International Journal of Network Dynamics and Intelligence Scilight

Article

Can AI See Bias in X-ray Images?

Kwasniewska Alicja*, and Szankin Maciej*

[1] SiMa Technologies Inc. 226 Airport Parkway, Suite 550 San Jose, CA 95110, United States
[2] Intel Corporation 16409 W Bernardo Dr Suite 100, San Diego, CA 92127, United States
* Correspondence: alicja.kwasniewska@sima.ai; maciej.szankin@intel.com

Received: 24 October 2022
Accepted: 13 December 2022
Published: 22 December 2022

Abstract: Recent advances in artificial intelligence (AI) have shown promising results in various image-based systems, improving accuracy and throughput, while reducing latency. All these factors are crucial in healthcare and have generated increased interest in this technology. However, there are also multiple challenges integrating AI in existing systems, such as poor explainability, data imbalance and bias. These challenges affect the reliability of the neural networks used in AI applications. The limitations may significantly affect the quality and cost of medical care by introducing false positive diagnosis. The false positives subsequently lead to increased stress in patients and necessitate additional testing and procedures. Lack of rich data representing all socio-economic groups can also undermine reliable decisions for underrepresented groups. Although various studies discussed techniques that may help with bias mitigation, to the best of our knowledge, no practical experiments have been conducted so far that compare different reweighting approaches using convolutional neural networks (CNN). This work focuses on in-depth explanatory analysis of chest X-ray datasets to understand and quantify the problem of class imbalance and bias. After that, various topologies of binary classifications are compared, followed by practical applications of loss reweighting techniques and comparison of their influence of privileged, underprivileged, and overall population. Experiments proved that high classification accuracy can be achieved using an efficient model topology suitable for embedded devices, making it possible to run locally without the need for cloud processing. Preliminary results showed that performance of the model for the underprivileged class can be improved by 15% if proper weighting factors are obtained and applied during the training procedure.

Keywords: artificial intelligence; bias; computer vision; medical imaging; edge computing.

1. Introduction

Although artificial intelligence (AI) has become a transformative force, enabling innovations across numerous market segments that redefined their capabilities [1], it can also be seen as a Pandora's box, leading to multiple challenges which, when unaddressed, can have significant deleterious consequences. Some examples include the black box nature of AI [2], models' generalization capabilities [3], or sensitivity to the four Vs of Big Data (Volume, Variety, Velocity, Veracity) [4]. These issues continued to be addressed in various studies on improving explainability, training procedures (including new optimizers [5], data enhancements and augmentations [6].

In this study, we focus on the data imbalance problem, which is one of the biggest limitations of AI in healthcare. Data imbalance impacts neural network reliability and performance even though the model reaches the convergence step. The data imbalance occurs when a dataset contains most examples representing the examined condition and no examples present for healthy participants [7], as well as various bias, such as sampling bias (e.g., labels inconsistency) or selection bias (e.g., overrepresentation of certain groups). Data imbalance and bias may significantly affect the quality and cost of medical care by introducing false positive diagnosis [8] and thus lead to increased stress in patients and a need for performing additional procedures [9]. Lack of rich data representing all socio-economic groups can undermine reliable decision making for the underrepresented group [10] and inequality in estimated prediction relevance [11].

Processing of medical data with deep neural networks (DNN) and convolutional neural network (CNN) in particular, is a well-studied problem. The current proposed solutions address analysis of various medical imaging modalities, including ultrasound [12], magnetic resonance imaging [13], X-ray [14], computed tomography [15], thermal imaging [16], and other human system interaction interfaces [17]. The application of AI technology in medicine has clear benefits, such as decreased cost and time of decision processes [18], possibility for combining data from different modalities [19], automated analysis of multiple inputs and frames in a few seconds [20], decision support and guidance for professionals [21], as well as access to telehealth, which is crucial for rural areas [22]. In addition, recently popular precision medicine [23] can provide even more advantages, e.g., possibility of tailoring services to individuals based on disease history, demographics, or treatment response [24]. CNNs have shown particularly good results in image processing tasks, due to their basic concepts such as weight sharing and translation invariance, which is allowed for repurposing models to many tasks even with the limited size of available training samples [25].

Although many of the recent studies showed very promising preliminary results in detection and classification of medical data, many of them don't address the class imbalance and bias problems. Luckily, a separate branch of research is looking into solutions to mitigate these limitations. Norori N. et al. [26] discussed various ways of addressing bias in AI that includes sharing of data and algorithms as well as patient-centered AI development. New research directions to address bias present in healthcare were also analyzed by Ntoutsi E. et al. [27] covering areas such as instance weighting and selection, model regularization, score corrections, and others. Fletcher R. et al. [28] described three criteria for evaluation of AI systems to provide better guidance for healthcare, i.e., appropriateness, farness, and bias. Another survey on bias types and influence on medical decisions was presented by Mehrabi N. et al. [29]. Some initial attempts for distributions rebalancing have been also evaluated [30].

Various approaches have been already discussed in the literature to mitigate class imbalance and bias, however, to the best of our knowledge, no practical experiments have been performed with such approaches in CNN-based X-ray classification systems to prove their robustness. In the view of the foregoing, the contribution of this work is threefold: (1) first, a detailed analysis of chest X-ray datasets of 14 common lung disease categories is performed, including data cleaning, visualization, common features selection, and distribution inspection; (2) secondly, the presence of class imbalance and bias is examined by comparing the probability of favorable outcome for privileged and unprivileged instances; (3) and lastly, various CNN topologies are compared to select the best performing one and evaluate the possibility for bias mitigation using different weighting approaches.

The rest of the paper is structured as follows: Section 2 covers the exploratory data analysis focused on understanding the structure of the X-ray dataset commonly used in AI research. The analysis includes assessment of distribution of available samples across various classes, followed by evaluation of dataset fairness and classes reweighting transformations. Section 3 presents comparison of performance of different deep neural network topologies in the task of X-ray image classification. Additionally, an in-depth examination of the influence of various weighting operations on the model performance is summarized in this section. Finally, the preliminary results are discussed in Section 4 and the work is concluded in Section 5.

2. Exploratory Data Analysis

2.1. Dataset

The NIH Chest X-ray dataset [31] was used in this study. It contains over 100,000 samples gathered from over 30,000 patients, many of whom have been identified with advanced lung diseases. All samples have been annotated with one of 14 diseases: Infiltration, Atelectasis, Effusion, Nodule, Pneumothorax, Mass, Cardiomegaly, Pneumonia, Hernia, Emphysema, Pleural Thickening, Fibrosis, Consolidation, Edema.

A subset of images belongs to the 'No Finding' category. Some of the annotations include the location of the abnormalities, as well, as presented in Figure 1. Many of the labels were obtained using natural language processing techniques, so there could be some erroneous labels, but the NLP labeling accuracy is estimated to be >90%. The resolution of images is 1024x1024 and the set was downloaded more than 60,000 times so far.

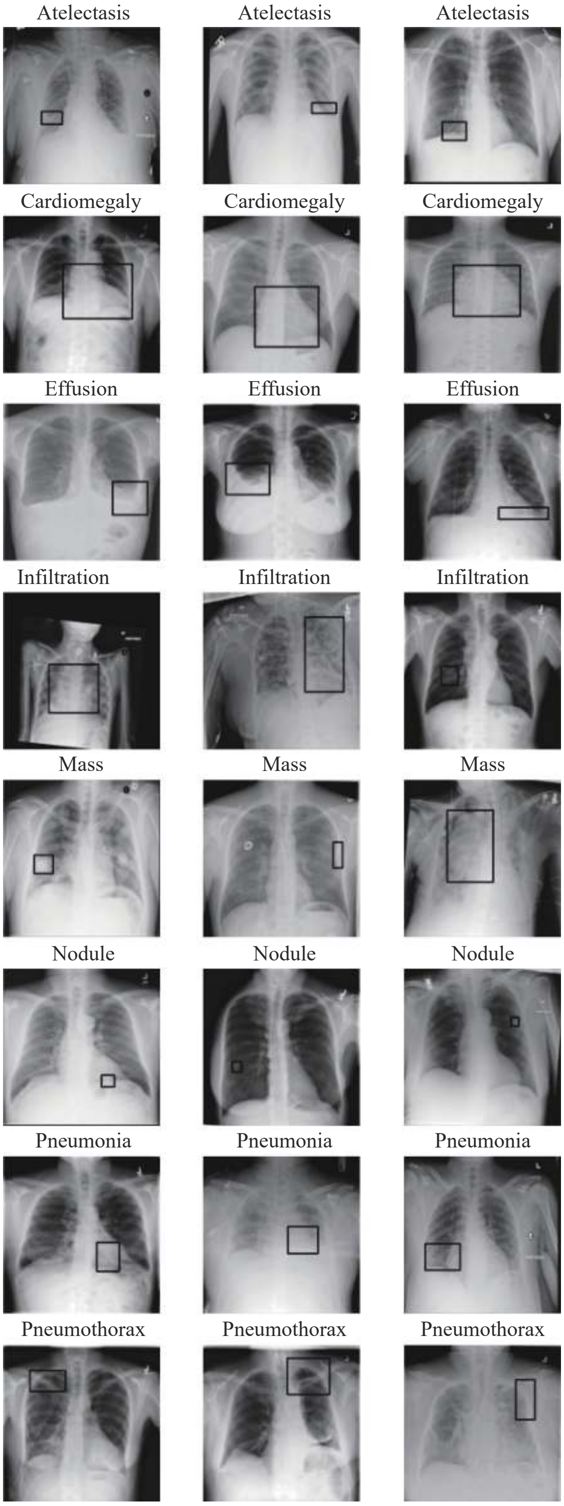

Figure 1. Visualization of samples with ground-truth bounding boxes across sample classes.

2.2. Modifications to the Dataset

In the initial phase, the dataset was carefully examined and cleaned-up to remove missing labels and simplify analysis of bias mitigation techniques. First, the samples with multiple labels were skipped. Secondly, images from the version 2 of the dataset were also removed. Finally, ~10300 random samples with the 'No Finding' category were preserved in the analyzed set, while the rest was dropped to reduce discrepancy between quantities of positive and negative predictions. Figures 2-4 show the distribution of samples in the dataset across different categories in all described steps of data preprocessing. Simultaneously, the label file (that contained metadata of recorded samples, such as patients' demographics, follow-up procedure codes, abnormality location, and others) was also examined and cleaned up from redundant or unused columns ('Unnamed: 0', 'Follow-up #', 'Patient ID', 'View Position', 'OriginalImage[Width', 'Height]', 'OriginalImagePixelSpacing[x', 'y]'). Additionally, the column representing disease classes was converted to one-hot encoded vectors required by the neural network training pipeline.

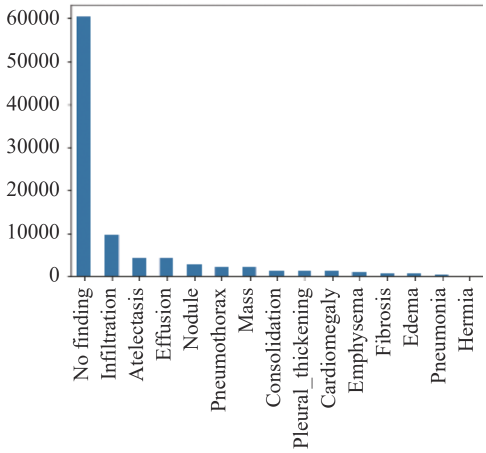

Figure 2. Number of samples per class in the Chest X-ray dataset. Samples with multiple labels were removed.

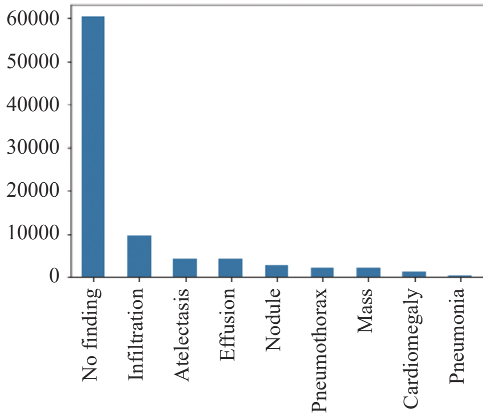

Figure 3. Number of samples per class in the Chest X-ray dataset. Dataset used in this work is a modified version, which excludes the classes introduced in the version 2 of the dataset (Hernia, Emphysema, Pleural Thickening, Fibrosis, Consolidation, Edema).

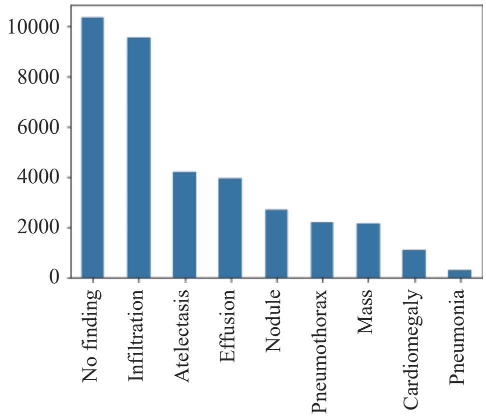

Figure 4. Number of samples per class in the Chest X-ray dataset after removing 50,000 random samples from the "No finding" class.

2.3. Exploratory Data Analysis

Data prepared in the previous step was then analyzed to evaluate class imbalance and bias problems as well as to obtain a better understanding of sample characteristics, such as the most dominant features, areas of the biggest variability per each class, and visual differences between images from various categories. A very common operation in various image-based diagnostics and monitoring solutions to reduce noise and motion artifacts is to analyze a mean image across a temporal window of the input sequence instead of each frame at a time [32]. Another way of reducing noise is to use denoising algorithms [33] or denoising neural networks [34]. The mean image can be also used to draw meaningful conclusions about abnormality characteristics for a specific disease by calculating the average of pixel values across samples collected from different subjects for each class [35, 36]. Similar analysis was performed in this study to evaluate the presence of areas of the average and the highest variability per each category in the chest X-ray images and determine if some meaningful differences between different classes are visible in the averaged frames. Mean images for each class calculated for random 25 samples are presented in Figure 5, followed by images representing absolute differences between the mean frame for each category and the mean image for the 'No finding' class (Figure 6).

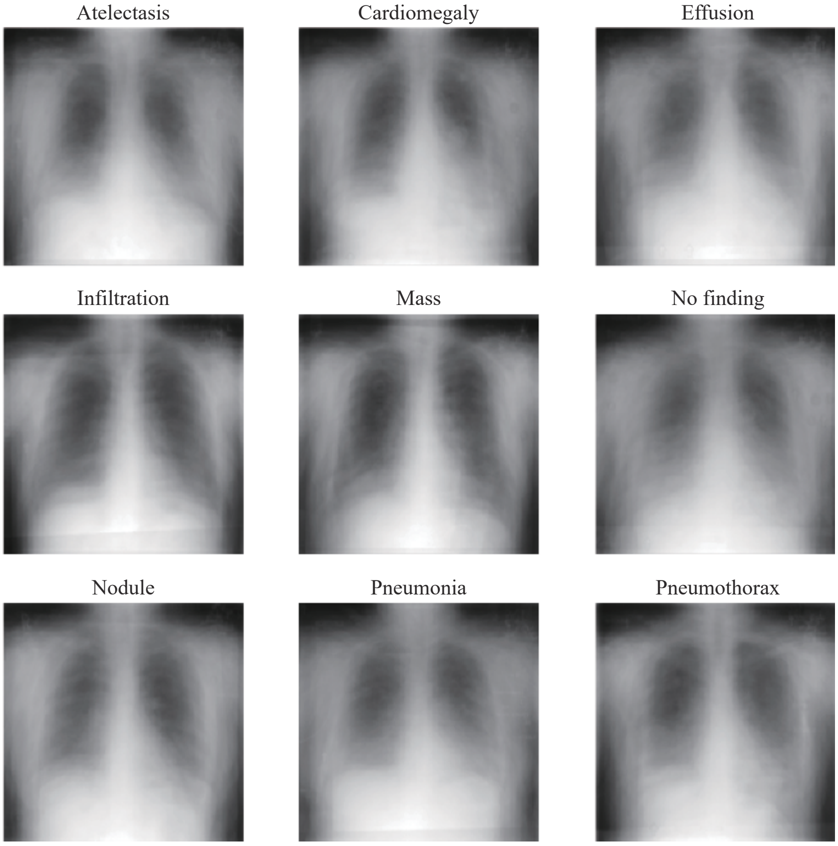

Figure 5. Mean image calculated for 25 random samples per each category.

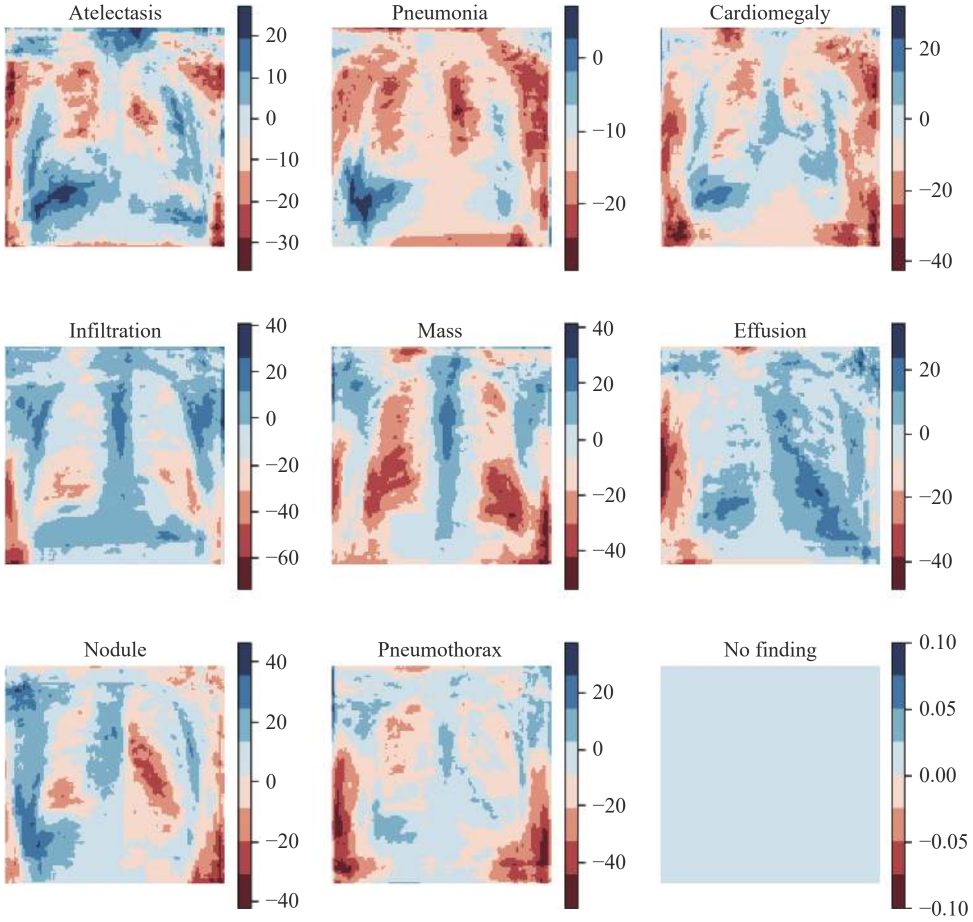

Figure 6. Absolute differences between the mean image calculated for 25 random samples per each category and the mean image calculated for 25 random samples for the 'No finding' class'.

Figures 7 and 8 show standard deviation frames and differences between the computed result frames and the standard deviation frames obtained for the 'No Finding' category, correspondingly. As can be seen, the areas of highest variability (dark blue or dark red regions) differ between classes. This confirms the possibility for CNN to learn such patterns to perform accurate classification of the analyzed lung conditions.

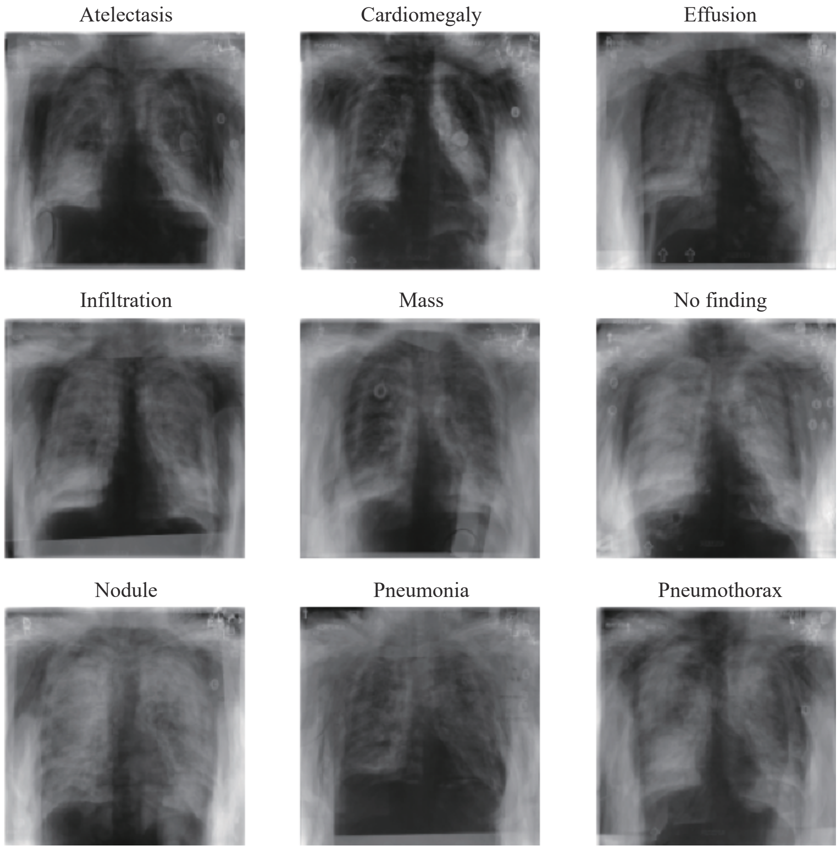

Figure 7. Standard deviation image calculated for 25 random samples per each category.

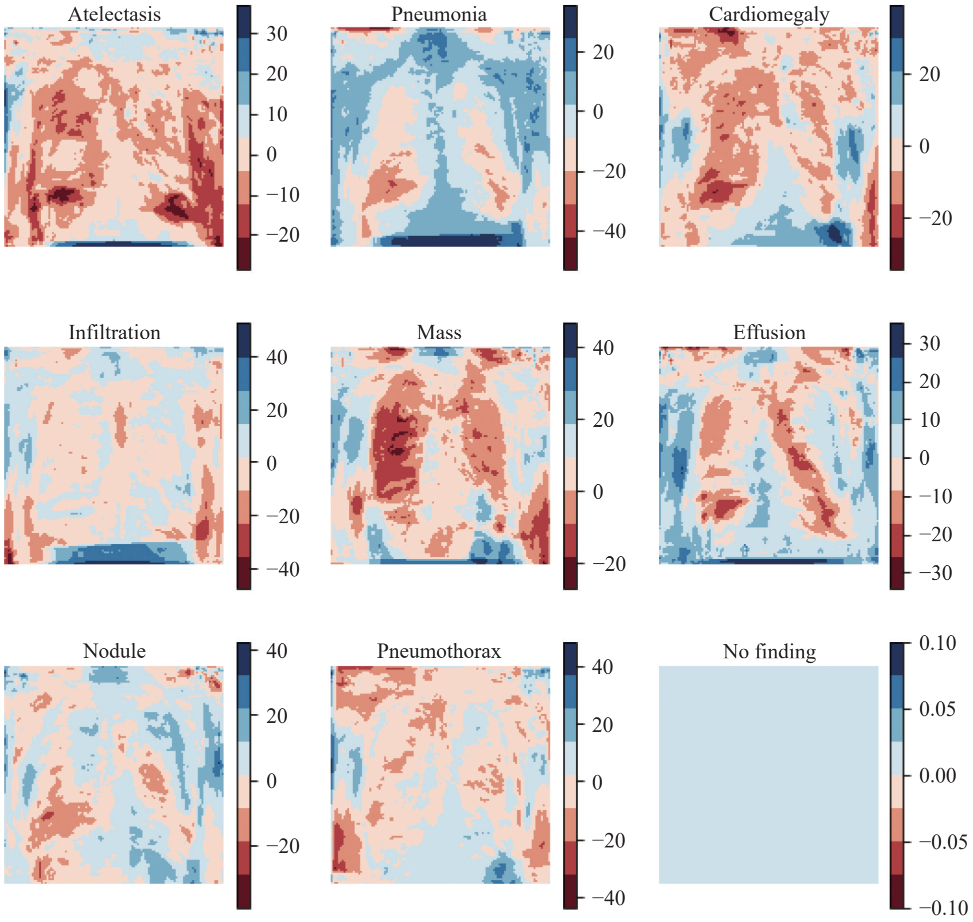

Figure 8. Absolute differences between the standard deviation image calculated for 25 random samples per each category and the standard deviation image calculated for 25 random samples for the 'No finding' class'.

Since mean and standard deviation frames were computed for different volunteers, the features such as height, weight, and gender may influence the values of result frames and potentially lead to misleading conclusions. Thus, this step was only performed to verify the presence of differences in spatial representation between different categories, yet the neural networks were trained with single samples only.

The next step of the Exploratory Data Analysis involved the use of principal component analysis (PCA) for computation of the most dominant features. Such summary of features was obtained by transforming data into fewer dimensions using singular value decomposition (SVD). Eigen images obtained for each category (40% of components kept) are plotted in Figure 9. Eigen images can be used to preserve only the features that correspond to the highest variance of the set, and in this way, simplify the convergence of the neural network and the risk of overfitting. Based on the presented results, the highest differences are present in the central and lower portion of lungs for most of the categories. Taking this into account, the images could be cropped to focus on this area.

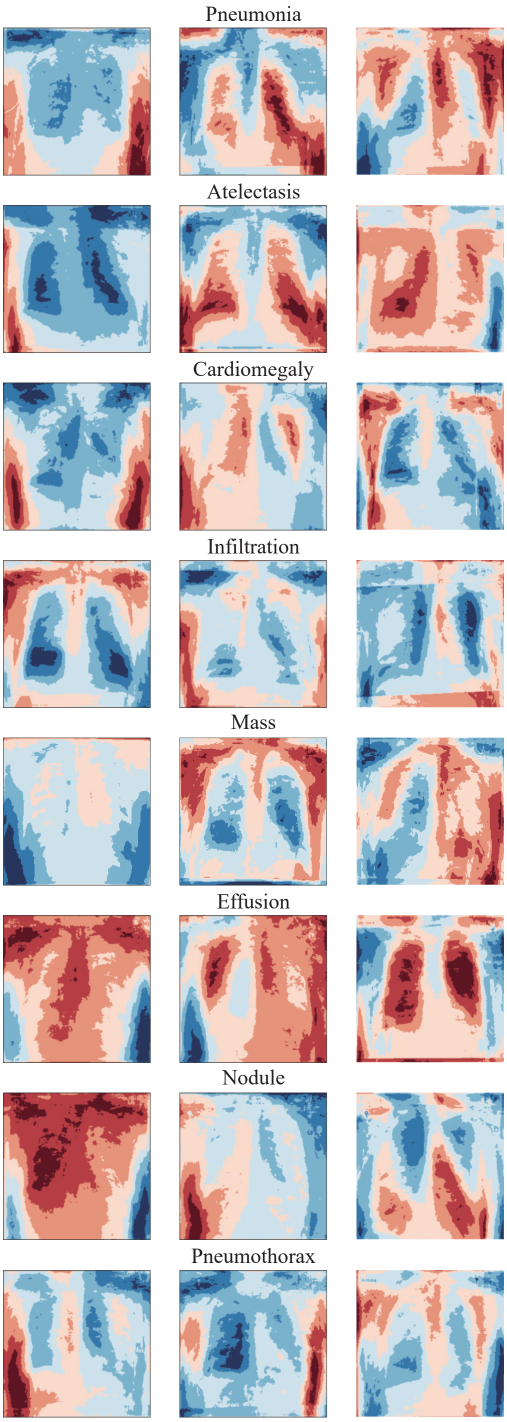

Figure 9. Eigen images obtained for each category using Singular Value Decomposition with number of components set to 40%.

2.4. Class Imbalance and Bias Evaluation

As previously mentioned, the goal of the work is to evaluate a possibility of mitigating bias and class imbalance in the task of CNN-based X-ray lung disease classification. To do this, a detailed analysis of class distribution was performed. At first, the count of samples in each category grouped by genders was obtained. Results are collected in Table 1 and plotted in Figure 10.

Table 1 Number of samples across Finding Labels categories grouped by genders

	Patient Gender	
Finding Labels	Female	Male
Atelectasis	1612	2603
Cardiomegaly	585	508
Effusion	1797	2158
Infiltration	4164	5383
Mass	838	1301
No Finding	4550	5811
Nodule	1181	1524
Pneumonia	128	194
Pneumothorax	1193	1001

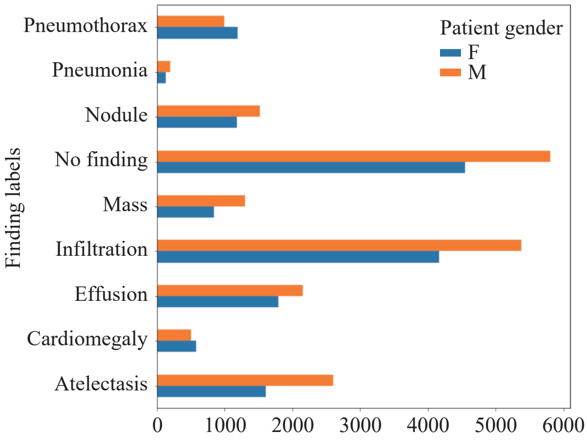

Figure 10. Distribution of samples across Finding Labels categories grouped by gender.

As can be seen, the number of samples is greater for the male category for the majority of categories except Pneumothorax and Cardiomegaly. Since the further experiments were performed for the binary classification, two categories were selected: one underrepresented by the male and the other one underrepresented by the female groups. An exemplary discussion is done for the ‘No Finding’ and ‘Pneumothorax’ classes but can be easily expanded to other categories. To amplify the imbalance and bias problems even further, random samples from the ‘No Finding’ class representing the female group were removed, preserving only 1% of samples. The obtained distribution of samples across the groups is shown in Figure 11.

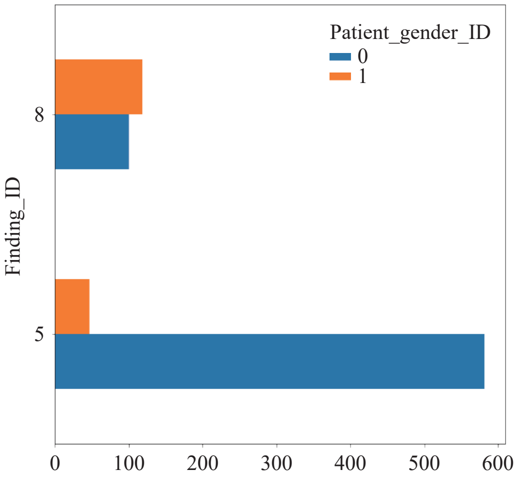

Figure 11. Distribution of samples across both categories grouped by gender after amplifying class imbalance and bias problems by removing majority of samples from one class. Finding_ID 5 and 8 correspond to the ‘No Finding’ and ‘ Pneumothorax’ categories, respectively, while gender_ID 0 and 1 correspond to male and female categories, respectively.

Evaluation of the impact of the bias and class imbalance problems on the possible favorable impact per each group was done using the IBM AI Fairness 360 Toolkit [37]. The toolkit can help to examine, report, and mitigate discrimination and bias in machine learning systems, supporting the entire application lifecycle. In our case, the fairness tool was used in the following steps:

- A *Binary Label Dataset* was created specifying ‘No Finding’ as the favorable outcome and gender as the protected attribute
- Gender female was selected as the unprivileged group, since most of the samples in the ‘No Finding’ category correspond to the male group
- Two metrics were used for evaluation of the bias and imbalance problems: Statistical parity difference (Equation (1)) and Disparate impact (Equation (2)).

Statistical parity difference (SPD) or in other words the absence of bias specifies the difference between probability (P) of the favorable outcome (O_f) for the unprivileged class (S^{up}) to the favorable outcome for the privileged class (S^p), and can be defined as:

$$SPD_c(X,S) = P\left(c(x) = O_f \middle| x \in S^{up}\right) - P\left(c(x) = O_f \middle| x \in S^p\right) \tag{1}$$

where X is the analyzed population, S is a subset of the population belonging to a specific group for which the bias is being estimated, c is the binary classifier and x is the feature input vector for the specific sample.

Disparate impact (DI) is another metric used for bias evaluation. It analyzes the same components as SPD but instead of computing their difference, the ratio between them is used, as defined in Equation (2):

$$DI_c(X,S) = \frac{P\left(c(x) = O_f \middle| x \in S^{up}\right)}{P\left(c(x) = O_f \middle| x \in S^p\right)} \tag{2}$$

Metrics calculated for the analyzed set are presented in Table 2. The negative value of the SPD metric indicates lower probability of the favorable income for the unprivileged group. Similarly, DI below 0.5 confirms the same finding.

Table 2 Bias metrics: statistical parity difference and disparate impact computed for the population from the X-ray set used in the study

Bias metrics	**value**
Statistical parity difference	-0.57
Disparate impact	0.33

To reduce the impact of the class imbalance and bias present in the analyzed population, the reweighting algorithm from the Fairness toolkit was used to compute the factors for weighing the dataset and transforming the sample weights. The factor for each class (c) and group (g) F_{gc} is calculated as:

$$F_{gc} = \frac{W(groups == g) * W(y == c)}{(W * W(g\_and\_c))} \tag{3}$$

where W is the sample weight for a specific subgroup of the set.

Computed reweighting factors for each gender and ‘Finding Labels’ class are presented in Table 3. As can be seen, the combination of the ‘No Finding’ and female received the highest factor, which compensates for the fewest number of samples in this group. These weights were later used for the neural network training to improve performance of the model for each class and ensure none of them will be favored.

Table 3 Number of samples across Finding Labels categories grouped by gender

Finding Labels	Gender	Reweighting factor
No Finding	F	2.658418
No Finding	M	0.868697
Pneumothorax	F	0.358931
Pneumothorax	M	1.762872

Additionally, a common practice is to compute class weights using a logarithm of the group size with respect to the total number of samples and dividing by a number of categories to keep the loss at the same magnitude, as defined in Equation (4).

$$F_{gc} = clip_{min}(\log\left(\frac{1}{groupsize} * \frac{totalsize}{2}\right), 1) \tag{4}$$

Factors calculated using this approach were also evaluated in our work. We will refer to them as the group size (GS) factors. Weights obtained with the first method will be referred to as fairness toolkit (FT) factors.

3. Neural Network Experimental Evaluation

This section describes the details of selecting a deep neural network for the task of X-ray data binary classification using the image set prepared and described in the previous sections.

Before feeding samples to the model, images were normalized using z-score normalization which is a usually encountered type of data spread in X-ray and radiology [38]. Also, to improve generalization capabilities, various data augmentation techniques were utilized, e.g., random rotation, shift, shear, and zoom. We also used the horizontal flip but did not apply vertical flip to preserve the standard orientation of lungs. Examples of such prepared samples from both categories are presented in Figure 12.

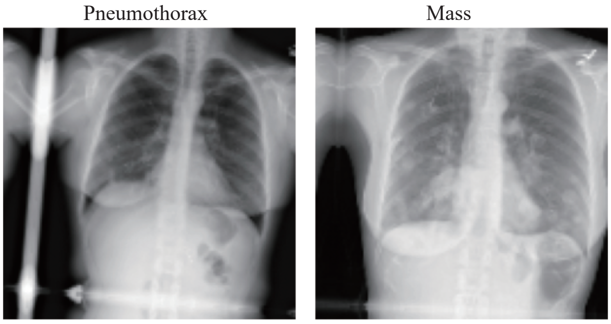

Figure 12. Examples of images from both categories after data pre-processing and augmentation.

For all experiments, the following hyperparameters were used: BATCH_SIZE = 32, IMAGE_SIZE = [128, 128], OPTIMIZER = Adam, LOSS = Binary Cross Entropy, EPOCHS = 10. The dataset split was 0.6:0.15:0.15 for the train, test, and validation sets. Various topologies of Convolutional Models were tested to determine the most accurate one. Additionally, different regularization techniques were also compared. Finally, the effect of class reweighting on the accuracy per each category was analyzed to determine if CNNs are sensitive to bias and whether it’ s possible to mitigate such limitations. The code and links to prepared datasets and checkpoints are available online for ease of result reproductivity[1]. The environment used for the experiment was Google Colab but can also be executed locally using Jupyter notebooks.

3.1. Evaluation of Various Model Topologies and Regularization Techniques

The first evaluated model was a custom shallow CNN with two convolutional layers, each followed by the max pooling operation, and three dense layers at the end. Various width and depths of each layer were evaluated. The final structure of the model is presented in Figure 13. The learning rate used in this configuration was 1e-4.

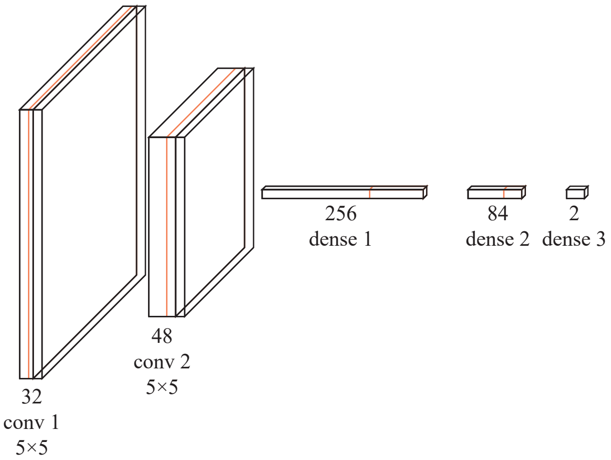

Figure 13. Topology of the shallow model, red slices indicate maxpool operator, orange slices indicate ReLU activation, at the end of the model the softmax activation was used.

Accuracy and loss functions are presented in Figure 14. After a few epochs, the validation loss increases, indicating that the model was overfitted. Thus, different regularization techniques were used. Figure 15 presents the corresponding plots for the case in which the dropout was applied. Figure 16 shows results achieved for dropouts with an additional L1, L2 kernel regularization (l1=1e-5, l2=1e-4) and L2 activity regularization (l2=1e-5). Additionally, the early stop was used to terminate the training procedure before the overfitting appears. Training curves obtained in this scenario are shown in e shown in Figure 17.

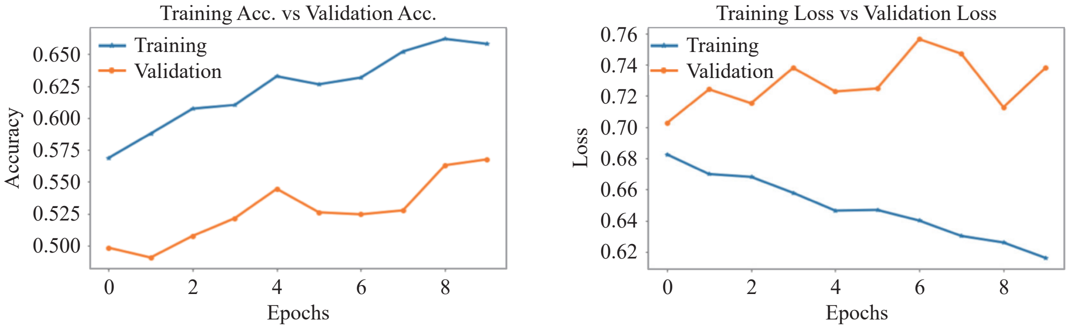

Figure 14. Accuracy and loss for the shallow model.

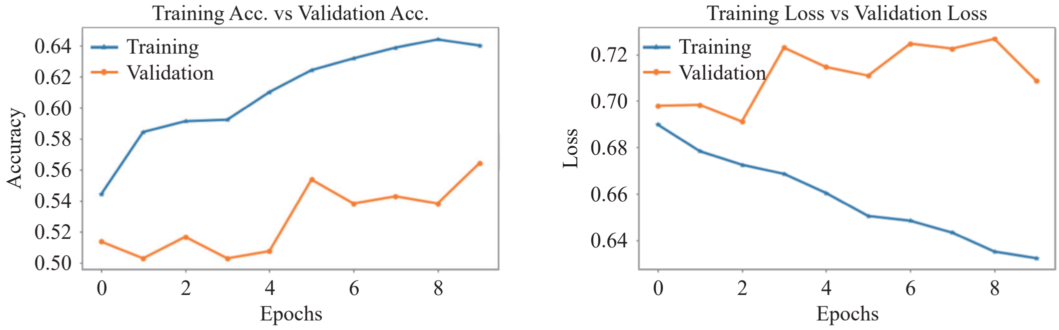

Figure 15. Accuracy and loss for the shallow model after applying the dropout.

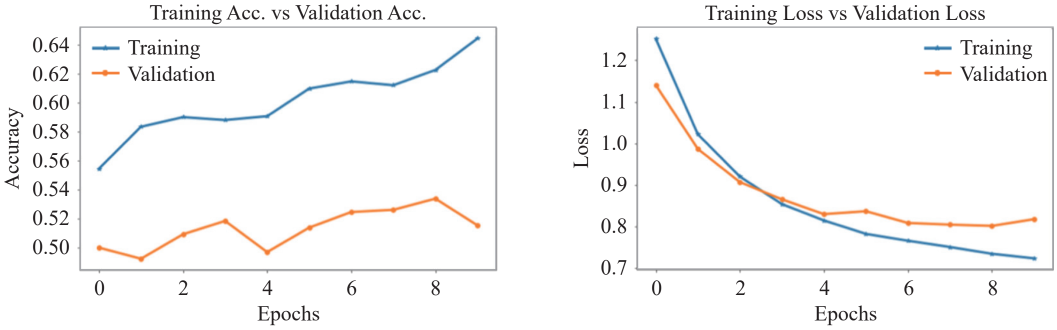

Figure 16. Accuracy and loss for the shallow model after applying the dropout and regularization.

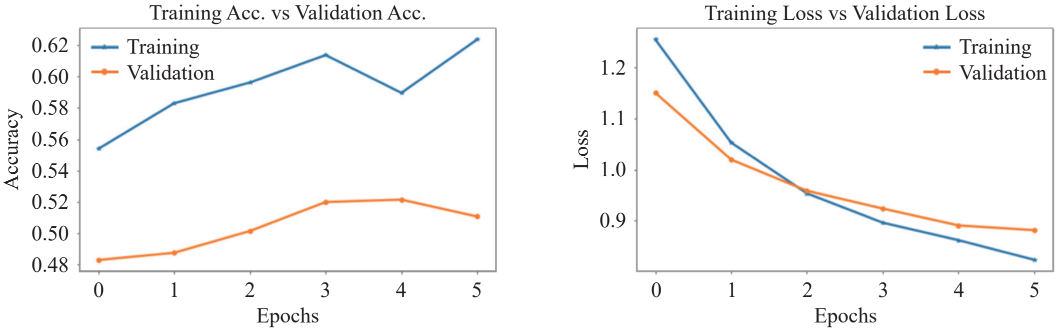

Figure 17. Accuracy and loss for the shallow model after applying the dropout and regularization.

Based on the performed analysis, it was confirmed that regularization techniques are crucial to avoid model overfitting and ensure model convergence. However, the accuracy of the model was still not satisfactory. Therefore, in the next step, the transfer learning approach was used. This method allows for repurposing the existing model, trained previously for a different task, to a new unknown problem [16]. It's achieved by reusing all layers but the last one, and fine tuning only the fully connected layer with a new dataset. The model selected for the study and fine-tuned to the binary classification of the X-ray data was EfficientNetB1 model [39]. This topology was used since it is the best model for achieving high accuracy, while being very efficient by scaling all dimensions with compound coefficients. All layers of the model pretrained on the ImageNet dataset remained unchanged, except the final classification dense layer. It was replaced with the global average pooling, followed by dense operators and softmax activation. The learning rate in the transfer learning approach should be reduced, thus, it was decreased to 1e-5. All regularization techniques used previously were also applied to this topology. Figure 18 presents accuracy and loss curves produced in this scenario.

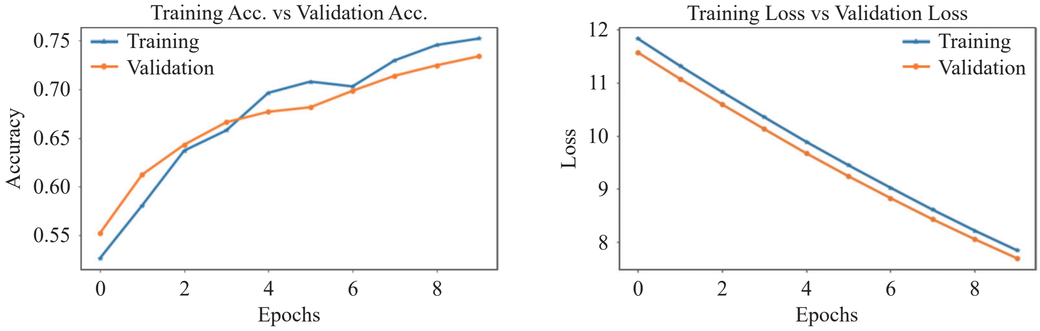

Figure 18. Accuracy and loss for the shallow model after applying the dropout and regularization.

The use of the deeper architecture and leveraging weights previously trained on larger datasets was crucial for achieving significantly better performance. Both training and validation accuracy were improved by more than 20%.

3.2. Evaluation of Class Weighting Approaches

As previously mentioned, the goal of the performed study is to (1) evaluate if CNN models are sensitive to bias and class imbalance problems and (2) compare different techniques used for bias mitigation. After comparing performance of various topologies, it was possible to move to the final experimental analysis, i.e., applications of two previously introduced reweighting factors: GS and FT factors. Since transfer learning techniques with bigger models can compensate for such issues, this step of the analysis was done by using the shallow network. Three scenarios were compared: (1) the shallow model without reweighting, (2) the shallow model with GS reweighting, and (3) the shallow model with FT reweighting. The idea of the reweighting technique is to weigh the loss computed for different samples based on whether they belong to the majority or the minority classes. A higher weight is assigned to the loss encountered by the samples associated with the minor class, as explained in the previous section. To better understand the influence of bias and class imbalance problems as well as evaluate potential performance gain from using reweighting algorithms, different metrics were used in the comparison: accuracy, precision, recall, and ROC curves. The reason for selecting ROC curves was that we wanted to consider true negatives as well, due to the class imbalance they may have influenced in the final performance of models. Results obtained for all three scenarios are presented in Figure 19 and Table 4.

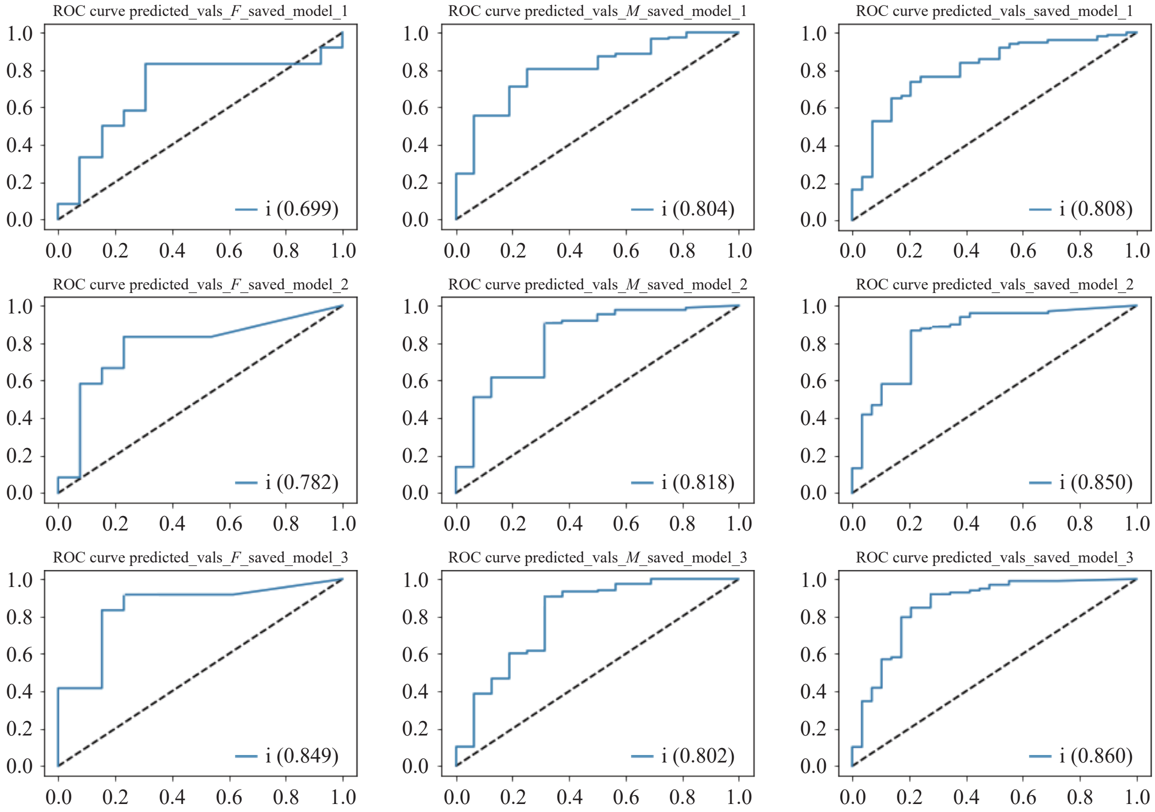

Figure 19. ROC curves for 3 analyzed scenarios. From the top row: model_1 no reweighting, model_2 reweighting using GS factors, model_3 reweighting using FT factors. From the left column: female subset only, male subset only, both genders combined.

Table 4 Performance metrics obtained for different weighting algorithms with the binary X-ray classification based on the custom shallow CNN

	Metric (validation subset, combined genders) [%]			
Scenario	**Accuracy**	**Precision**	**Recall**	**ROC**
No reweighting	0.70	**0.80**	0.76	0.80
Reweighting with GS	0.70	0.71	**0.99**	0.85
Reweighting with FT	**0.71**	0.71	**0.99**	**0.86**

4. Discussion

The presented study focuses on addressing an important problem of imparity difference, which is especially important in medical applications. Recent progress in AI shows promising results of processing medical data, however, without mitigating the bias and blackbox nature of AI, it may be difficult to convert state-of-the-art research into the state-of-the-art product.

Based on the performed experiments, the probability of a favorable outcome for the underprivileged group can be increased with reweighting approaches, which confirms the thesis of this study. As can be seen in Figure 19, the area under the ROC curve increased by 9% and 15% for GS and FT factors, respectively for the female group. This proves the necessity of bias quantification and mitigation using all available metadata and patient demographics, other than the limited categorical labels associated with the dataset. In addition, our study showed that simply taking the ratio of samples from both categories yields certain performance improvements, but may not be sufficient enough. Therefore, detailed analysis of various weighting factors is needed.

Furthermore, the selection of proper performance metrics is important in case of bias and class imbalance, as noted in Table 4. Even though the accuracy for all scenarios remains relatively unchanged, detailed analysis of threshold independent ROC curves (Figure 19) showed poor performance for the underrepresented classes, revealing the real performance and limitations of the trained classifier. Adaptation of different reweighting approaches helped to address this issue. As can be seen from Table 4 and Figure 19, the proposed reweighting approach with FT factors allowed for achieving high ROC for both genders separately (or combined). This is contrary to the case of no reweighting and GS weighting that resulted in low ROC for the female group.

Although the presented results are promising, they are only preliminary and further research is encouraged. The presented analysis was conducted for gender bias, but the similar reweighting could be performed for other protected attributes such as age bias. Figure 20 presents age distribution for each of the categories where some diseases impact only certain age groups, and this can be a bias source needing to be performed in future work.

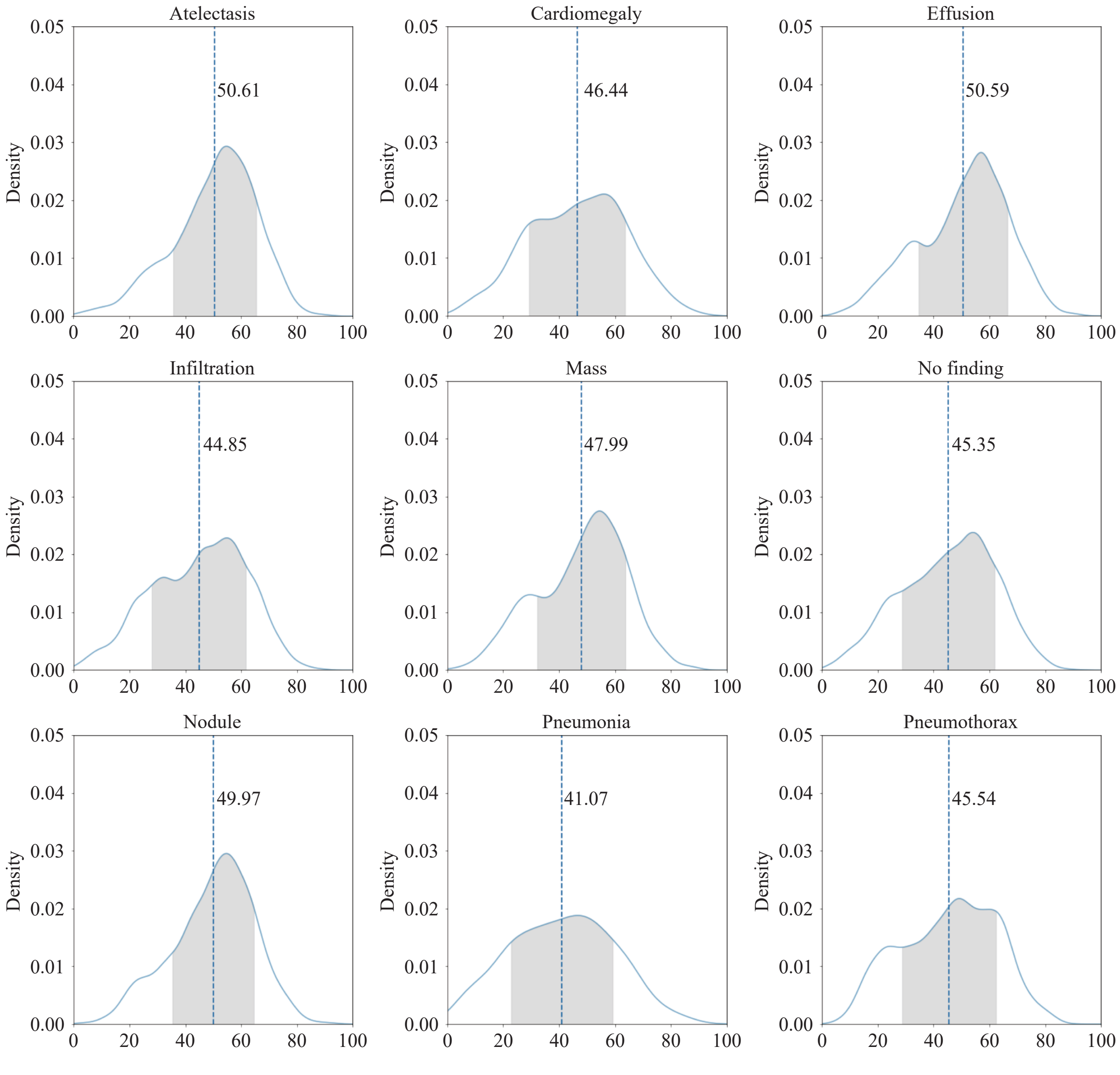

Figure 20. Relation between classes and age (mean and standard deviation marked for each class).

5. Conclusions

In this paper, an in-depth evaluation of techniques was performed for quantifying and mitigating bias in medical datasets used for CNN models. The study was performed based on the explanatory data analysis which revealed data imbalance and bias problems in the commonly used X-ray chest database. After comparison of different topologies, models were trained with the loss weighting technique, increasing importance of the underprivileged class. The preliminary results showed that such an approach is a necessity in image-based healthcare diagnostics and can improve prediction accuracy by ~15% for the minority class. Further work will focus on (1) examining other protected attributes that may suffer from the bias problem and (2) applying other methods that help with explainability and fairness of AI systems.

Data Availability Statement: All code, links to models and data are available here: https://github.com/issondl/from-data-to-solution-2021.

Author Contributions: Kwasniewska Alicja and Szankin Maciej: conceptualization; Kwasniewska Alicja and Szankin Maciej: methodology; Kwasniewska Alicja and Szankin Maciej: software; Kwasniewska Alicja: validation; Szankin Maciej: data curation; Kwasniewska Alicja and Szankin Maciej: writing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Acknowledgments: We thank all our colleagues from SiMa.ai and Intel Corporation for their invaluable feedback that greatly assisted the study.

References

1. Akter, S.; Michael, K.; Uddin, M.R.; *et al*. Transforming business using digital innovations: The application of AI, blockchain, cloud and data analytics. *Ann. Oper. Res.*, **2022**, *308*: 7−39.
2. Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, **2018**, *6*: 52138−52160.
3. Geirhos, R.; Temme, C.R.M.; Rauber, J.; et al. Generalisation in humans and deep neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, *Montréal Canada*, *3–8 December 2018*; ACM: Montréal Canada, 2018; pp. 7549–7561. doi:10.5555/3327757.3327854
4. Kwaśniewska, A.; Giczewska, A.; Rumiński, J. Big data significance in remote medical diagnostics based on deep learning techniques. *Task Quart.*, **2017**, *21*: 309−319.
5. Zhang, Z.J. Improved Adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*, *Banff*, *AB*, *Canada*, *4–6 June 2018*; IEEE: Banff, AB, Canada, 2018; pp. 1–2. doi:10.1109/IWQoS.2018.8624183
6. Mikołajczyk, A.; Grochowski, M. Data augmentation for improving deep learning in image classification problem. In *2018 International Interdisciplinary PhD Workshop (IIPhDW), Swinoujście, Poland, 9–12 May 2018*; IEEE: Swinoujście, Poland, 2018; pp. 117–122. doi:10.1109/IIPHDW.2018.8388338
7. Wang, L.; Han, M.; Li, X.J.; *et al*. Review of classification methods on unbalanced data sets. *IEEE Access*, **2021**, *9*: 64606−64628.
8. Gervasi, S.S.; Chen, I.Y.; Smith-McLallen, A.; *et al*. The potential for bias in machine learning and opportunities for health insurers to address it. *Health Aff.*, **2022**, *41*: 212−218.
9. Chin, J.C.; Seidensticker, D.F.; Lin, A.H.; *et al*. Limited use of outpatient stress testing in young patients with atypical chest pain. *Fed. Pract.*, **2018**, *35*: S30−S34.
10. Panch, T.; Mattie, H.; Atun, R. Artificial intelligence and algorithmic bias: Implications for health systems. *J. Glob. Health*, **2019**, *9*: 010318.
11. Sun, W.L.; Nasraoui, O.; Shafto, P. Evolution and impact of bias in human and machine learning algorithm interaction. *PLoS One*, **2020**, *15*: e0235502.
12. Zeimarani, B.; Costa, M.G.F.; Nurani, N.Z.; *et al*. Breast lesion classification in ultrasound images using deep convolutional neural network. *IEEE Access*, **2020**, *8*: 133349−133359.
13. Bhanumathi, V.; Sangeetha, R. CNN based training and classification of MRI brain images. In *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, *Coimbatore*, *India*, *15–16 March 2019*; IEEE: Coimbatore, India, 2019; pp. 129–133. doi:10.1109/ICACCS.2019.8728447
14. Singh, S.; Sapra, P.; Garg, A.; et al. CNN based Covid-aid: Covid 19 detection using chest X-ray. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, *Erode*, *India*, *8–10 April 2021*; IEEE: Erode, India, 2021; pp. 1791–1797. doi:10.1109/ICCMC51019.2021.9418407
15. Garud, S.; Dhage, S. Lung cancer detection using CT images and CNN algorithm. In *2021 International Conference on Advances in Computing*, *Communication*, *and Control (ICAC3)*, *Mumbai*, *India*, *3–4 December 2021*; IEEE: Mumbai, India, 2021; pp. 1–6. doi:10.1109/ICAC353642.2021.9697158
16. Kwaśniewska, A.; Rumiński, J.; Rad, P. Deep features class activation map for thermal face detection and tracking. In *2017 10Th international conference on human system interactions (HSI)*, *Ulsan*, *Korea (South)*, *17-19 July 2017*; IEEE: Ulsan, Korea (South), 2017; pp. 41–47. doi:10.1109/HSI.2017.8004993
17. Xia, Y.F.; Yu, H.; Wang, F.Y. Accurate and robust eye center localization via fully convolutional networks. *IEEE/CAA J. Autom. Sin.*, **2019**, *6*: 1127−1138.
18. Shi, J.; Fan, X.L.; Wu, J.Z.; et al. DeepDiagnosis: DNN-based diagnosis prediction from pediatric big healthcare data. In *2018 Sixth International Conference on Advanced Cloud and Big Data (CBD)*, *Lanzhou*, *China*, *12–15 August 2018*; IEEE: Lanzhou, China,

2018; pp. 287–292. doi:10.1109/CBD.2018.00058

19. Dai, Y.; Gao, Y.F.; Liu, F.Y. TransMed: Transformers advance multi-modal medical image classification. *Diagnostics*, **2021**, *11*: 1384.
20. Seedat, N.; Aharonson, V.; Schlesinger, I. Automated machine vision enabled detection of movement disorders from hand drawn spirals. In *2020 IEEE International Conference on Healthcare Informatics (ICHI)*, *Oldenburg*, *Germany*, *30 November 2020–3 December 2020*; IEEE: Oldenburg, Germany, 2020; pp. 1–5. doi:10.1109/ICHI48887.2020.9374333
21. Lakshmanaprabu, S.K.; Mohanty, S.N.; Rani S, S.; *et al*. Online clinical decision support system using optimal deep neural networks. *Appl. Soft Comput.*, **2019**, *81*: 105487.
22. Kwasniewska, A.; Ruminski, J.; Szankin, M. Improving accuracy of contactless respiratory rate estimation by enhancing thermal sequences with deep neural networks. *Appl. Sci.*, **2019**, *9*: 4405.
23. Kamala, Y.L.; Rao, K.V.S.N.R.; Josephine, B.M. Comparison and evaluation of studies on precision medicine using AI. In *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, *Erode*, *India*, *7–9 April 2022*; IEEE: Erode, India, 2022; pp. 330–335. doi:10.1109/ICSCDS53736.2022.9760969
24. Bohr, A.; Memarzadeh, K. The rise of artificial intelligence in healthcare applications. In *Artificial Intelligence in Healthcare*; Bohr, A.; Memarzadeh, K., Eds.; Elsevier: Amsterdam, 2020; pp. 25–60. doi:10.1016/B978-0-12-818438-7.00002-2
25. Szankin, M.; Kwasniewska, A.; Sirlapu, T.; et al. Long distance vital signs monitoring with person identification for smart home solutions. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, *Honolulu*, *USA*, *18–21 July 2018*; IEEE: Honolulu, USA, 2018; pp. 1558–1561. doi:10.1109/EMBC.2018.8512509
26. Norori, N.; Hu, Q.Y.; Aellen, F.M.; *et al*. Addressing bias in big data and AI for health care: A call for open science. *Patterns*, **2021**, *2*: 100347.
27. Ntoutsi, E.; Fafalios, P.; Gadiraju, U.; *et al*. Bias in data ‐ driven artificial intelligence systems—An introductory survey. *WIREs Data Min. Knowl. Discovery*, **2020**, *10*: e1356.
28. Fletcher, R.R.; Nakeshimana, A.; Olubeko, O. Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. *Front. Artif. Intell.*, **2021**, *3*: 561802.
29. Mehrabi, N.; Morstatter, F.; Saxena, N.; *et al*. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, **2022**, *54*: 50.
30. Chakraborty, J.; Majumder, S.; Menzies, T. Bias in machine learning software: Why? How? What to do? In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, *Athens Greece*, *23–28 August 2021*; ACM: Athens Greece, 2021; pp. 429–440. doi:10.1145/3468264.3468537
31. Wang, X.S.; Peng, Y.F.; Lu, L.; et al. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, *Honolulu*, *USA*, *21–26 July 2017*; IEEE: Honolulu, USA, 2017; pp. 3462–3471. doi:10.1109/CVPR.2017.369
32. Kwaśniewska, A.; Rumiński, J.; Wtorek, J. The motion influence on respiration rate estimation from low-resolution thermal sequences during attention focusing tasks. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, *Jeju*, *Korea (South)*, *11– 15 July 2017*; IEEE: Jeju, Korea (South), 2017; pp. 1421 –1424. doi:10.1109/EMBC.2017.8037100
33. Dhannawat, R.; Patankar, A.B. Improvement to blind image denoising by using local pixel grouping with SVD. *Procedia Comput. Sci.*, **2016**, *79*: 314−320.
34. Ilesanmi, A.E.; Ilesanmi, T.O. Methods for image denoising using convolutional neural network: A review. *Complex Intell. Syst.*, **2021**, *7*: 2179−2198.
35. Bramich, D.M. A new algorithm for difference image analysis. *Mon. Not. Roy. Astrono. Soc. Lett.*, **2008**, *386*: L77−L81.
36. Sanders, J.G.; Jenkins, R. Individual differences in hyper-realistic mask detection. *Cognit. Res. Princ. Implic.*, **2018**, *3*: 24.
37. Bellamy, R.K.E.; Dey, K.; Hind, M.; *et al*. AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM J. Res. Dev.*, **2019**, *63*: 4.
38. Kim, K.T.; Lee, D.R. Probabilistic parameter estimation using a Gaussian mixture density network: Application to X-ray reflectivity data curve fitting. *J. Appl. Cryst.*, **2021**, *54*: 1572−1579.
39. Tan, M.X.; Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, *Long Beach*, *California*, *9–15 June 2019*; PMLR: Long Beach, California, 2019; pp. 6105–6114.

Citation: Alicja, K.; Maciej, S. Can AI See Bias in X-ray Images?. *International Journal of Network Dynamics and Intelligence*. https://doi.org/10.53941/ijndi0101005