, YOLO (You Only Look Once), specifically utilizing its latest iteration at the time of our research, YOLOv8. YOLO, an acronym for "You Only Look Once," was proposed initially by Redmon J. et al. [14] and is a series of computer vision algorithms capable of performing various computer vision tasks. The YOLO framework is distinguished by its unique method of processing images, where it divides the image into a grid and predicts bounding boxes and class probabilities for each grid cell simultaneously. Since its introduction, YOLO has undergone several iterations and improvements. Each version, from YOLOv1 to the latest YOLOv8, has brought enhancements in accuracy and speeds. YOLOv8 represents the latest development in this series, offering further optimizations in performance and accuracy.

In our research, we utilized YOLOv8 Nano, initialized with weights pre-trained on the MS COCO dataset [15], to create a chest X-ray image classifier for distinguishing between normal and pneumonia cases. YOLOv8 is characterized by its distinctive architecture, incorporating an anchor-free approach combined with a decoupled head. This design allows the independent handling of objectness, classification and regression tasks. Our focus particularly lies on the aspect of classification, where YOLOv8 leverages the softmax function to compute class probabilities.

The fundamental unit of the architecturein Yolov8 called C2f block, incorporates the core elements of CBS – a composition of a Convolutional layer (Conv), Batch Normalization (4) and a SiLU activation layer (5). This C2f module, characterized by "f" representing the total feature count. A significant aspect of the C2f unit is its adoption of a 3 × 3 × 3 kernel size in the convolution along with the bottleneck within this unit also utilizes a 3 × 3 × 3 kernel.

$$\mathrm{BN}(x) = \gamma \left(\frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} \right) + \beta \tag{4}$$

where x is the input, μ and σ^2 are the mean and variance, γ and β are learnable parameters, and ϵ is a small constant for numerical stability.

$$\mathrm{SiLU}(x) = x \cdot \sigma(x) \tag{5}$$

where $\sigma(x)$ is the sigmoid function, defined as $\sigma(x) = \frac{1}{1+e^{-x}}$. In this equation, x is the input to the activation function. The sigmoid function outputs a value between 0 and 1, which scales the input x.

We started by leveraging the feature extraction capabilities of YOLOv8 Nano's backbone, which was already trained on the diverse MS COCO dataset. The primary modification in our approach was the adaptation of the loss function to suit binary classification. Given the nature of our task, we employed binary cross-entropy as our loss function (3).

The focus of our training efforts was on the later layers of the YOLOv8 Nano model. These layers, more specialized for the specific task of classification, were fine-tuned to adapt to the nuances of our dataset. By retraining these layers, the model was able to better interpret the features extracted from the X-ray images and make accurate distinctions between normal and pneumonia cases.

C. SPINE Framework

The SPINE Evaluation Framework is the cornerstone of our research on evaluating synthetic medical data, focusing particularly on pulmonary imaging (lung X-rays). This framework operates on three pivotal axes: expert domain assessment, statistical data analysis and adversarial evaluation, encompassing a variety of specific criteria essential for determining the suitability of synthetic images for medical applications.

The output of each criterion for the first two axes is primarily the labeling of images as ‘correct’ or ‘erroneous’ indicating if the synthetic image aligns with expert knowledge and statistical analysis criteria. This binary classification allows for a straightforward assessment of the synthetic data initial quality. The proposed logic behind this framework suggests a continuous application of these criteria, with the aim of systematically excluding data samples labeled as ‘erroneous’ from our synthetic data repository. In the pursuit of these first two axes, we have consciously chosen not to employ sophisticated machine learning techniques. Instead, our focus is on enhancing the interpretability, reproducibility and explainability of our processes. This is achieved through the implementation of straightforward image manipulation techniques. These methods, while simpler, offer clear advantages in terms of understanding and replicating the results. Furthermore, it’s important to note that even though our specific use case and dataset focus on classifying lung X-rays between normal and pneumonia, the criteria we have developed are versatile and applicable to any synthetic lung X-ray imaging scenario, regardless of the specific medical task at hand.

On the other hand, the third axis of the framework involves a qualitative evaluation of the synthetic data, necessitating human intervention to compare the evaluation results of the inference on the baseline classifier for synthetic versus real-world data. This process focuses on assessing the predictive value of the synthetic data by thoroughly reviewing the results according to clear guidelines. This holistic approach is critical for ensuring that only the most accurate and reliable synthetic data is used, complementing the expert knowledge-based assessment of the first axis and the statistical analysis of the second axis.

The major goal of this framework is to add additional safety and evaluation layers to the use of synthetic data in medical applications, focusing particularly on excluding erroneous data. Although the dataset utilized for generating synthetic data primarily consisted of images from young children [4], the criteria established in the framework are versatile and broad-ranging, enabling their effective application across all age groups. Furthermore, wherever possible, we intend to validate these criteria against real-world data. This crucial step aims to confirm the robustness and reliability of our framework.

1) [Axis 1: Experts Knowledge] Criterion 1.1 - Thoracic Field Completeness: The thoracic field completeness is a fundamental criterion for evaluating synthetic lung X-rays, as it ensures the integrity and clinical usability of the image. In a diagnostic-quality chest X-ray, we expect to see clear delineation of the thoracic cavity, with distinct visualization of the lung fields, heart, diaphragm, and bony structures such as the ribs and spine. Typically, bones appear white or light due to their high density, contrasting with the darker appearance of the lungs and air-filled spaces, which possess lower X-ray absorption characteristics. The background, ideally homogeneous and dark, provides a contrast that helps in identifying these anatomical landmarks [16].

Algorithm 1: Thoracic Field Completeness Check.

```
1: Input: Synthetic Lung X-ray images
2: Output: Thoracic Field Completeness Check
3: procedure APPLYING OTSU’S THRESHOLDING
4:   for each image do
5:     Apply Otsu’s thresholding to create a binary image
6:     Distinguish darker areas (lungs and background)
       from brighter areas (bones and tissues)
7:   end for
8: end procedure
9: procedure LUNG-BACKGROUND SEPARATION CHECK
10:  Identify background as darker region touching image
     margins
11:  Check if segmented lung area contacts the background
     area
12: end procedure
13: procedure DETERMINING IMAGE VALIDITY
14:  if lung area is not distinct from background then
15:    Mark image as erroneous
16:  else
17:    Mark image as correct
18:  end if
19: end procedure
```

Otsu’s thresholding [17] is a classical and widely-used automatic thresholding technique in image processing to separate the foreground from the background. Otsu’s thresholding is fundamental to our research and will be used as the base technique in the Thoracic Field Completeness criterion of the SPINE framework. The method involves exhaustively searching for the threshold that minimizes the within-class variance or, equivalently, maximizes the between-class variance.

$$\sigma_B^2(t) = \omega_0(t)\omega_1(t)[\mu_0(t) - \mu_1(t)]^2 \tag{6}$$

In the (6), $\sigma_B^2(t)$ is the between-class variance, $\omega_0(t)$ and $\omega_1(t)$ are the probabilities of the two classes separated by a threshold t, and $\mu_0(t)$ and $\mu_1(t)$ are the class means. The method iteratively calculates t for each threshold level and selects the one yields the maximum $\sigma_B^2(t)$. This method is effective in distinguishing areas of interest (like lung fields) in medical images due to its ability to adaptively adjust the threshold based on the image content.

To assess the specific criterion of thoracic field completeness in synthetic chest X-ray images, we followed a structured approach as outlined in Algorithm 1. This algorithm encapsulates a series of critical steps, including Otsu’s thresholding application, lung-background separation checks, image validity determination and visualization with colored masks.

2) [Axis 1: Experts Knowledge] Criterion 1.2 - Diaphragm Existence: In a typical chest X-ray, which includes the thoracic vertebrae, lungs, heart and other thoracic structures, the diaphragm plays a vital role in the interpretation of the image [18]. Based on Thitiporn S. et al. [19] the right hemidiaphragm dome is typically positioned at 9.7 ± 0.8 cm

Algorithm 2: Diaphragm Existence Check.

```
Input: Synthetic Lung X-ray images
Output: Diaphragm Existence Check
procedure APPLYING OTSU'S THRESHOLDING
  for each image do
    Apply Otsu's thresholding to create a binary image
    Distinguish darker areas (lungs and background)
    from brighter areas (bones and tissues)
  end for
end procedure
procedure DIAPHRAGM LINE IDENTIFICATION
  for each image do
    Divide the image horizontally, focusing on the
    bottom 1/6
  end for
end procedure
procedure DETERMINING IMAGE VALIDITY
  for each image do
    Select the area below the diaphragm line
    Calculate the percentage of brighter pixels (1 s in the
    binary image) for each half of the bottom 1/6
    Determine if brighter areas exceed 50% in each half
    of the bottom part
    if brighter areas are more than 50% then
      Mark image as correct
    else
      Mark image as erroneous
    end if
  end for
end procedure
```

thoracic vertebral levels below the top of the first thoracic vertebra, with the left hemidiaphragm dome slightly lower at 10.2 ± 0.8 cm vertebral levels. The hemidiaphragms, appearing as domed structures, should be well-defined and visible up to the midline on a frontal view, with the right diaphragm visible all the way to the anterior chest wall and the left diaphragm visible up to the point where it borders the heart.

Based on the above, in terms of its proportion in a chest X-ray image, the diaphragm typically occupies the lower 1/6 of the image when split horizontally. This estimation considers the diaphragm's relative position to the liver, which elevates the right hemidiaphragm, and its relationship with the heart, which limits the visibility of the left hemidiaphragm.

On a grayscale chest X-ray, the diaphragm is distinguished by its appearance and intensity. The diaphragm should present as a well-defined, dome-shaped structure with a consistent density. It contrasts against the air-filled lungs, which appear darker due to their lower density, and the abdominal structures beneath, which are denser and thus appear lighter in color. The right hemidiaphragm, bordered by the air in the lungs and the soft tissue of the liver, provides a clear interface visible to the anterior chest wall. The left hemidiaphragm, however, becomes less distinct where it borders the heart, as both structures have similar densities. This distinction in densities and the well-defined contours of the diaphragm are key aspects in evaluating its normalcy in a chest X-ray [18].

In order to assess the presence of the diaphragm in synthetic chest X-ray images, we employed Algorithm 2. This Algorithm 2 is crafted to scrutinize each image, ensuring the precise identification of the diaphragm by labelling the 'correct" or 'erroneous' images accordingly.

3) [Axis 2: Data Statistics] Criterion 2.1 - Diagnostic Noise Level Check: To evaluate the statistical properties of our synthetic lung X-ray images, we have broadened our dataset to include the National Institutes of Health (NIH) ChestX-ray14 dataset [20]. This integration is not only crucial for enhancing the statistical robustness of our evaluation but could ensure that there is an alignment with real-world clinical standards. The ChestX-ray14 dataset is a comprehensive collection, consisting of 112,120 frontal-view X-ray images from 30,805 unique patients, each annotated with 14 different thorax disease categories.

In the process of calculating the noise in our datasets, we utilized Immerkaer's method [21], a technique in the field of image processing for estimating the noise variance. Immerkaer's method operates by calculating the local mean around each pixel in an image. It does so by considering a 3x3 window centered on each pixel. The core of this method is the estimation of variance, which it achieves by analyzing the difference between the intensity of each pixel and its local mean (7). This difference is indicative of the noise level at each pixel. The method then involves squaring these differences and summing them across the entire image.

$$\text{Noise Variance} = \frac{1}{M}\sum_{i=1}^{H-1}\sum_{j=1}^{W-1}\left(I(i,j) - \frac{1}{4}\sum_{\text{neighbors}} I(n)\right)^2 \tag{7}$$

In (7), Noise Variance is the estimated variance due to noise in the image. M is the total number of pixels in the image, adjusted for border pixels. H and W are the height and width of the image, respectively. $I(i,j)$ denotes the intensity of the pixel at position (i,j). The inner summation calculates the local mean around each pixel by averaging the intensities of its four immediate neighbors (top, bottom, left, and right), effectively capturing the local variability in pixel intensity that is indicative of noise. The method's primary objective is to quantify the noise level in an image by providing a numerical value for the noise variance. This value is a representation of how much the intensity of each pixel varies from its local mean due to noise.

In Fig. 2, we present the noise distributions of the two datasets of lung X-rays, alongside their concatenated distribution. Mendeley dataset [4] exhibits a mean noise level of 107.58, a median of 108.72, and a standard deviation of 14.04, while NIH dataset [20] has a mean of 110.62, a median of 114.12, and a standard deviation of 14.89. Despite these slight variations, the noise levels in both datasets are predominantly concentrated around similar values. This observation underscores a notable consistency in the noise characteristics of lung X-ray images across different datasets. Such consistency is crucial, especially in medical imaging, where the reliability and quality of images are paramount for accurate diagnoses.

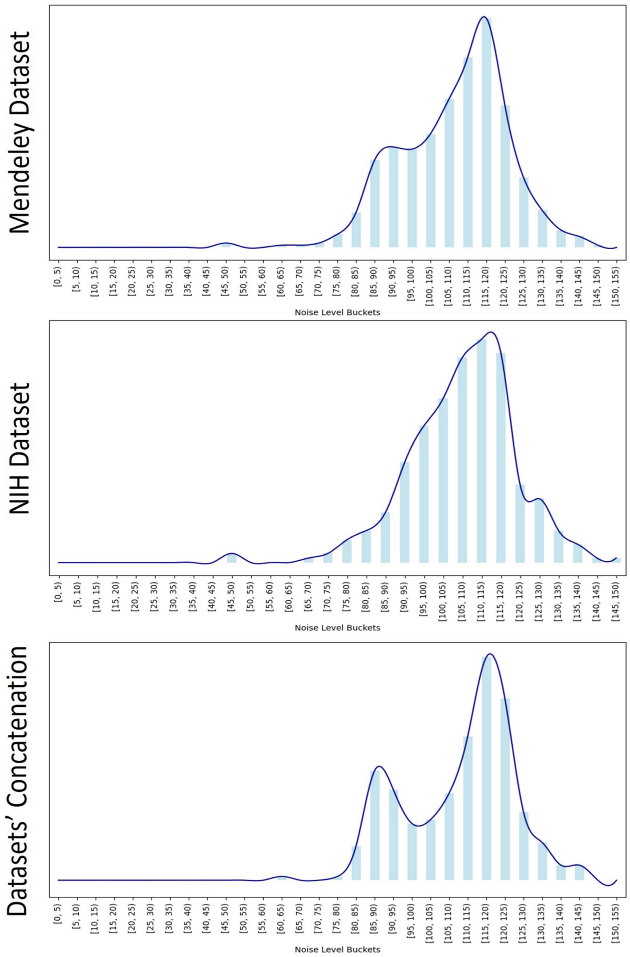

Fig. 2. Noise Distribution along Datasets.

Based on our analysis and the visual evidence from Fig. 2, we can deduce that synthetic lung X-ray images exhibiting noise levels greater than 155 are likely to be erroneous. Conversely, images with noise levels below this threshold can generally be considered correct. This assertion is grounded in the understanding that excessive noise in medical images, especially X-rays, can significantly degrade their quality. High noise levels obscure critical details, reduce image clarity, and compromise the contrast necessary for distinguishing vital anatomical structures and pathological indications. In lung X-rays, where subtleties in tissue textures and densities are often key indicators of various conditions, maintaining low noise levels is imperative to ensure the images' diagnostic utility and accuracy. Consequently, identifying and flagging synthetic images with anomalously high noise levels as erroneous becomes a crucial step in ensuring the quality and reliability of medical imaging data.

Based on the analysis of noise levels in lung X-ray images, we have developed Algorithm 3. This algorithm categorizes synthetic lung X-ray images as either correct or erroneous by assessing their noise levels using Immerkaer's method.

4) [Axis 2: Data Statistics] Criterion 2.2 - Diagnostic Sharpness Level Check: Sharpness, which essentially translates to the level of detail in medical imaging, particularly in chest X-rays, is essential. For diagnosing conditions like pneumonia, the ability to discern subtle changes in lung density is paramount. These changes, often manifesting as increased whiteness on the radiograph, are key indicators of the disease's presence and severity. According to Cleverley et al. [22], identifying these nuances, which can be exceedingly subtle, is integral to an accurate diagnosis. This reliance on fine detail accentuates the necessity of sharp, high-quality images, without which the risk of misinterpretation escalates significantly. Furthermore, the quality of chest X-rays is a significant determinant of their interpretability. Factors that degrade image quality, such as under-exposure or poor imaging due to the patient's positioning, can obscure critical details.

Algorithm 3: Diagnostic Noise Level Check.

```
Input: Synthetic Lung X-ray images
Output: Diagnostic Noise Level Check
procedure ESTIMATE NOISE USING IMMERKAER'S METHOD
    for each image do
        Estimate the noise level of the image using Immerkaer's method
    end for
end procedure
procedure DETERMINING IMAGE VALIDITY
    for each image do
        Retrieve the estimated noise level for the image
        if noise level $> 155$ then
            Mark image as erroneous
        else
            Mark image as correct
        end if
    end for
end procedure
```

$$\text{Laplacian Variance} = \text{Var}(L) \tag{8}$$

where the variance Var is calculated over the Laplacian image L, quantifying the spread of the edge information and L is calculated as:

$$L = \nabla^2 I = \frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2} \tag{9}$$

where I is the original image, ∇^2 denotes the Laplacian operator, and $\frac{\partial^2}{\partial x^2}$ and $\frac{\partial^2}{\partial y^2}$ are the second partial derivatives of the image intensity I with respect to the x and y coordinates.

For assessing the sharpness of images, a technique known as the Laplacian variance used, as in (8) and (9). This approach is well-suited for evaluating the level of detail in an image. In the context of chest X-rays, edges and fine details are paramount in revealing the structural integrity of lung tissues and the presence of pathological changes. By applying the Laplacian operator to an image, we essentially accentuate these critical edges, making it easier to assess the image's overall sharpness. A higher variance indicates a greater degree of sharpness, implying that

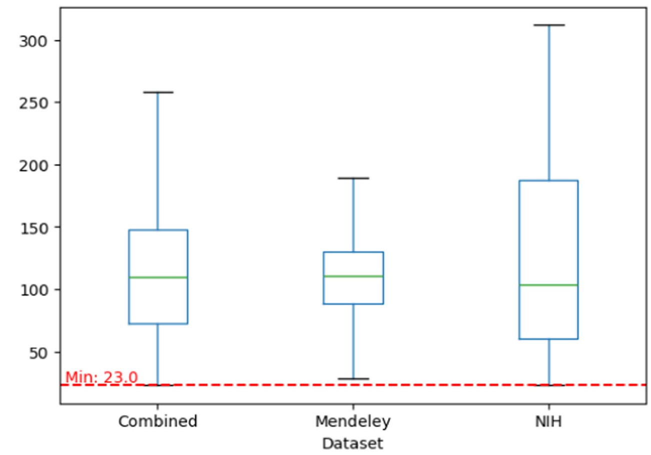

Fig. 3. Sharpness levels in lung X-ray datasets.

Algorithm 4: Diagnostic Sharpness Level Check.

```
Input: Synthetic Lung X-ray images
Output: Diagnostic Sharpness Level Check
procedure ESTIMATE SHARPNESS USING LAPLACIAN VARIANCE
    for each image do
        Estimate the sharpness level of the image using Laplacian variance
    end for
end procedure
procedure DETERMINING IMAGE VALIDITY
    for each image do
        Retrieve the estimated sharpness level for the image
        if noise level $< 23$ then
            Mark image as erroneous
        else
            Mark image as correct
        end if
    end for
end procedure
```

the image has more defined edges and, consequently, a higher level of detail.

In Fig. 3, we showcase the boxplots representing sharpness levels across our lung X-ray datasets, including the Mendeley [4] and NIH datasets [20], as well as their concatenation as "Combined". This visual representation underscores the sharpness characteristics of the datasets, integral to the quality and diagnostic viability of the images. A crucial finding from our analysis is the establishment of a minimum sharpness threshold of 23, derived from these datasets. Synthetic lung X-ray images falling below this threshold are considered inadequate for medical applications and are therefore excluded from further analysis or use. The threshold of 23 for sharpness acts as a decisive marker in evaluating the utility of synthetic images. Those falling below this marker lack the essential clarity and detail required for accurate medical interpretation.

To operationalize this criterion in a systematic and automated manner, we have developed Algorithm 4. This algorithm assesses the sharpness levels of synthetic lung X-ray images, categorizing them based on their adherence to the established sharpness threshold. Images that meet or exceed the threshold are deemed suitable for medical purposes, while those falling below it are flagged and excluded.

5) [Axis 3: Adversarial Evaluation]: This evaluation axis, called "Adversarial Evaluation" involves the use of a trained baseline classifier, in our case coming from the Preparatory Task 2, to discern between normal and pneumonia-afflicted lung conditions. The classifier, trained on real-world data, serves as a benchmark tool for assessing the predictive value of real against synthetic datasets.

The core objective of this adversarial evaluation is to analyze the extent to which synthetic data can mimic the predictive characteristics of real-world data when subjected to the same diagnostic algorithm. By inferencing the classifier with both real and synthetic data, we aim to perform a comparative analysis across various evaluation metrics. The comparison will include not only a general assessment across all data but also a more detailed analysis on a class-by-class basis. This approach allows us to delve deeper into how well synthetic data represent each category – normal and pneumonia.

A critical aspect of this evaluation is determining any disparity in the predictive value of synthetic data. For instance, if the synthetic data consistently underperform in accurately classifying pneumonia cases when compared to real data, this would indicate a lack of fidelity in the synthetic dataset's representation of pathological features critical for diagnosis. Conversely, comparable performance between synthetic and real data would suggest that the synthetic images possess a high degree of realism and diagnostic value.

IV. Experimental Outcomes

A. System Configuration

In our research setup, we chose the GPU V100 over a CPU due to its superior computational power. GPUs excel in handling parallel tasks, making them ideal for training complex models like cGANs and image classifiers, which involve complex computations. This parallel processing capability significantly accelerates training and analysis especially with intricate neural network architectures.

B. Dataset Description

In the context of this research, the Mendeley Chest X-ray dataset [4], derived from Guangzhou Women and Children's Medical Center, Guangzhou, plays a pivotal role. It encompasses 5856 anterior-posterior chest X-ray images of pediatric patients aged one to five years, as shown in Fig. 4. These images, integral for studying pulmonary conditions, were collected as part of the patients' routine clinical care. The dataset underwent thorough quality control, where scans of inferior quality were excluded and the remaining images were meticulously graded by two

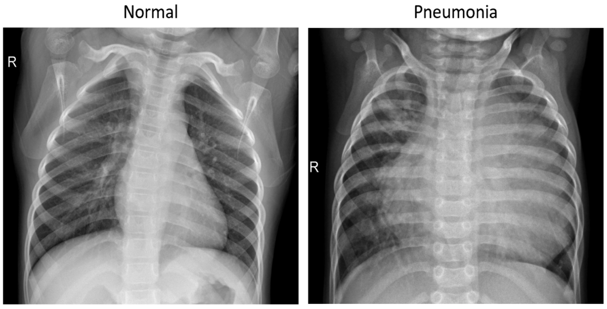

Fig. 4. Illustrative examples of chest X-rays.

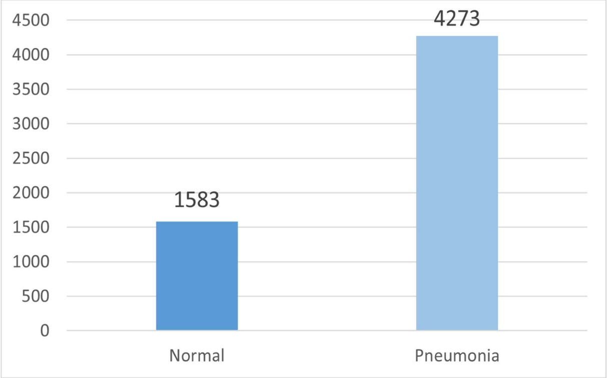

Fig. 5. Distribution of Chest X-ray Images by Class.

expert physicians, with a third providing additional review for the evaluation set.

For the purposes of this study, the two general classes were used: Normal and Pneumonia, as in Fig. 5, without distinguishing between the subcategories of bacterial and viral pneumonia. The dataset is comprised of 4273 images indicating pneumonia and 1583 images classified as normal.

Regarding data management, as shown in Fig. 1, our initial dataset was stratified and split into two distinct sets: “Real_main” and “Real_test,” with a 90/10 ratio. The “Real_main” dataset was utilized for data synthesis, resulting in the “Synthetic” dataset. Furthermore, “Real_main” underwent another stratified split into “Real_sub1” and “Real_sub2” with a 95/5 ratio. “Rel_sub1” was employed in the training of our classifier, while “Real_sub2” played a crucial role in the third axis of our core task, specifically for adversarial evaluation. Additionally the evaluation report for the classifier for the validation of the effectiveness of our classifier was generated by inferencing on the “Real_test” dataset, providing insights into the classifier’s performance metrics.

C. Evaluation Metrics

To assess the quality and efficacy of our development process we employed a diverse range of evaluation metrics. For the evaluation of the synthetic data quality during the preparatory task of data synthesis, we used the Structural Similarity Index Measure (SSIM) as in (10), a metric proposed by Wang et al. [23].

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{10}$$

In (10) μ_x and μ_y are the average values of images x and y, σ_x and σ_y are the variances, σ_{xy} is the covariance and c_1 with c_2 are the constants to stabilize the division.

For the evaluation of both the classifier during the second preparatory task of classifier development but also in the concept of adversarial evaluation in the 3 rd axis of framework, we used metrics based on the confusion matrix. A confusion matrix is a tabular representation of the performance of a classification algorithm. In our binary classification scenario for normal and pneumonia cases, the confusion matrix elements are True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). TP represents correctly identified pneumonia cases, TN denotes correctly identified normal cases, FP indicates normal cases incorrectly identified as pneumonia, and FN represents pneumonia cases incorrectly identified as normal.

From this confusion matrix, we derived critical metrics such as accuracy (11), recall (12), precision (13) and the F1 score (14). Accuracy measures the overall correctness of the model, recall (or sensitivity) assesses how well the model identifies positive cases, precision evaluates the correctness of positive predictions, and the F1 score provides a balance between precision and recall, especially important in uneven class distributions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{14}$$

D. Results

1) Preparatory Task 1 - Data Synthesis: In our study, we engaged in a systematic training regimen for the cGAN model, extending over 70 epochs, to produce the synthetic data that our SPINE framework will evaluate as part of the end-to-end data management. This training process involved regular monitoring of the model’s key components: the generator and the discriminator. Special attention was paid to their respective losses, a critical factor in ensuring optimal model performance and learning accuracy.

These observations are depicted in Fig. 6, offering insight into the dynamic interplay between the generator and discriminator during the training process.

A pivotal aspect of our training methodology was the use of a latent dimension of 200 in the cGAN model. The latent dimension in generative models like cGANs refers to the size of the input noise vector. Additionally, we adopted a batch size of 8 for the training process. Furthermore, a crucial aspect of our training process involved the periodic plotting of the generated

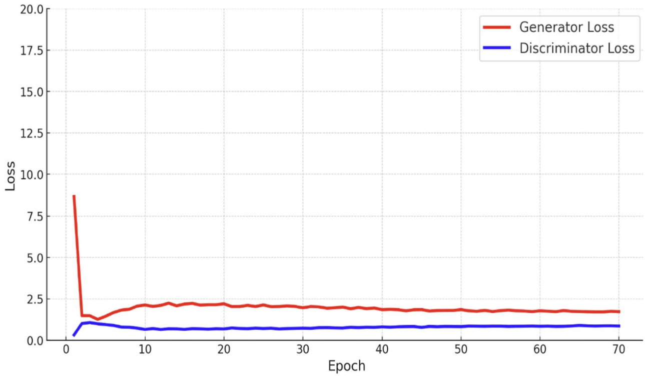

Fig. 6. Generator and discriminator losses.

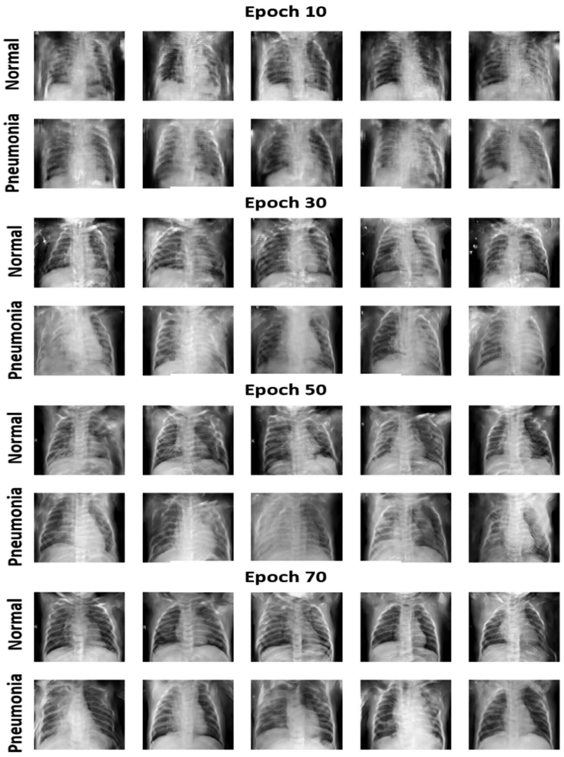

Fig. 7. Generated images.

synthetic images. By examining these images at regular intervals, we could qualitatively assess the evolution and refinement of the model's image generation capabilities, as shown in Fig. 7.

The synthetic data generated underwent an evaluation using SSIM comparing "Real_main" and "Synthetic" datasets of Fig. 1. The results of this evaluation were quite revealing, as we achieved an Average SSIM score of 0.55%. For the 'Normal' class, we achieved an Average SSIM of 0.48%. On the other hand, the 'Pneumonia' class exhibited a higher level of fidelity, as evidenced by an Average SSIM of 0.58%.

2) Preparatory Task 2 - Classifier Development: In our study, as part of the end-to-end model management, we created a baseline classifier which is crucial for the adversarial evaluation axis of our SPINE methodology. This classifier would be capable of categorize the X-ray lung images to normal/pneumonia. For this reason, we resized all chest X-ray images to 256 × 256 pixels as a preprocessing step. We trained our model for 25 epochs, with a batch size of 4, selecting the best-performing iteration based on the lowest validation loss, as shown in Fig. 8.

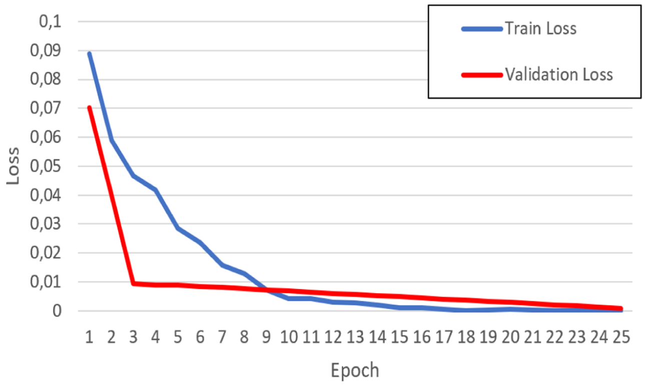

Fig. 8. Train and validation losses - YOLOv8.

TABLE I
CLASSIFICATION REPORT OF YOLOV8

Class	Total Metrics	Precision	Recall	F1	Support
Normal		0.99	0.43	0.60	234
Pneumonia		0.74	1.00	0.85	390
	Macro Avg	0.87	0.71	0.72	624
	Weighted Avg	0.84	0.78	0.76	624
Total Accuracy: 0.78					

To have an indication about the performance of the classifier, our baseline model was evaluated on the "Real_test" set (Fig. 1) producing the classification report of Table I. The evaluation revealed a high precision (0.99) for "Normal" class but a lower recall (0.43), indicating accurate but not comprehensive detection of normal cases. In contrast, it showed strong performance in identifying "Pneumonia" cases, with high precision (0.74) and perfect recall (1.00). The overall accuracy stood at 78%, with macro and weighted F1-scores of 0.72 and 0.76, respectively, suggesting a balanced but improvable performance across both classes.

3) [Axis 1: Experts Knowledge] Criterion 1.1 - Thoracic Field Completeness: In Fig. 9, we present two indicative cases following the application of Algorithm 1 applying red mask for background and blue for lung areas. This figure illustrates the outcomes of our thoracic field completeness check, showcasing one example of a correctly processed synthetic chest X-ray image and another example deemed erroneous.

To validate the effectiveness of Algorithm 1, we applied it to the "Real_sub2" dataset of Fig. 1. All images in this dataset successfully passed the check, with each being labeled as 'correct'.

4) [Axis 1: Experts Knowledge] Criterion 1.2 - Diaphragm Existence: In Fig. 10, two representative examples showcase the outcomes post-application of Algorithm 2, where the area subject to analysis is demarcated with a red line.

This process involves computing the ratio of bright to dark pixels within the specified region. Furthermore, a validation

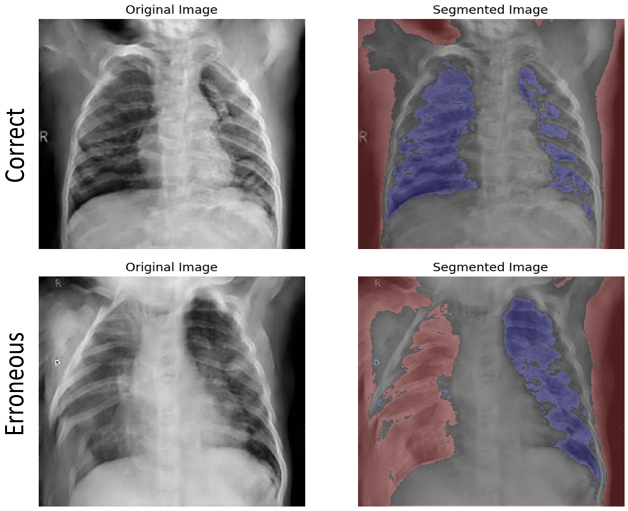

Fig. 9. Thoracic field completeness check.

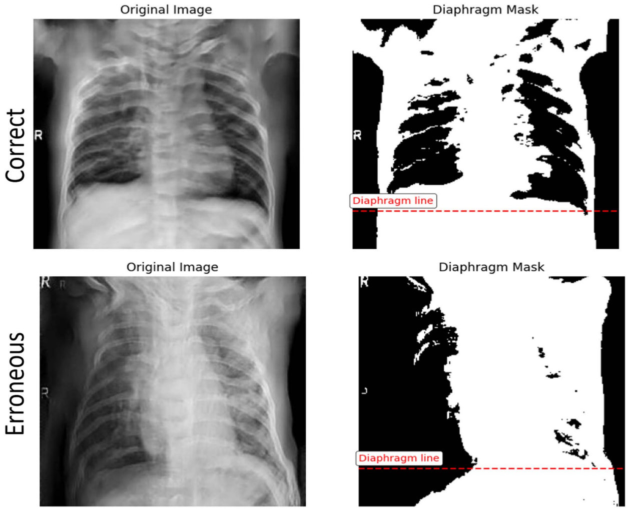

Fig. 10. Diaphragm existence check.

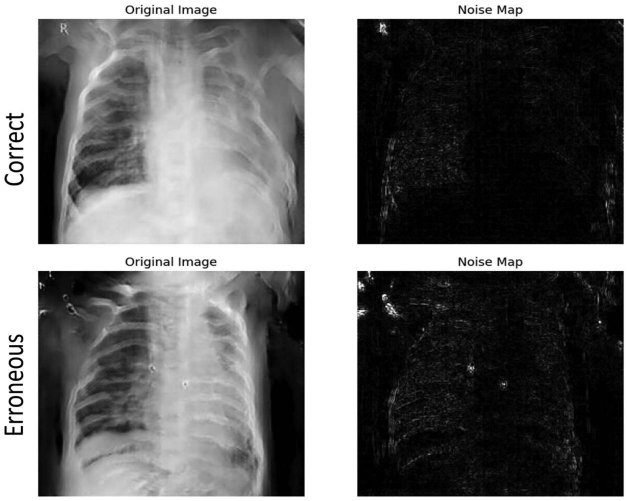

Fig. 11. Diagnostic noise level check.

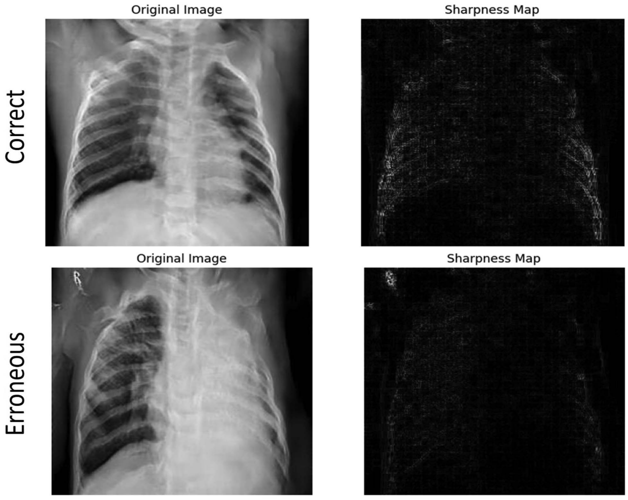

Fig. 12. Diagnostic sharpness level check.

of Algorithm 2 was conducted on the “Real_sub2” dataset, as depicted in Fig. 1. This validation exercise confirmed the algorithm’s robustness, with all images in the dataset being classified as ‘correct’.

5) [Axis 2: Data Statistics] Criterion 2.1 - Diagnostic Noise Level Check: In Fig. 11, we present side-by-side examples of a ‘correct’ and an ‘erroneous’ synthetic lung X-ray image from our dataset, following the application of Algorithm 3, accompanied by their respective noise maps. The ‘erroneous’ image exhibits a marked reduction in clarity with a wider blurry area. This lack of detail is reflected in its noise map, which shows a greater abundance of white spots, signifying higher noise levels.

6) [Axis 2: Data Statistics] Criterion 2.2 - Diagnostic Sharpness Level Check: The comparison of a ‘correct’ and an ‘erroneous’ synthetic lung X-ray image based on Algorithm 4 could be seen in Fig. 12.

Fig. 12 images are accompanied with their respective sharpness maps. The sharpness map of the ‘correct’ image reveals a high level of detail, with clearly defined edges and structures. On the other hand, the ‘erroneous’ image, marked by its low sharpness, displays a significant loss of detail, evidenced by blurred edges and a lack of clarity in its corresponding sharpness map.

7) [Axis 3: Adversarial Evaluation]: Regarding the adversarial evaluation of the synthetic chest X-ray images, Table II is presenting the comparative analysis of “Real_sub2”(Real in Table II) and “Synthetic”(Synthetic in Table II) datasets, as per Fig. 1, through the trained classifier of Preparatory Task 2. The evaluation focused on recall, precision, F1 score and overall accuracy metrics for both normal and pneumonia conditions in each dataset.

The recall rates for Real data were high, with 0.93 for normal and 0.96 for pneumonia conditions, indicating the classifier’s proficient capability in correctly identifying true positives. In

TABLE II
COMPARATIVE EVALUATION OF CLASSIFICATION METRICS FOR REAL AND SYNTHETIC DATA

Metric	Class	Real Data	Synthetic Data
Recall	Normal	0.93	0.84
	Pneumonia	0.96	0.93
Precision	Normal	0.89	0.81
	Pneumonia	0.97	0.94
F1-score	Normal	0.91	0.82
	Pneumonia	0.97	0.94
Accuracy		0.95	0.91

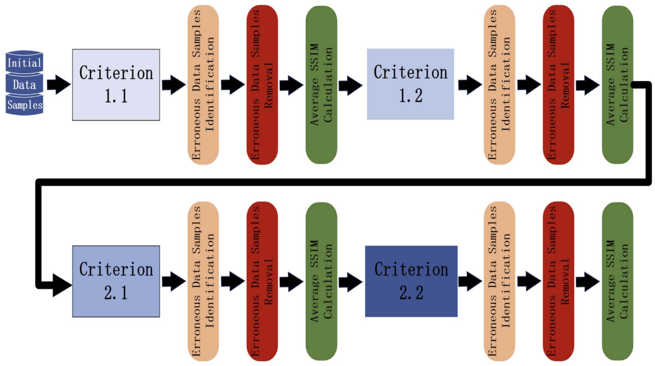
Initial Data Samples
Criterion 1.1
Erroneous Data Samples Identification
Erroneous Data Samples Removal
Average SSIM Calculation
Criterion 1.2
Erroneous Data Samples Identification
Erroneous Data Samples Removal
Average SSIM Calculation
Criterion 2.1
Erroneous Data Samples Identification
Erroneous Data Samples Removal
Average SSIM Calculation
Criterion 2.2
Erroneous Data Samples Identification
Erroneous Data Samples Removal
Average SSIM Calculation

Fig. 13. Workflow diagram of ablation study.

contrast, Synthetic data showed lower recall rates of 0.84 for normal and 0.93 for pneumonia conditions. This discrepancy suggests that while Synthetic data can mimic real scenarios to a certain extent, it is less reliable, especially in identifying TP in normal conditions. Precision values further accentuated this trend. Real data exhibited high precision values of 0.89 for normal and 0.97 for pneumonia conditions. Synthetic data, however, lagged with precision values of 0.81 for normal and 0.94 for pneumonia conditions, pointing to a higher incidence of FP, particularly in the normal category. F1 score, further confirm the above outcomes. Finally, overall accuracy metrics highlighted also the differences as Real data achieved an overall accuracy of 0.95, compared to the 0.91 accuracy of synthetic data. This difference underscores that Synthetic data, while accurate, doesn't match the performance level of real-world data. From these observations, it's evident that Synthetic data does not completely replicate the predictive value of real-world data, with its limitations being more pronounced in the normal condition category.

V. Ablation Study

In the ablation study conducted to evaluate synthetic medical images, a comprehensive approach was taken to understand the impact of the four criteria on the quality of the dataset, as shown in Fig. 13. The primary objective was to refine the synthetic dataset ("Synthetic" in Fig. 1) by sequentially applying these criteria and observing changes in the SSIM against the real-world dataset("Real_sub2" in Fig. 1). This process demonstrated the effectiveness of each criterion in enhancing the dataset's accuracy.

TABLE III
IMPACT OF SEQUENTIAL CRITERIA APPLICATION ON SSIM IMPROVEMENT IN SYNTHETIC MEDICAL IMAGES

Criterion	Initial Data Samples	Erroneous Data Samples	Remaining Data Samples	Average SSIM
	261		261	0.5536
1.1	261	76	185	0.5652
1.2	185	20	165	0.5739
2.1	165	4	161	0.5746
2.2	161	3	158	0.5758

Table III, encapsulates this sequential criterion application presenting at each row the synthetic data samples before the application of criterion ("Initial Data Samples" in Table III), the number of synthetic data samples considered erroneous after the application of the criterion ("Erroneous Data Samples" in Table III), the correct images remained after the application of the criterion ("Remaining Data Samples" in Table III) and finally the SSIM score of the remaining synthetic data samples against the real-world data.

The initial dataset comprised 261 synthetic images. The baseline SSIM between this complete synthetic dataset and the real-world dataset ("Synthetic" and "Real_sub2" in Fig. 1) was calculated as 0.5536, indicated also in results of Preparatory Task 1.

Criterion 1.1("Thoracic Field Completeness"), was then applied, leading to the identification and exclusion of 76 erroneous data samples. This exclusion reduced the dataset to 185 images and it resulted in an improvement in the average SSIM score to 0.5652. Following this, Criterion 1.2 ("Diaphragm Existence") was implemented. This led to the further exclusion of 20 images, bringing the dataset down to 165 images. The application of this criterion brought another increase in the SSIM score, rising to 0.5739. The study proceeded with the application of the Criterion 2.1("Diagnostic Noise Level Check"), which identified and excluded 4 more images, leaving 161 images in the dataset. Post this exclusion, a slight increase in the SSIM score was observed, reaching 0.5746. Finally, the Criterion 2.2("Diagnostic Sharpness Level Check") was applied. This last criterion resulted in the exclusion of 3 additional images, culminating in a dataset of 158 images. The final SSIM score after this round of refinement was 0.5758, the highest throughout the study.

The sequential application of these criteria and the corresponding increase in the SSIM score with each step effectively demonstrated that the removal of specific erroneous images led to a gradual but consistent enhancement in the resemblance of the synthetic dataset to the real-world dataset. This progression validated the utility of the SPINE framework in improving the quality of synthetic medical images, making the dataset more representative and accurate for potential applications in medical imaging and analysis.

VI. Conclusion

This paper introduced the SPINE framework, an innovative and comprehensive system crafted to evaluate synthetic pulmonary imaging. Our approach marks a substantial leap forward in utilizing synthetic data for medical diagnostics and machine learning, tackling key concerns related to data quality and applicability. To ensure thoroughness and precision at every stage of this evaluation, we went beyond merely leveraging an open-source synthetic dataset. We systematically developed all the necessary components as preparatory steps, encompassing the creation of synthetic data and the training of a classifier that is integral to the SPINE evaluation axis. This post-market approach followed at the SPINE framework, combining domain expertise, statistical data analysis and adversarial evaluations, allowed us to critically examine the utility and reliability of synthetic images compared to real-world data. Furthermore, the ablation study, a critical component of our research, demonstrated the effectiveness of the SPINE framework in refining the quality of synthetic medical images. By systematically applying the Algorithms developed under our 4 main Criteria and tracking improvements in SSIM scores, we provided concrete evidence of our methodology's ability to enhance the fidelity of synthetic datasets to real-world scenarios.

In conclusion, this research not only contributes a novel framework for evaluating synthetic medical images but also lays the groundwork for future advancements in this rapidly evolving field. By continually refining and expanding the SPINE framework, we can further bridge the gap between synthetic data and real-world applicability, enhancing the reliability and effectiveness of synthetic data in medical imaging and machine learning.

Acknowledgment

Conflict of Interest: The authors declare that there is no conflict of interest.

Author Contributions: N. Ntampakis carried out the experiment and wrote the manuscript. V. Argyriou and K. Diamantaras helped supervise the project. All authors provided critical feedback and helped shape the manuscript.

References

[1] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: Methodological failures and recommendations for the future," *Digit. Med.*, vol. 5, 2022, Art. no. 48.

[2] R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. K. Williamson, and F. Mahmood, "Synthetic data in machine learning for medicine and healthcare," *Nature Biomed. Eng.*, vol. 5, pp. 493–497, 2021.

[3] X. Xing, Y. Nan, F. Felder, S. Walsh, and G. Yang, "The beauty or the beast: Which aspect of synthetic medical images deserves our focus?," *Alzheimer's Dement.*, 2023.

[4] D. Kermany, K. Zhang, and M. Goldbaum, "Labeled optical coherence tomography (OCT) and chest X-ray images for classification," *Mendeley Data*, vol. 2, no. 2, 2018, Art. no. 651.

[5] A. Goncalves, P. Ray, B. Soper, J. Stevens, L. Coyle, and A. P. Sales, "Generation and evaluation of synthetic patient data," *BMC Med. Res. Methodol.*, vol 20, no. 108, pp. 1–40, 2020.

[6] C. Yan et al., "A Multifaceted benchmarking of synthetic electronic health record generation models," *Nature Commun.*, vol. 13, no. 1, 2022, Art. no. 7609.

[7] K. E. Emam, L. Mosquera, X. Fang, and A. El-Hussuna, "Utility metrics for evaluating synthetic health data generation methods: Validation study," *JMIR Med. Inform.*, vol. 10, no. 4, 2022, Art. no. e35734.

[8] L. Lan et al., "Generative adversarial networks and its applications in biomedical informatics," *Public Health*, vol. 8, 2020, Art. no. 164.

[9] I. J. Goodfellow et al., "Generative adversarial networks," 2014, *arXiv:1406.2661 [stat.ML]*.

[10] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[11] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3431–3440.

[12] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. 30th Int. Conf. Mach. Learn.*, vol. 28, 2013, p. 3.

[13] D. P. Kingma and J. Ba, "Adam: A methodforstochastic optimization," 2014, *arXiv:1412.6980*.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[15] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," 2014, *arXiv:1405.0312 [cs.CV]*.

[16] D. D. F. Gumieri and I. d. S. Marques, "Evaluation of chest X-ray quality parameters," *Int. J. Radiol. Imag. Techn.*, vol. 7, 2021, Art. no. 082.

[17] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

[18] R. Smithuis and O. v. Delden, "AChest X-ray - basic interpretation,"*Radiol. Dept. Alrijne Hospital, Leiderdorp Academical Medical Centre*, Amsterdam, The Netherlands, Feb. 2013. [Online]. Available: https://radiologyassistant.nl/chest/chest-x-ray/basic-interpretation

[19] T. Suwatanapongched, D. S. Gierada, R. M. Slone, T. K. Pilgram, and G. P. Tuteur, "Variation in diaphragm position and shape in adults with normal pulmonary function," *Chest.*, vol. 123, no. 6, Jun. 2003, pp. 2019–27, doi: 10.1378/chest.123.6.2019.

[20] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," *IEEE Conf. Comput. Vision Pattern Recognition (CVPR)* , Honolulu, HI, USA, 2017, pp. 3462–3471, doi: 10.1109/CVPR.2017.369.

[21] J. Immerkær, "Fast noise variance estimation," *Comput. Vis. Image Understanding*, vol. 64, no. 2, pp. 300–302, 1996.

[22] J. Cleverley, J. Piper, and M. M. Jones, "The role of chest radiography in confirming COVID-19 pneumonia," *BMJ.*, vol. 370, Jul. 2020, Art. no. m2426, doi: 10.1136/bmj.m2426.

[23] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.