International Journal of Reliability, Risk and Safety: Theory and Application

Online ISSN: 2676-3346

Vol. 7/ Issue 2/ Year 2024/ pp. 15-27
DOI: 10.22034/IJRRS.2024.7.2.2
Received: 06 June 2024, Revised: 09 September 2024, Accepted: 11 September 2024

Available online at: https://www.ijrrs.com

Original Research Article

Evaluation and Risk Management Strategies for Developing AI-based Medical Image Products

Shahrzad Oveisi [1*], Marjan Goudarzi [1] Mohammad Shahram Moein [1]

1. Innovation and Development of Artificial Intelligence Center, ICT Research Institute, Tehran, Iran

*** shahrzad.oveisi@gmail.com**

Abstract

AI-based products, particularly in medical diagnosis, have become increasingly popular. However, with the rise of AI technologies, there is a critical need for quality assurance and risk assessment to ensure the reliability and impartiality of these systems. One crucial application of AI in the medical field is the diagnosis of diseases through imaging techniques, such as chest X-rays. Chest X-rays are commonly used by physicians to diagnose respiratory diseases quickly and cost-effectively. Yet, interpreting chest X-rays can be challenging, and errors in diagnosis can have severe consequences, especially for life-threatening conditions like pneumonia. Given the high mortality rate associated with pneumonia, accurate and timely diagnosis is essential. It is vital to prioritize quality assurance and risk assessment in the development and implementation of AI-based products, particularly in critical areas like medical diagnosis.

In this paper, we utilized a deep CNN network to diagnose pneumonia from chest X-ray images. We also introduced two criteria, bias and transparency, to evaluate these products. For assessing these criteria, we provided methods based on checklists and quantitative assessment approaches for our data. We successfully implemented these solutions on our data and even achieved a robust model by applying data augmentation techniques, raising accuracy above 90 percent. Additionally, to validate our data, we used two tests: the pressure test and the crystal test, which yielded an accuracy of over 70 percent. We also completed all the checklist-based methods and were able to obtain validation for these data in medical products.

Keywords: Artificial Intelligence; Diagnostic Imaging; Deep Learning; Risk Management.

Nomenclature

CNN	Convolutional Neural Networks
AI	Artificial Intelligence
DNN	Deep Neural Networks
COPD	Chronic Obstructive Pulmonary Disease
DL	Deep Learning Models

1. Introduction and Background Research

Artificial intelligence (AI) with its impressive capabilities has led to numerous innovations in almost every aspect of our society and economy, from business and health to transportation and cybersecurity [1]. There have been several arguments regarding the failures of AI. According to a 2019 IDC survey, "most organizations have reported AI failures in their projects, with a quarter having a failure rate of up to 50% [2]. These failures have created a strong and tolerable record of appropriate software testing, and industry surveys indicate that AI is one of the most important trends in software testing." Therefore, due to the significance of these systems and the high costs associated with their failure, the requirements for the production of artificial intelligence products have been formulated in the form of standards. For example, the production of standard AI-based systems has resulted in a net profit of 17,000,000€ for Germany [3,4]. AI is assessed using checklists based on standards and criteria; a comprehensive and detailed review of AI systems by checklists prevents the occurrence of accidents or possible negative consequences of artificial intelligence systems. According to experts, transparency and bias are the most important and prioritized criteria [5,6].

As mentioned, AI-based systems are of high importance in the field of health and medical devices. In developing regions where billions of people suffer from energy poverty and rely on polluting energy sources, there is a high risk of contracting pneumonia, especially in terms of lung injury diagnosis [7,8]. The World

How to cite this article:
S. Oveisi, M. Goodarzi, and M.S. Moein, "Evaluation and risk management strategies for developing AI-based medical image products," *International Journal of Reliability, Risk and Safety: Theory and Application*, vol. 7, no. 2, pp. 15-27, 2024, doi: 10.22034/IJRRS.2024.7.2.2.

Health Organization estimates that more than 4,000,000 premature deaths occur annually due to diseases related to indoor air pollution, including pneumonia. Each year, over 150,000,000 people, especially children under 5 years of age, are faced with pneumonia. In regions such as Africa, a shortage of medical resources and medical personnel can exacerbate the problem. For these populations, an accurate and rapid diagnosis is of high importance because it enables their access to timely treatment, thereby reducing the time and cost so valuable for those living in poverty [9,10]. In recent times, deep learning algorithms inspired by convolutional neural networks have become the standard choice for medical image classification; however, these techniques require high processing power. As an alternative, our study proposes a simple but optimal network model for solving the pneumonia classification problem [6]. Convolutional neural networks (CNN) have an advantage over deep neural networks (DNN) from the standpoint of image processing, which is equivalent to that of humans.

CNNs have an optimal structure for controlling 2D and 3D images and shapes and can extract abstract 2D features through learning. The max-pooling layer of CNN is effective in assimilating variable shapes and includes sparse connections along with common weights. Compared to fully connected networks of equivalent size, convolutional neural networks have significantly fewer parameters. Most importantly, gradient-based learning algorithms are used in CNN training and are less prone to the gradient reduction problem because the gradient-based algorithm is responsible for training the entire network to directly reduce a common weights map.

In this paper, after the introduction and research background, as shown in Figure 1, a workflow of the article's structure is presented. Accordingly, in the research fundamentals section, we address the basics of research and image data for the evaluation and significance of testing and assessment. Subsequently, in the methods section, we provide various methods and strategies for evaluating these products and data. In the third section, all the proposed strategies are tested and evaluated on our dataset.

2. Research Fundamentals

In this section, research fundamentals are discussed (Figure 1).

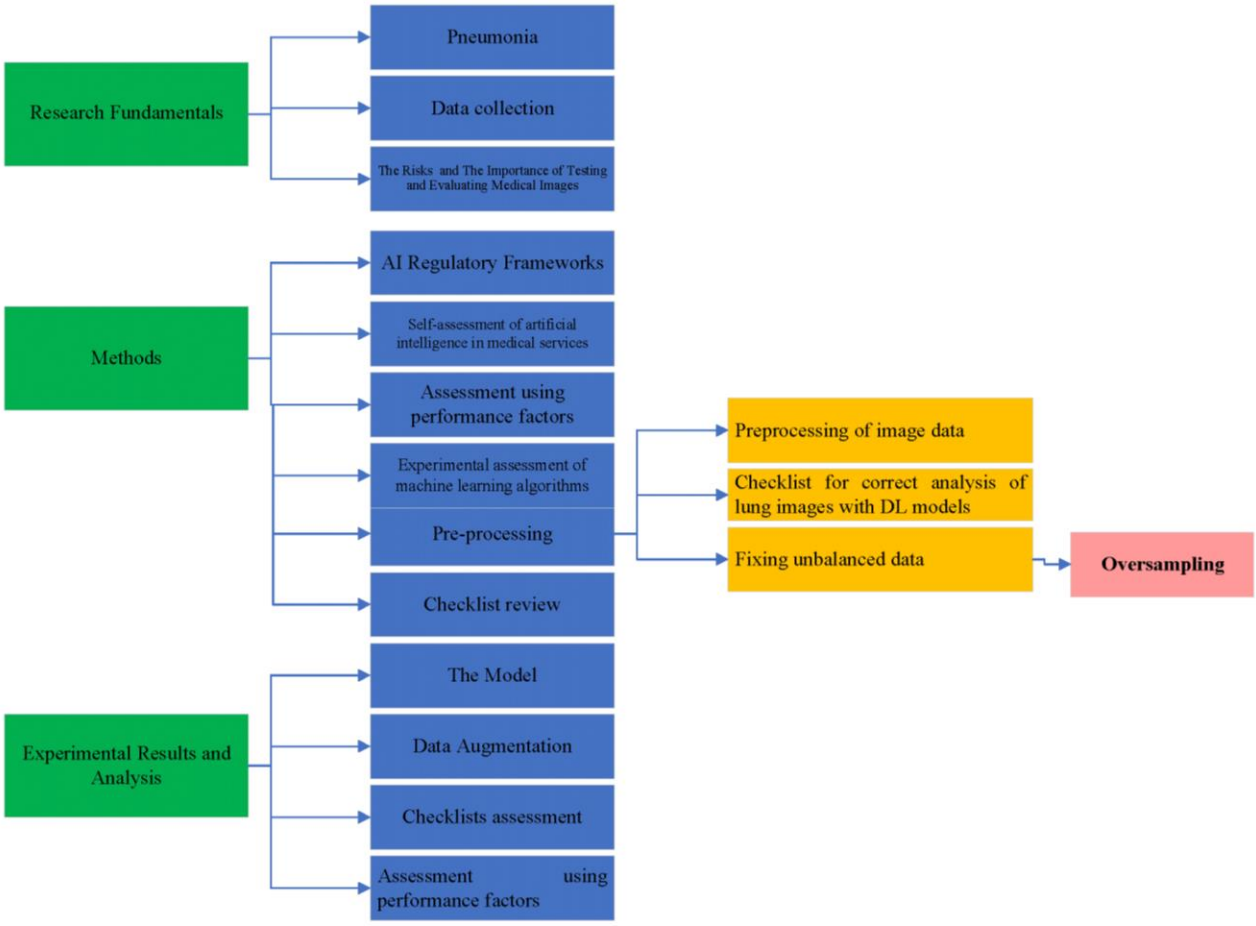

Figure 1. Workflow of paper

2.1 Pneumonia

Pneumonia is an inflammatory state of the lungs that impacts the alveoli, which are tiny air sacs. It commonly presents symptoms such as a dry cough, chest pain, fever, and breathing difficulties. The severity of the condition varies among individuals. Pneumonia is primarily caused by viral or bacterial infections, although it can also be triggered by other microorganisms, certain medications, or autoimmune disorders.

Risk factors include cystic fibrosis, chronic obstructive pulmonary disease (COPD), asthma, diabetes, heart failure, a history of smoking, decreased cough reflex after a stroke, and a weakened immune system (Figure 2).

Diagnosis is typically established through symptom assessment and physical examination. Additional diagnostic methods may include chest X-rays, blood tests, and sputum cultures to confirm the diagnosis [11,12].

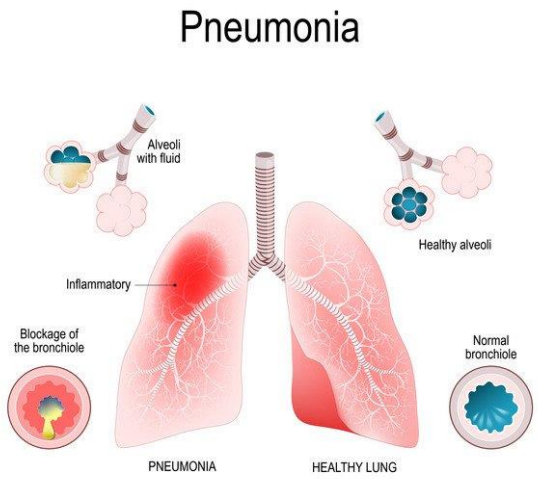

Figure 2. Pneumonia

2.2 Data Collection

The main dataset includes three main folders (i.e., training, test and validation) and two subfolders containing chest X-ray images with alternative (P) and normal (N) designations. All 5,856 chest X-ray images from a previous chest were carefully selected. The images were taken from 1–5-year-old children as part of the daily medical care of patients. The main classification of data was changed to balance the proportion of data for training and validation sets. We transformed the entire dataset into training and validation sets. To improve the validation accuracy, a total of 3,722 and 2,134 images were assigned to training and validation sets, respectively. It should be noted that this dataset was prepared and compiled from the validated Kaggle[1] website.

2.3 The Risks of AI In Health Care and The Importance of Testing and Evaluating Medical Images

In a paper published more than 50 years ago, William B. Schwartz stated: "Computer science is likely to exert its main effects by enhancing the mental matters of physicians and, in some cases, even by massively replacing them" [7]. Despite the effective role of artificial intelligence in the diagnosis of patients, experts have expressed concern about the potential adverse consequences of AI in medicine, which we will examine in detail in this chapter.

2.3.1 Risk of AI Bias in Medicine and The Persistence of Inequalities

A Canadian study in 2020 (Figure 3) investigated the validity of state-of-the-art deep learning algorithms for detecting abnormalities such as fractures, pulmonary lesions, nodules, pneumonia and so forth in chest radiographic images [8]. This study showed that the highest rate of misdiagnosis occurs in young women (aged 0-20 years), blacks, as well as low-income patients covered by public health insurance. In addition, the highest rate of non-diagnosis was observed in patients with intersectional identities (for example, a low-income Hispanic patient with health insurance). The authors concluded that "models trained on large data sets are usually not equivalent and lead to potential care inequalities if not corrected before application."

The most common cause of inequality in AI-based medicine is the bias of data that is used to train machine learning models, which has been hotly debated. Marzieh Ghassemi from the University of Toronto states in her recent presentation on AI in healthcare [9]: “Currently, bias is part of the clinical landscape; therefore, it is not the case that machine learning seeks to harm us. Indeed, when we train on data that was generated, labelled, and interpreted by humans, we may find some of the biases that humans have injected into that data."

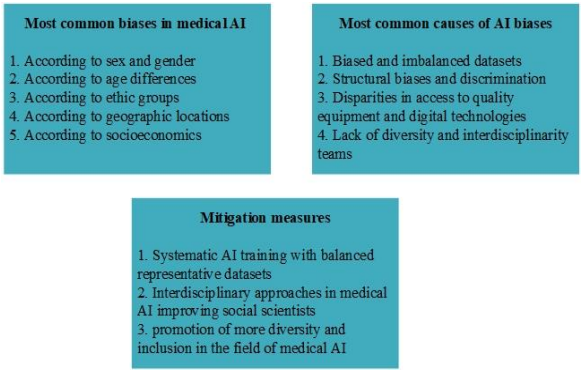

Figure 3. The most common AI biases in medicine, their causes and potential mitigation measures to develop fairer AI algorithms

1. Data Availability Statement: The dataset used in this study is openly available on Kaggle and is titled "Chest X-Ray Images (Pneumonia)" (https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia) This dataset has been available since June 1, 2018.

2.3.2 Lack of Transparency

Lack of transparency is a critical issue in the development and application of current AI tools in healthcare (Figure 4). This trend appears to undermine public trust in artificial intelligence, especially in sensitive fields such as medicine and healthcare, which focus on the health and well-being of citizens. Nevertheless, mistrust affects the acceptability of emerging AI algorithms among patients, physicians, and healthcare systems. Transparency of AI has a close relationship with its traceability and explainability, and transparency is necessary at both distinct levels:

1. Transparency of AI development and application processes (traceability)
2. Transparency of artificial intelligence decisions (explainability).

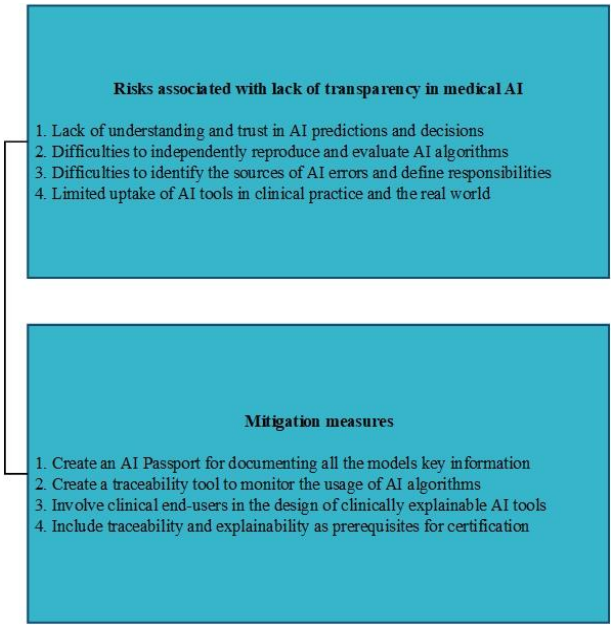

Figure 4. Main risks related to lack of transparency in current AI algorithms and possible mitigation measures

The reliability of artificial intelligence demands traceability, which involves maintaining a complete and transparent record of the AI development process. This includes tracking how the artificial intelligence model operates under real-world conditions. Specifically, achieving traceability requires:

Understanding Model Details: Knowledge of the model's specifics, such as its intended application, algorithm type (e.g., neural network), hyperparameters, and any pre- and post-processing steps. 2. Data Recognition: Awareness of data used for training and validation. This encompasses the data collection process, integration, learning protocols, and labelling. 3.AI Tool Monitoring: Regular evaluation of performance metrics, identification of defects, and periodic assessments.

In the healthcare domain, there are only a few AI tools with comprehensive tracking capabilities. Companies often hesitate to fully disclose their algorithms, presenting them as obscure tools that independent parties find challenging to comprehend and review. Consequently, their reliability and practical acceptance diminish in the real world.

There are several ways to improve the transparency of AI technologies used in healthcare. First, there should be an "AI passport" for each algorithm to record all the key information of that model. There should also be tools to track and monitor the use of AI algorithms to record possible errors and performance drops as well as implementing periodic inspections. Artificial intelligence developers should involve the end users of clinical services in the development process from the beginning to choose the best justification approach for each application, ensure that these justifications are effective and acceptable at the bedside and thereby improve the justification of AI algorithms. Finally, regulatory bodies can play an essential role by mandating traceability and justifiability of AI tools as prerequisites for certification.

3. Methods

In this section, we examine and evaluate risk assessment methods [13,14,15].

3.1 AI Regulatory Frameworks

AI risks can be described and classified based on the severity of damage they cause, as well as the probability and frequency of the damage. In medical services, the risks of artificial intelligence are significantly different, from rare or small risks with limited and manageable damage to patients and citizens to frequent and severe risks inflicting loss or irreversible damage.

Therefore, to minimize the risks of AI and maximize its benefits in future healthcare services, it is of high importance to identify, analyze, understand and monitor potential risks on a case-by-case basis for each application and new AI algorithm. In this proposed plan, specific requirements and obligations for proper risk management of high-risk AI are proposed, which are presented in Table 1.

Table 1. Requirements and obligations of high-risk artificial intelligence tools according to the plan

• Using quality data for training, validation and testing (relevant data and representative data).
• Preparation of technical documentation and adjustment of system logging capabilities (traceability and auditability).
• Ensuring the appropriate level of transparency and providing information about the capabilities and limitations of the system and its usage mode to the users.
• Ensuring human supervision (measures available inside the system or actions that must be implemented by users).
• Ensuring strength, precision and cybersecurity

3.2 Self-assessment of Artificial Intelligence in Medical Services

Self-Assessment of AI in Health Services was published in 2021 by a multidisciplinary team of Australian

researchers. The purpose of this inventory was to help clinicians assess the readiness of algorithms to be used in daily care and to identify areas that require further development and fine-tuning before system deployment. This list was compiled based on several narrative review articles on AI in health care, which were summarized around ten general questions as assessment questions that are presented in Table 2.

Table 2. Questions of medical AI assessment list

• What is the purpose and context of the algorithm? • How appropriate was the data used to train the algorithm? • Was there enough data to train the algorithm? • How well does the algorithm perform? • Can the algorithm be transferred to a new clinical setting? • Are the outputs of the algorithm clinically understandable? • How does the algorithm fit with the current work routine and how does it complement it? • Has the use of the algorithm been found to improve patient outcomes and patient care? • Can the algorithm harm the patient? • Does the use of the algorithm cause ethical, legal or social concerns?

3.3 Assessment using performance factors

The assessment of AI algorithms in the medical field is increasingly recognized as needing to encompass factors beyond traditional measures like model accuracy. **Table 3** presents examples of performance factors for AI-based diagnostic algorithms in radiology, highlighting the need for a comprehensive approach to assessment. These factors include: 1- Reliably 2- Applicability 3- Transparency 4-Monitorability 5-Usability

Table 3. Examples of performance factors for AI imaging algorithms

Precision	The algorithm must be able to accurately perform all the diagnostic tasks for which it was designed
Reliability	The algorithm must maintain its accuracy even when exposed to the kinds of changes that might reasonably occur in a clinical situation, such as fluctuations in image quality.
Executability	The accuracy of the algorithm must be kept across all different models, as well as models of imaging modalities and patient populations for which it is designed.
Certainty	When the algorithm is applied to the same image multiple times and in different settings, it should provide the same result each time.
Irreversibility	It is necessary that the algorithm extract the most important information from the image without being influenced by irrelevant or redundant aspects of the image data.
Awareness of limitations	The algorithm must have the means to recognize when it is within or beyond its capabilities, either due to inherent limitations of the model, its clinical applicability, or limitations imposed by clinical variations such as unexpected patient anatomy or image quality.
Lack of failure	The algorithm must be able to recognize when an incorrect result has been reached and have the ability to ensure that any errors are detected and prevented before they are published in a clinical setting.
Clear logic	The user interface should allow the operator to clearly understand the relationship between input and output, including what data was analyzed, what options were considered, and why certain possibilities were excluded so that the algorithm can correctly accept or reject conclusions in each given case.
Transparent rate of confidence	This algorithm should inform the user of the confidence level of its assessment for each case. In addition to verifying the confidence of the model itself, it is also important to check the accuracy of the confidence of the model.
Controllability	The algorithm should communicate its performance data to the user to enable continuous monitoring of individual as well as aggregate samples and quickly highlight any significant performance discrepancies.
Editability	A method should be provided to monitor the continuous performance of the algorithm to guide the appropriate intervention. This method can include regular quality control checks similar to those performed by operators on imaging equipment.
Intuitive user interface	The user interface should allow the operator to directly understand how to use the algorithm with the least possible training and should impose the least possible cognitive pressure on the user.

Among the works that have directly examined the fairness of AI in the medical field, a recent study assessed advanced deep neural networks on a public chest X-ray dataset concerning the patient's gender, age, race, and type of insurance, the latter acting as a proxy for socioeconomic status. According to the research findings, "Models trained on large datasets do not naturally provide equal opportunity and instead, if applied without modification, potentially lead to disparities in care".

3.4 Experimental assessment of machine learning algorithms

In this section, we will examine the experimental assessment methods of our algorithm.

3.4.1 Preprocessing of image data

Due to the large volume of data, it is necessary to automate the preprocessing process. Preprocessing should not alter the image to remove or add relevant data. The purpose of preprocessing is to make it impossible to identify calibration components of the machine or machine characteristics such as radiation dose. In general, preprocessing is an important stage that precedes preparation. Preprocessing should reduce the possibility of bias and ensure the presentation of more homogeneous images in which no medically important components are removed.

3.4.2 Checklist for correct analysis of lung images with DL models

Based on the information provided, it seems that the development of a model for analyzing lung images is indeed a complex process. A checklist has been prepared based on analyzed studies and the errors found in them [Table 4]. This checklist is aimed at significantly improving the quality of the modelling process and helping to prevent or quickly identify and fix errors.

Table 4. Lung image analysis checklists

Image Preprocessing	Data Source
In the below list, the letter R indicates that the point should be consulted with a field specialist/radiologist, and the letter D shows that the point should be consulted with a model developer.	
D Is data preprocessing described? D Are artifacts (such as subtitles) removed? Increase data (if needed) D Are the lungs completely present after transformations? R Are lung structures visible after brightness or contrast conversion? D Are only reasonable conversions applied?	D Do the data and related information provide sufficient diagnostic quality? If the images are in DICOM, does the header provide the required information? If not, is it presented in another way? R Are low-quality images (i.e., blurry, too dark, or too bright) rejected? D Is the data set gender and age balanced? R Does the dataset contain one type of image (CT or X-ray)? R Are lung structures (lung window) visible on CT images? D Are images of children and adults labelled as such in the dataset? R Are the images correctly classified concerning the pathology class? D Are AP/PA projections described for each x-ray image?

3.4.3 Fixing unbalanced data

This problem is common in medical imaging data centers. Because data is collected from so many different sources and since not all diseases have equal incidence, these data centers contain an imbalance.

Will it be a problem to train a neural network based on an unbalanced data center? In response, it should be said that the neural network will try to learn more from classes that have more images (compared to classes with fewer images). For this reason, it is possible that the model predicts more images of "bacterial pneumonia" while the images are from the other two classes, and the output will be unfavourable when working with medical images.

It should also be kept in mind that when interpreting medical images, the final accuracy (both accuracy in training and validation) of a model is not the right parameter to measure the performance of a model because if a model performs poorly in a particular class but shows a good performance in classes with a high number of images, the accuracy of this model will still be shown to be high. We want the model to perform very well in all classes.

It is also necessary to have separate batches of images on which the model is neither trained nor validated so that with these batches of images, we can test the performance of the model on images it has not seen before. This is an important and mandatory step to analyze the performance of a model.

There are different approaches to dealing with the problem of imbalance in classes. The best way to collect more images for classes that are in the minority. However, this method is not applicable in all situations, for which generally three techniques can be useful:

A) Error function with weighted loss;
B) Undersampling;
C) Oversampling.

3.4.3.1 Oversampling

In our research, oversampling is a process that adds more images to lower classes so that the number of images in lower classes can be compared with higher ones.

This can be done by duplication of images in lower classes. Direct duplication of the same image can cause overfitting. Therefore, to reduce overfitting, techniques (Table 5) of data augmentation can be used to create more images for lower classes (This also causes overfitting, but is a better technique than directly duplicating the original images).

Table 5. Data augmentation techniques

Data Augmentation Technique
Affine transformations Affine transformations involve applying a combination of translation, rotation, scaling, and shearing to an image. These transformations preserve parallel lines and ratios of distances between points.
Rotation Rotation refers to rotating an image by a certain angle around its center. It can be clockwise or counterclockwise.
Scaling/ Zooming Scaling or zooming involves resizing an image either by increasing or decreasing its dimensions.
Flip Flipping an image horizontally or vertically involves reflecting the image along the vertical or horizontal axis, respectively.
Horizontal Horizontal augmentation is a data augmentation technique used in computer vision tasks. It involves flipping an image horizontally along the vertical axis, effectively creating a mirror image of the original. This technique helps to increase the variability of the dataset and improve the model's ability to generalize to different orientations or viewpoints.
Vertical Vertical augmentation is another data augmentation technique used in computer vision. It involves flipping an image vertically along the horizontal axis. This creates an inverted version of the original image, which can help the model learn to recognize objects or patterns in different orientations.

Data Augmentation Technique
Shifting/Translation Shifting or translating an image involves moving it horizontally or vertically by a certain number of pixels.
Shearing Shearing is a data augmentation technique used in computer vision to introduce geometric distortions to an image. It involves tilting or slanting the image along a particular axis.
Brightness change Brightness change involves adjusting the overall brightness or darkness of an image.
Crop Cropping an image involves removing a portion of the image by selecting a specific region of interest.
Contrast change Contrast change involves modifying the difference between the light and dark areas of an image.
Gaussian noise Gaussian noise is a type of random noise that can be added to an image to simulate real-world variations and improve model robustness.
ZCA whitening transformation ZCA whitening is a technique used to decorrelate the features of an image by applying a linear transformation.
Elastic transformation Elastic transformation involves applying local deformations to an image, simulating the effect of elastic materials.
Grid distortion Grid distortion involves warping an image by manipulating a grid of control points.
Optical distortion Optical distortion refers to correcting lens distortions in images, such as barrel distortion or pincushion distortion.
Warping Warping refers to morphing or transforming an image into a different shape or appearance.
Multiple patches from each image This technique involves extracting multiple patches or sub-images from a single original image, allowing for more diverse training examples.
Class inherent transformations Network This refers to using a neural network architecture that is specifically designed to handle inherent transformations within the classes of a dataset.
Augmentation is used but parameters are not specified This indicates that data augmentation techniques were applied, but the specific parameters or values used for each transformation were not mentioned.
No augmentation used This means that no data augmentation techniques were applied, and the original dataset was used as is.

3.5 Checklist review

Table 6 summarizes which points from the checklist were met by the reviewed studies.

Table 6. List of studied checklists

Checklist/ Study
Image Preprocessing [D] Is the data preprocessing described? [D] Are artifacts (such as captions) removed?
Data Augmentation (If Needed) [D] Are the lungs fully present after transformations? [R] Are lung structures visible after brightness or contrast transformations? [D] Are only sensible transformations applied?
Model Performance [D] Are at least a few metrics proposed in use? [D] Is the model validated on a different database than the one used for training?
Domain Quality of Model Explanations [R] Are other structures (bowel loops) [R] All the areas marked as highly explanatory are located inside the lungs? [R] Are artifacts misidentified as part of the explanations? [R] Are areas indicated as explanations consistent with the opinions of radiologists? [R] Do explanations accuracy indicate lesions?

4. Experimental Results and evaluation

In this section, the results obtained from the CNN method are reviewed and evaluated, as well as those of risk assessment and management.

4.1 Pre-processing

We used several dataset-quality preprocessing methods. This process helps solve overfitting problems and increases the generalization ability of the model during training. The settings used in the image enhancement are shown in Table 7, and the data can be seen in Figure 5 according to the applied pre-processing.

Table 7. Data Preprocessing Methods

Preprocessing	Description
Resize to the Same	Resizing images¶ We proceeded to resize all images from an original size of (2090 x 1858) pixels to (20 x 20) pixels as well as cropping the borders, significantly reducing computational time.
Grayscale Normalization	Grayscale normalization refers to the process of adjusting the intensity values of an image to enhance its visual appearance or facilitate further analysis. In grayscale images, each pixel typically represents a single intensity value ranging from 0 to 255, with 0 being black and 255 being white. Normalization involves scaling these intensity values to a desired range, such as 0 to 1 or -1 to 1, to improve the contrast and dynamic range of the image. This can be achieved by dividing each intensity value by the maximum value in the image or applying more sophisticated normalization techniques, such as histogram equalization or contrast stretching.
Creating Data Frame	We initialize a data frame and fit every picture from the Normal Group into a row in the data frame. This process is called linearizing data.

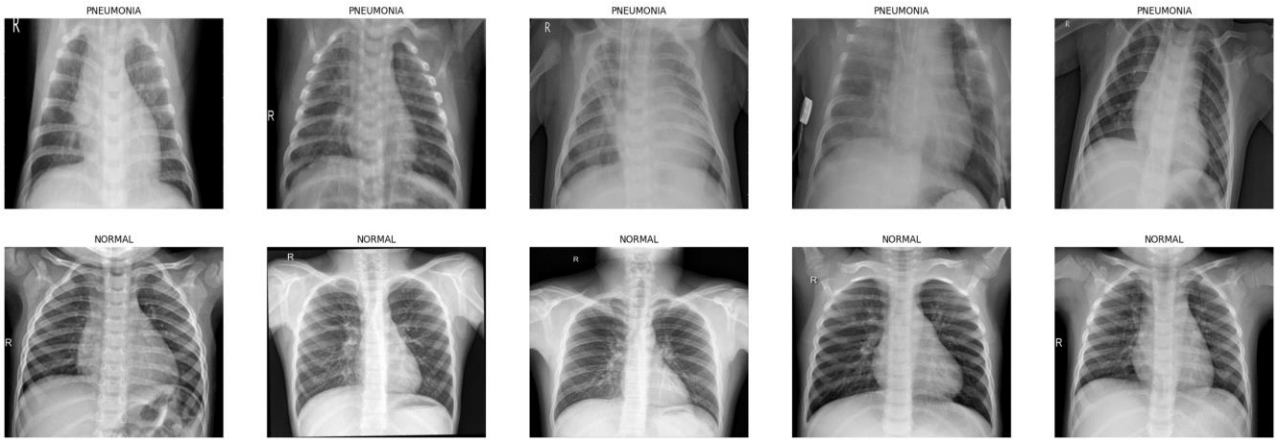

Figure 5. Normal/Pneumonia Image Visualization for Training Dataset

4.2 The Model

The network consists of several convolutional layers, normalization layers, pooling layers, and additional components that effectively process and analyze images. With increased depth, the implementation of Dropout and Batch Normalization aims to improve accuracy while mitigating overfitting. At the end of the network, a Dense layer is incorporated to make final predictions by analyzing all the extracted features. Overall, this network has 1,246,401 trainable parameters, enabling it to learn from input data effectively (Figure 6).

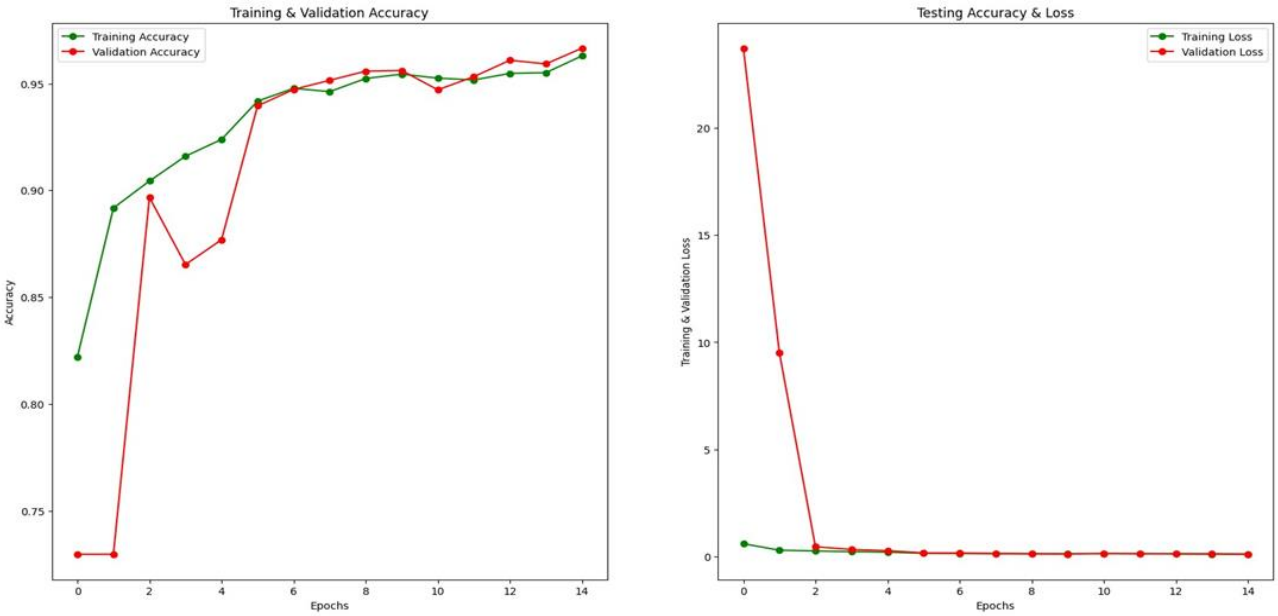

Figure 6. Accuracy and Training Validation with Primary Setting

The columns are defined as follows:

1. **Layer (type):** This column indicates the name and type of each layer. For instance, the table includes convolutional layers (Conv2D), batch normalization layers (Batch Normalization), and pooling layers (MaxPooling2D). Dropout layers are also utilized to help prevent overfitting.
2. **Output Shape:** This column presents the output shape of each layer. For example, the output of the first layer (conv2d_5) is represented as (None, 150, 150, 32), indicating that there are 32 feature maps at a resolution of 150x150 pixels, with an unspecified number of samples (None).
3. **Param:** This column denotes the number of trainable parameters in each layer. For example, the conv2d_5 layer contains 320 parameters.

Table 8. Suggestion Model for CNN Layers

Layer (type)	Output Shape	Param
conv2d_5 (Conv2D)	(None, 150, 150, 32)	320
batch_normalization_5 (BatchNormalization)	(None, 150, 150, 32)	128
max_pooling2d_5 (MaxPooling2D)	(None, 75, 75, 32)	0
conv2d_6 (Conv2D)	(None, 75, 75, 64)	18,496
dropout_4 (Dropout)	(None, 75, 75, 64)	0
batch_normalization_6 (BatchNormalization)	(None, 75, 75, 64)	256
max_pooling2d_6 (MaxPooling2D)	(None, 38, 38, 64)	0
conv2d_7 (Conv2D)	(None, 38, 38, 64)	36,928
batch_normalization_7 (BatchNormalization)	(None, 38, 38, 64)	256
max_pooling2d_7 (MaxPooling2D)	(None, 38, 38, 64)	0
conv2d_8 (Conv2D)	(None, 38, 38, 64)	73,856
dropout_5 (Dropout)	(None, 19, 19, 128)	0
batch_normalization_8 (BatchNormalization)	(None, 19, 19, 128)	512
max_pooling2d_8 (MaxPooling2D)	(None, 10, 10, 128)	0
conv2d_9 (Conv2D)	(None, 10, 10, 256)	295,168
dropout_6 (Dropout)	(None, 10, 10, 256)	0
batch_normalization_9 (BatchNormalization)	(None, 10, 10, 256)	1,024
max_pooling2d_9 (MaxPooling2D)	(None, 10, 10, 128)	0
flatten_1 (Flatten)	(None, 6400)	0
dense_2 (Dense)	(None, 128)	819,328
dropout_7 (Dropout)	(None, 128)	0
dense_3 (Dense)	(None, 1)	129

4.3 Data Augmentation

The chest X-ray imaging data center shows an imbalance in Figures 7 and 8. Here, we have shown this imbalance with two pie charts and a histogram based on their class. In the following, using data augmentation, we will increase the dataset of "normal" images to correct this imbalance.

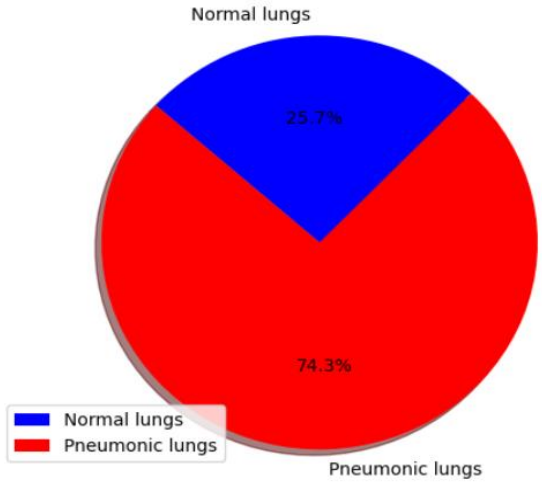

Figure 7. The imbalance in the images of the two classes

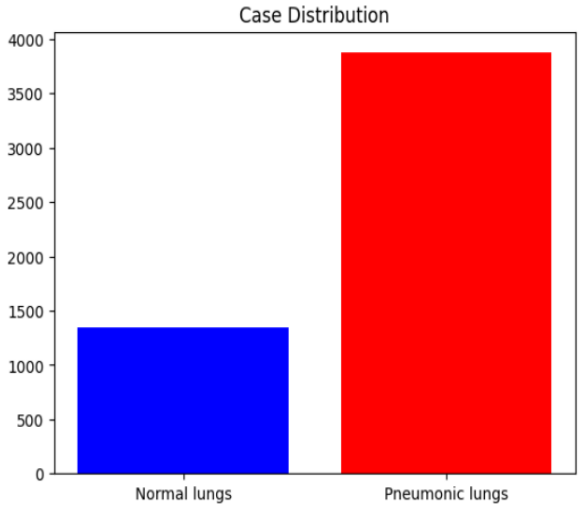

Figure 8. The imbalance in the images of the two classes

Using data augmentation techniques to expand the dataset is a widely recognized method to mitigate overfitting in machine learning models. By applying small transformations to the training data while keeping the labels unchanged, variations in the dataset can be effectively reproduced, thereby creating a more robust model. Some popular data augmentation techniques include:

1- Grayscales; 2- Horizontal flips; 3- Vertical flips; 4- Random crops; 5- Dithering, 6- Offsets and 7- Rotations.

By implementing these transformations, the dataset can be effectively enlarged, potentially doubling or tripling the number of training samples We have considered initial settings for data augmentation techniques. Are shown in Table 9.

The charts related to training and validation accuracy and loss can provide valuable insights into the model's performance. Generally, with the increase in training data through data augmentation techniques

(based on the initial settings shown in Table 9), we will observe a better trend in training and validation accuracy, as illustrated in Figure 8.

Also, the results presented in Table 10 examine the impact of different data augmentation and normalization methods on the performance of a model in recognizing pneumonia and normal conditions. Here, a summary of the results found in the table and their interpretation is provided:

Table 9. Initial setting of data augmentation

Data Augmentation Technique	
Rotation	**30 degrees**
Scaling/ Zooming	**20% of some training images**
Shifting/Translation	**10% of the width** **10% of the height**

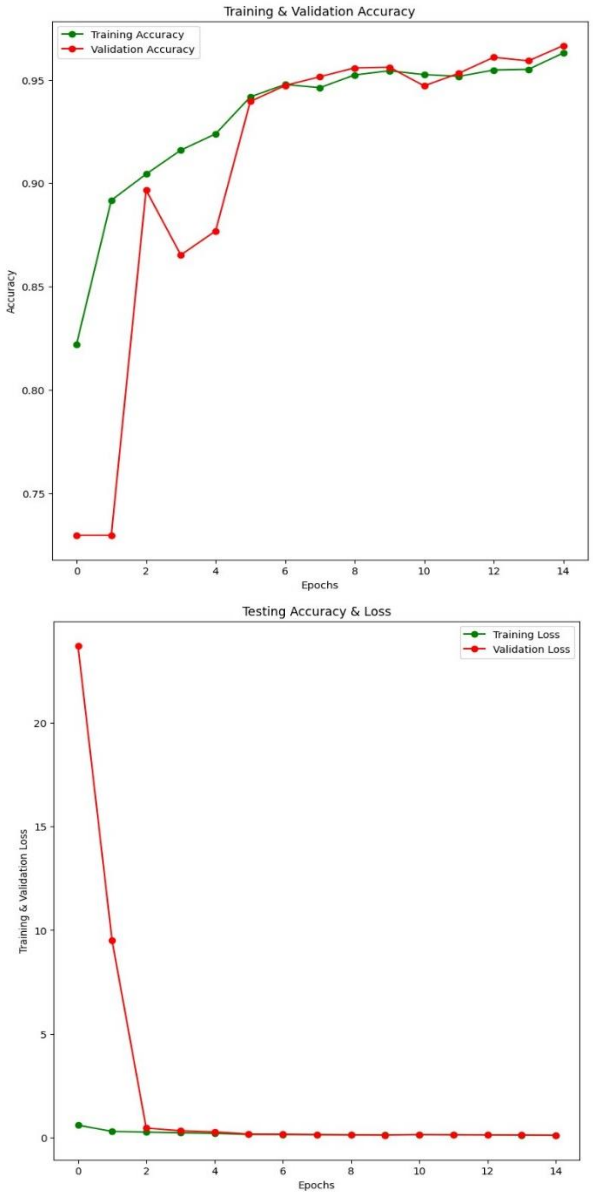

Figure 8. Accuracy and Training Validation with Primary Setting

Table 10. Results with the primary setting

	Precision	Recall	F1-Score
Results with the primary setting			
Pneumonia (Class 0)	0.99	0.95	**0.97**
Normal (Class 1)	0.87	0.97	**0.92**
accuracy			**0.95**
Macro avg	0.93	0.96	**0.94**
Weighted avg	0.96	0.95	**0.95**
Results with samplewise_center method			
Pneumonia (Class 0)	0.74	1.00	**0.85**
Normal (Class 1)	0.85	0.04	**0.07**
accuracy			**0.74**
Macro avg	0.79	0.52	**0.46**
Weighted avg	0.77	0.74	**0.64**
Results with featurewise_std_normalization method			
Pneumonia (Class 0)	0.73	1.00	**0.84**
Normal (Class 1)	0.00	0.00	**0.00**
accuracy			**0.73**
Macro avg	0.36	0.50	**0.42**
Weighted avg	0.53	0.73	**0.62**
Results with samplewise_std_normalization method			
Pneumonia (Class 0)	0.73	1.00	**0.84**
Normal (Class 1)	0.00	0.00	**0.00**
accuracy			**0.73**
Macro avg	0.36	0.50	**0.42**
Weighted avg	0.53	0.73	**0.62**
Results with rotation_range =80			
Pneumonia (Class 0)	0.99	0.86	**0.92**
Normal (Class 1)	0.73	0.98	**0.84**
accuracy			**0.90**
Macro avg	0.86	0.92	**0.88**
Weighted avg	0.92	0.90	**0.90**
Results with zoom_range = 0.8			
Pneumonia (Class 0)	0.98	0.95	**0.96**
Normal (Class 1)	0.86	0.94	**0.90**
accuracy			**0.94**
Macro avg	0.92	0.94	**0.93**
Weighted avg	0.95	0.94	**0.94**
Results with_shift_range			
Pneumonia (Class 0)	0.98	0.96	**0.97**
Normal (Class 1)	0.90	0.95	**0.92**
accuracy			**0.96**
Macro avg	0.94	0.95	**0.95**
Weighted avg	0.96	0.96	**0.96**
Results with vertical flip			
Pneumonia (Class 0)	0.99	0.87	**0.93**
Normal (Class 1)	0.74	0.98	**0.84**
accuracy			**0.90**
Macro avg	0.86	0.92	**0.88**
Weighted avg	0.92	0.90	**0.90**

4.4 Checklists assessment

We have taken steps to assess the quality of the presented methods. First, according to Table 11, we checked (the requirements and obligations of AI tools) in our algorithm, and then in Table 12, we reviewed the assessment list of medical AI tools in our algorithm.

Table 11. Requirements and Obligations of high-risk artificial intelligence tools

• Using quality data for training, validation and testing (relevant data and representative data)
To balance the proportion of data for the training and validation sets, the main classification of data was changed. We transformed the entire dataset into training and validation sets. A total of 3,722 images were assigned to the training set and 2,134 images were assigned to the validation set to improve the validation accuracy. It should be noted that this dataset was prepared and compiled from the validated Kaggle site.

• Preparation of technical documentation and adjustment of system logging capabilities (traceability and auditability).
The system features and the way the system works are prepared and adjusted in the requirements analysis phase with a group of experts.

• Ensuring the appropriate level of transparency and providing information about the capabilities and limitations of the system and its usage mode for users
The system features and the way the system works are prepared and adjusted during product development with a group of experts.

• Ensuring the existence of human supervision (measures available inside the system or actions that must be implemented by users)
Human supervision is done in the product maintenance phase after the end of the product development process

• Ensuring strength, precision and cyber security
The accuracy of the system is checked in this paper

Table 12. Assessment list of medical AI device

• What is the purpose and context of the algorithm?
Disease diagnosis using image data (lung damage diagnosis)

• Was there enough data to train the algorithm? For appropriate validation and accuracy, we used a total of 3,722 training images and 2,134 validation images.

• How well does the algorithm perform?
The performance of the algorithm is checked in **Table 6**.

• Can the algorithm be transferred to a new clinical environmen
Yes. Image analysis methods and algorithms used for other similar cases apply to similar cases.

• Are the outputs of the algorithm clinically understandable?
Yes. According to the expression of the results schematically and numerically, the outputs of the algorithm are clinically understandable.

• How does the algorithm fit with the current work routine and how does it complement it?
This method complements the doctors' diagnosis.

• Has the use of the algorithm been found to improve patient outcomes and patient care?
These methods are complementary to disease diagnosis and patient care and help doctors diagnose or predict outcomes more accurately.

• Can the algorithm harm the patient?
If the diagnosis of the disease is wrong, it will lead to the prescription of the wrong drugs and as a result harm to the patient.

• Does the use of the algorithm cause ethical, legal or social concerns?
All the data used should be in such a way that the privacy of the patients is preserved. Similarly, in case of wrong diagnosis, moral and legal concerns are raised.

In continuation of our investigations, Table 14 indicates the pre-processing of the data of this data set and Table 13 shows the checklists.

Table 13. Checklists assessed on the use case

Checklist/ Study	
Image Preprocessing [D] Is the data preprocessing described?	✓
[D] Are artifacts (such as captions) removed?	✓
Data Augmentation (If Needed) [D] Are the lungs fully present after transformations?	✓
[R] Are lung structures visible after brightness or contrast transformations?	✓
[D] Are only sensible transformations applied?	✓
Model Performance [D] Are at least a few metrics proposed in use?	✓
[D] Is the model validated on a different database than the one used for training?	✓
Domain Quality of Model Explanations [R] Are other structures (bowel loops)	✓
[R] All the areas marked as highly explanatory are located inside the lungs?	✓
[R] Are artifacts misidentified as part of the explanations?	✓
[R] Are areas indicared as explanations consistent with the opinions of radiologists?	✓
[R] Do explanations accuracy indicate lesions?	✓

Table 14. Pre-processing performed on the data set

D Do the data and related information provide sufficient diagnostic quality? If the images are in DICOM, does the header provide the required information? If not, is it presented in another way?
In the Pneumonia dataset, detailed information about the diagnostic quality of data and associated information is not available.

R Are low-quality images (i.e., blurry, too dark, or too bright) rejected?
For the analysis of chest x-ray images, all chest radiographs were initially screened for quality control by removing all low-quality or unreadable scans.

D Is the data set gender and age balanced?
Gender and age are not considered characteristics of the data set.

R Does the dataset contain one type of image (CT or X-ray)?
There are only chest X-ray images.

R Are lung structures (lung window) visible on CT images?
Yes, it is completely visible**.**

D Are images of children and adults labelled as such in the dataset?
No, this labelling is not done in this data set.

R Are the images correctly classified about the pathology class?
It turns out that a total of 5863 images are available in the challenge, which is divided into two categories and normally they are also divided into training/testing and validation sets.

D Are AP/PA projections described for each x-ray image?
The original dataset consists of chest X-ray images and other information such as the type of prognosis (AP or PA) is not available for each X-ray image.

4.5 Assessment using performance factors

In this section, we examined the qualitative assessments presented in sections (3-5 of Table 3) on the data of our study and showed the results in Table 15.

The analysis of Tables 16 and 17 illustrates the impact of adversarial attacks on the performance of an AI model. While these attacks may increase recall, they often come at the expense of precision and overall accuracy. Understanding and mitigating these effects can significantly enhance the resilience of the model. Continuous assessment and rigorous testing in diverse contexts are crucial for comprehending and improving AI systems, especially when they are vulnerable to adversarial influences.

Table 15. Assessment using functional factors

Precision	The precision is shown in Table 10.
Reliability	According to the tests performed, crystal test and pressure test (Tables 16 and 17), it can be seen that the system was able to preserve the quality.
Executability	Executability was observed to be maintained for all different models and imaging modalities as well as all patient populations for which it was designed.
Certainty	The algorithm was executed several times and it was seen that the same results were obtained from it.
Irreversibility	-
Awareness of limitations	-
Lack of failure	If the accuracy obtained is less than the specified rate, the result is considered to be wrong.
Clear logic	The step-by-step algorithm describes the stages of preprocessing, classification and accuracy.
Transparent rate of confidence	The algorithm informs the user about the accuracy of the desired results after each execution.
Controllability	-
Editability	The algorithm declares the accuracy after each execution and the stated accuracy is checked continuously.
Intuitive user interface	-

Table 16. Crystal test of test data

	Precision	Recall	F1-score
Pneumonia	0.77	0.97	0.86
Normal	0.74	0.21	0.32
accuracy			0.77

Table 17. Adversarial attack on test data

	Precision	Recall	F1-Score
Pneumonia	0.73	1.00	0.84
Normal			
accuracy			0.73
Loss			0.60

5. Conclusion

In recent years, the use of artificial intelligence (AI) in medicine and healthcare has been praised for the great promise it offers, but it has also been at the center of heated controversy. This article discusses how artificial intelligence can influence the future of healthcare, especially increasing the efficiency of doctors, and improving diagnosis and medical treatment. It examines the main risks of AI applications in healthcare and suggests mitigation measures to minimize these risks. The World Health Organization estimates that more than 4,000,000 premature deaths occur annually from air pollution-related illnesses, including pneumonia. For these populations, an accurate and rapid diagnosis is of high importance as it can provide them with access to timely treatment, thus reducing the time and cost that is so valuable for people living in poverty.

Prioritizing quality assurance and risk assessment in the development and implementation of AI-based products, particularly in sensitive domains such as medical diagnosis, is of paramount importance. Addressing issues related to bias and a lack of transparency in AI systems is crucial for ensuring their validity and efficacy in medical applications. In this article, we designed the Panoumia diagnostic system and subsequently evaluated it using both quantitative metrics and a qualitative checklist. We provided methods for evaluating products based on medical images (using methods based on evaluation checklists and quantitative methods to examine data imbalance. To achieve a more robust model, we employed data augmentation techniques, which allowed us to achieve an accuracy level exceeding 90% in most of these methods). Furthermore, in the evaluation section with functional factors, we conducted two tests: the pressure and crystal tests, both of which achieved an accuracy of over 70%. In the checklist evaluation section, our images and data were meticulously assessed and reviewed using the checklists.

For future research, we recommend exploring additional evaluation methods [16] and AI-driven products to assess and improve the reliability and maintenance of these systems.

Conflict of Interests

No conflict of interest has been expressed by the authors.

6. Reference

[1] S. Secinaro, D. Calandra, A. Secinaro, V. Muthurangu, and P. Biancone, "The role of artificial intelligence in healthcare: a structured literature review," *BMC Medical Informatics and Decision Making*, vol. 21, 2021, Art. no. 125, doi: https://doi.org/10.1186/s12911-021-01488-9.

[2] G. Verma and S. Prakash, "Pneumonia classification using deep learning in healthcare," *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, no. 4, pp. 1715-1723, 2020, doi: http://doi.org/10.35940/ijitee.D1599.029420.

[3] M. A. Al-Antari, "Artificial intelligence for medical diagnostics-existing and future AI technology!," *Diagnostics*, vol. 13, no. 4, p. 688, 2023, doi: https://doi.org/10.3390/diagnostics13040688.

[4] S. Kaur *et al*., "Medical diagnostic systems using artificial intelligence (AI) algorithms: Principles and perspectives," *IEEE Access*, vol. 8, pp. 228049-228069, 2020, doi: https://doi.org/10.1109/ACCESS.2020.3042273.

[5] M. Mirbabaie, S. Stieglitz, and N. R. Frick, "Artificial intelligence in disease diagnostics: A critical review and classification of the current state of research guiding future direction," *Health Technology*, vol. 11, no. 4, pp. 693-731, 2021, doi: https://doi.org/10.1007/s12553-021-00555-5.

[6] Y. Kumar, A. Koul, R. Singla, and M. F. Ijaz, "Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework, and future research agenda," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 7, pp. 8459-8486, 2023, doi: https://doi.org/10.1007/s12652-021-03612-z.

[7] W. B. Schwartz, "Medicine and the computer: the promise and problems of change," in *Use and Impact of Computers in Clinical Medicine*, New York: Springer, 1970, pp. 321-335, doi: https://doi.org/10.1007/978-1-4613-8674-2_20.

[8] L. Seyyed-Kalantari, G. Liu, M. McDermott, I. Y. Chen, and M. Ghassemi, "CheXclusion: fairness gaps in deep chest X-ray classifiers," in *Pacific Symposium on Biocomputing*, 2021, pp. 232-243, doi: https://doi.org/10.1142/9789811232701_0022.

[9] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *The Lancet Digital Health*, vol. 3, no. 11, pp. e745-e750, 2021, doi: https://doi.org/10.1016/S2589-7500(21)00208-9.

[10] A. Çınar, M. Yıldırım, and Y. Eroğlu, "Classification of pneumonia cell images using improved ResNet50 model," *Traitement du Signal*, vol. 38, no. 1, pp. 165-173, 2021, doi: http://dx.doi.org/10.18280/ts.380117.

[11] K. El Asnaoui, Y. Chawki, and A. Idri, "Automated methods for detection and classification pneumonia based on x-ray images using deep learning," in *Artificial Intelligence and Blockchain for Future Cybersecurity Applications*, Cham: Springer, 2021, pp. 257-284, doi: https://doi.org/10.1007/978-3-030-74575-2_14.

[12] A. U. Ibrahim, M. Ozsoz, S. Serte, F. Al-Turjman, and P. S. Yakoi, "Pneumonia classification using deep learning from chest X-ray images during COVID-19," Cognitive Computation, vol. 16, pp. 1589–1601, 2024, doi: https://doi.org/10.1007/s12559-020-09787-5.

[13] F. Jiang et al., "Artificial intelligence in healthcare: past, present and future," *Stroke and Vascular Neurology*, vol. 2, no. 4, p. 230, 2017, doi: https://doi.org/10.1136/svn-2017-000101.

[14] G. Rong, A. Mendez, E. Bou Assi, B. Zhao, and M. Sawan, "Artificial intelligence in healthcare: review and prediction case studies," *Engineering*, vol. 6, no. 3, pp. 291-301, 2020, doi: https://doi.org/10.1016/j.eng.2019.08.015.

[15] W. Hryniewska, P. Bombiński, P. Szatkowski, P. Tomaszewska, A. Przelaskowski, and P. Biecek, "Checklist for responsible deep learning modeling of medical images based on COVID-19 detection studies," *Pattern Recognition*, vol. 118, p. 108035, 2021, doi: https://doi.org/10.1016/j.patcog.2021.108035.

[16] S. Oveisi, F. Gholamrezaie, N. Qajari, M. S. Moein, and M. Goodarzi, "Review of artificial intelligence-based systems: evaluation, standards, and methods," *Advances in the Standards & Applied Sciences*, vol. 2, no. 2, pp. 4-29, 2024, doi: https://doi.org/10.22034/asas.2024.450378.1055.