Refining Tuberculosis Detection in CXR Imaging: Addressing Bias in Deep Neural Networks via Interpretability

1st Özgür Acar Güler
Department of Informatics
University of Zurich
Zurich, Switzerland
oezgueracar.gueler@gmail.ch

2nd Manuel Günther
Department of Informatics
University of Zurich
Zurich, Switzerland
manuel.guenther@uzh.ch

3rd André Anjos
Idiap Research Institute
Martigny, Switzerland
andre.anjos@idiap.ch

arXiv:2407.14064v1 [cs.CV] 19 Jul 2024

Abstract—**Automatic classification of active tuberculosis from chest X-ray images has the potential to save lives, especially in low- and mid-income countries where skilled human experts can be scarce. Given the lack of available labeled data to train such systems and the unbalanced nature of publicly available datasets, we argue that the reliability of deep learning models is limited, even if they can be shown to obtain perfect classification accuracy on the test data. One way of evaluating the reliability of such systems is to ensure that models use the same regions of input images for predictions as medical experts would. In this paper, we show that pre-training a deep neural network on a large-scale proxy task, as well as using mixed objective optimization network (MOON), a technique to balance different classes during pre-training and fine-tuning, can improve the alignment of decision foundations between models and experts, as compared to a model directly trained on the target dataset. At the same time, these approaches keep perfect classification accuracy according to the area under the receiver operating characteristic curve (AUROC) on the test set, and improve generalization on an independent, unseen dataset. For the purpose of reproducibility, our source code is made available online.[1]**

Index Terms—**Computer-aided diagnosis, Tuberculosis, Interpretability, Saliency mapping, Label balancing, Bias**

I. INTRODUCTION

Chest radiography (CXR) has been a pivotal tool in diagnosing and managing tuberculosis (TB) for over a century, while its effectiveness largely depends on the availability and expertise of human interpreters. Recent advancements in artificial intelligence, particularly in computer-aided detection (CAD) software, have revolutionized CXR analysis for TB detection. These software applications not only automate CXR interpretation for TB but also identify other non-TB radiographic abnormalities [1]. Recognizing the potential of CAD, the World Health Organization (WHO) in 2021 endorsed a conditional recommendation for using CAD solutions as a substitute for human readers in TB screening and triage for individuals aged 15 and above [2]. This endorsement underscores the growing importance of CAD in enhancing the accuracy and efficiency of TB diagnosis, especially in contexts where skilled human readers are scarce [3].

However, challenges arise in the deployment of CAD software in medical diagnostics. The core algorithms of these CAD products are often deep neural networks (DNN), which are perceived as *black boxes* since their decision-making processes are not transparent or not easily understandable. This opacity in algorithmic functioning makes it difficult to assess and validate existing solutions and limits the trust in algorithmic forecasts [4]. Saliency mapping techniques can play a key role in addressing the black-box nature of deep learning-based CAD software for CXR interpretation. By visually highlighting detected radiological findings, these techniques can assist radiologists in providing more accurate diagnoses [4], [5]. Commercially available CAD software applications for TB interpretation typically include saliency map visualizations, as such approaches have been shown to improve diagnostic capabilities, particularly in complex cases with multiple abnormalities [6], [7].

The lack of publicly available and well-curated datasets for training DNNs for CXR interpretation in TB applications hinders the ability to evaluate and develop

---

[1]https://biosignal.pages.idiap.ch/software/paper-euvip24-refine-cad-tb/

1

models that can be deployed responsibly. Furthermore, the small amount of specifically annotated data renders it difficult to analyze if existing models, which claim high accuracy, are actually exploiting spurious biases in the data instead of more desirable and meaningful clinically explainable factors.

In this context, we pose the question if naively training a classifier on the largest publicly available dataset for TB classification, TBX11K [8], containing more than 11'000 images, is sufficient to develop a model that produces reasonable interpretation traits that resemble human judgment. We further explore the best approaches to achieve this using only publicly available data. Our main contribution is methodological: training (or pre-training) of DNN models while compensating for data imbalances lead to improved interpretation alignment with humans, while retaining comparable generalization capabilities on test data. Concretely, we show that: i) Naively training a DNN leads to a highly accurate model that latches to undesirable image biases; ii) Pre-training the weights of a classifier with a related proxy multi-objective classification task containing far more data reduces the interpretation biases and, finally, iii) Balancing classes during pre-training and fine-tuning can further align the final classifier to human expectations.

II. RELATED WORK

Before the Coronavirus Disease 2019 (COVID-19) pandemic, TB was responsible for more deaths per year than any other infectious disease [9]. Albeit the alarming scenario, the number of public datasets that can be used for developing and validating models for CXR interpretation in this context remains relatively low. Essentially, there exist four public datasets for the development and evaluation of TB-related CAD tools, three of which are relatively small, containing only a few hundred images each and are, therefore, of limited application to the development of DNN models. The more recent TBX11K dataset [8] contains 11'200 CXR images with corresponding bounding box annotations for image areas that corroborate the attributed image classes, which can be used to evaluate both classification and radiological sign localization. The authors of the dataset propose the development of two separate models for each task, with relatively high performance. Compared to that work, our proposed workflow only relies on a single model trained for classification, which can adjacently explain reasoning through saliency maps.

Research in automatic detection of Pulmonary TB from CXR images has progressed, with DNNs such as convolutional neural networks (CNNs) leading current advancements, demonstrating high accuracy and strong correlation with ground-truth labels, typically derived from skin or sputum tests [10]. Despite all improvements, realistic scenarios where CAD from CXR imaging for TB could be useful are rather different from lab conditions [11]. In high-burden countries, for example, TB must be screened against the general population, with individuals potentially presenting various other (pulmonary) diseases. Patients with positive skin or sputum tests, in different stages of the disease, or due to other co-morbidities (*e. g.* HIV-positive), may not present classical TB symptoms clearly visible on CXR images. Therefore, from a (human) interpretation point of view, one may argue that CAD for TB, based exclusively on CXR images, should be limited to identifying factual and reportable radiological signs that can be detected on the original CXR images. The availability of reproducible baselines also remains low, as typical datasets used to produce published results are not released due to privacy restrictions that are especially prominent in medical data.

Saliency mapping techniques serve as post-hoc interpretative tools that elucidate the decision-making process of a model, thereby enhancing user trust and understanding. While these maps do not reveal the intricate internal workings of a model, they contribute to its interpretability by highlighting key areas in the input data that influence the model's outcomes. This feature is particularly beneficial in applications like medical diagnosis where it helps professionals by revealing biases and crucial decision-influencing factors. However, the degree of usefulness varies across different saliency mapping techniques, necessitating a detailed examination of selected methods to understand their specific contributions and limitations [12].

Given the shortage of public data for training DNNs to perform classification of CXRs for TB, it should be possible to re-use similar proxy tasks with far more data, to pre-train these models before fine-tuning on existing, smaller TB datasets. For chest X-ray (CXR) data, the NIH-CXR14 dataset [13] can be useful as such a proxy dataset since it represents the largest publicly available repository to date, encompassing more than 112,000 images containing up to 14 different labels extracted from radiology reports. These labels are highly unbalanced and do not align with signs that are typically related to pulmonary TB onset, but pre-training on this task can still help our final task. Fine-tuning this model involves adjusting the weights of the pre-trained network to make it more suited for our specific task. This can be

TABLE I: MODELS. This table lists training details of our models. Pre-training was performed on the *proxy* dataset NIH-CXR14, while fine-tuning was performed on the *target* dataset TBX11K.

Model	M_U	M_B	$M_{U,U}$	$M_{U,B}$	$M_{B,B}$
Pre-training	—	—	unbalanced	unbalanced	balanced
Fine-tuning	unbalanced	balanced	unbalanced	balanced	balanced

done by retraining the entire network or just the final layers.

Rudd *et al.* [14] addressed multi-label imbalance in DNNs and introduced the mixed objective optimization network (MOON). MOON re-weights losses for highly imbalanced datasets, effectively balancing training across classes for each objective individually.[2] Compared to other resampling strategies like over- or undersampling, this method refrains from introducing additional biases by retaining the inherent correlations between labels and avoiding information loss in case of undersampling, while also being more efficient during training than oversampling. Weights w_i^+ and w_i^- are adapted based on the counts of samples in the positive S_i^+ and negative class S_i^- for each binary objective i:

$$w_i^+ = \begin{cases} 1 & \text{if } S_i^- > S_i^+ \\ \frac{S_i^-}{S_i^+} & \text{otherwise} \end{cases} \quad w_i^- = \begin{cases} 1 & \text{if } S_i^+ > S_i^- \\ \frac{S_i^+}{S_i^-} & \text{otherwise} \end{cases} \quad (1)$$

In [14], these weights were used as probabilities to sample the objectives for which loss values are back-propagated, but they can also improve interpretability when used as loss weights [15] in binary cross-entropy:

$$\mathcal{J} = -\sum_{i=1}^{M} w_i^{t_i} \left[t_i \log f_i(x) + (1-t_i) \log(1-f_i(x)) \right] \quad (2)$$

where i iterates all objectives, t_i is the binary ground-truth label, and $f_i(x)$ is the prediction of objective i in sample x.

From this setup, we pose the question if it is beneficial to re-use the partially unrelated NIH-CXR14 dataset to pre-train a DNN system to perform accurate and interpretable TB readout from CXR. We hypothesize that: i) Pre-training with NIH-CXR14 helps improve generalization compared to only using the largest TB dataset available (*i.e.* TBX11K); ii) Balancing classes during training will further remove spurious biases and improve human interpretability of saliency maps produced by DNNs.

---

[2]Objectives in [14] were to classify the presence of different attributes in facial images.

III. DATA AND METHODS

In this work, we make use of 3 publicly available datasets: NIH-CXR14 [16], TBX11K [8] and Shenzhen [17]. Our *proxy* dataset NIH-CXR14 contains 112'120 images of 30'805 unique patients provided as 8-bit grayscale images with a resolution of 1024×1024. The images of the dataset are split into 3 patient-disjoint partitions for training (98'637), validation (6'350) and testing (4'054). The 14 image-level labels in this dataset are: atelectasis, consolidation, infiltration, pneumothorax, edema, emphysema, fibrosis, effusion, pneumonia, pleural thickening, cardiomegaly, nodule, mass and hernia.

The *target* dataset TBX11K is used to either directly train a model from scratch or fine-tune an NIH-CXR14 pre-trained variant. This dataset consists of 11'702 24-bit RGB recordings of CXRs of resolution 512×512. Each sample is from a unique individual. There are 4 types of labels for existing samples, stratifying the images into: healthy, sick (non-TB), active TB, and latent TB. Here, we only consider healthy (3'800) and active TB cases (630), splitting existing data into 3 partitions for training (2'767), validation (706) and testing (957), while preserving healthy/active TB stratification. Beyond labels, samples in the TBX11K dataset also contain bounding boxes in the original CXR image where radiological signs corroborate label assignment for TB cases (and only in those cases). The precise type of radiological sign is not further annotated, except for its relationship to active TB or latent TB infection (*e. g.* scars from previous active TB sickness).

Our *external* dataset is Shenzhen and consists of 662 8-bit RGB CXR recordings of variable resolution (up to 3000×3000). The samples are classified as either healthy (326) or active TB (336). This dataset is not partitioned into subsets and is only used for evaluating the generalization of our DNN models.

In this work, we use a standard DenseNet-121 model architecture [18], for its excellent accuracy in image-related problems and availability in various deep learning software libraries. This model architecture has also shown to produce more accurate visualizations in con-

TABLE II: RESULTS. AUROC of all models on the test set of the *target* TBX11K dataset, and on the entire *external* Shenzhen dataset. Additionally, the medians of Proportional Energy on the *target* test set are shown, after applying Grad-CAM, HiResCAM, and Score-CAM to the last convolutional layer of all 5 DenseNet-121 models.

Metric	AUROC		Proportional Energy		
	Target	External	Grad-CAM	HiResCAM	Score-CAM
M_U	1.00	0.79	0.077	0.077	0.088
M_B	1.00	0.73	*0.060*	*0.058*	*0.007*
$M_{U,U}$	1.00	0.86	0.109	0.118	0.155
$M_{U,B}$	1.00	0.88	0.234	0.237	0.182
$M_{B,B}$	1.00	0.88	**0.295**	**0.301**	**0.326**

juction with current saliency mapping techniques in TB detection compared to the architecture proposed by Pasa *et al.* [19]. Before training, each network is pre-initialized with readily available ImageNet weights. Training is done through a stock Adam optimizer with default parameters, guided by the weighted binary-cross entropy loss (2) in both the binary *target* task and the multi-objective binary *proxy* task, either using *unbalanced training* with $\forall i : w_i^+ = w_i^- = 1$ or *balanced training* using class weights according to (1). Data augmentations included horizontal flips with 50% probability during *proxy* pre-training and elastic deformation [20] with 80% probability for the binary *target* task. Fine-tuning followed the same training technique with the same learning rate (1×10^{-4}) as for the NIH-CXR14 pre-training. Sample importance balancing was performed using the loss-weighting technique (2). In total, five models were trained, as listed in Tab. I.

To evaluate classification performance of TB vs. healthy CXR images, we report the area under the receiver operating characteristic curve (AUROC) on the TBX11K test set. To assess generalization, we report AUROC on the *external* Shenzhen dataset. The level of interpretability of saliency maps for a given DNN model is measured through the median Proportional Energy [12] of the test set, where we utilize the ground-truth bounding boxes provided in our *target* dataset. Saliency maps are produced using Grad-CAM [21] as a baseline for comparison with HiResCAM [22] and Score-CAM [12], which from previous experience produce maps that best correlate with human interpretability [23].

IV. RESULTS AND DISCUSSION

We have trained all five models, preserving the one reaching the lowest loss on the validation set in the respective setups. We then observed that all models achieved an AUROC of 1.0 (perfect scoring) on the *target* test set (TBX11K). Subsequently, we cross-evaluated all models against the full *external* dataset (Shenzhen) to explore their generalization capabilities. As shown in Tab. II, the unbalanced model M_U reaches an AUROC of 0.79 when cross-evaluated, whereas the balanced model M_B AUROC for a cross-dataset evaluation slightly drops to 0.73. All the other models achieved an AUROC of at least 0.86, demonstrating higher generalization once exposed to a larger dataset, with or without balancing.

We also evaluated each model's (human-like) interpretability using their median Proportional Energy over all test samples with active TB of the *target* dataset (for obvious reasons there exist no bounding-box labels for the healthy cases), listed in Tab. II. As can be observed, interpretability increases as more data and balancing are introduced to the models' training process, from below 0.1 for M_U to about 0.3 for all the saliency mapping techniques explored when balancing is applied to both *proxy* training and *target* fine-tuning. Remarkably, compared to model M_U, the addition of balancing to the simple binary classifier without adding more data (M_B) leads to a lower interpretability. Grad-CAM and HiResCAM perform similarly on all models, whereas Score-CAM shows more extreme penalization on model M_B with a median Proportional Energy of 0.007, and slightly better improvement on model $M_{B,B}$ with 0.326. We note that while a perfectly aligned model would have its median equal to 1.0, the use of bounding boxes to represent natural radiological signs compromises this metric as such signs are rarely rectangular in nature.

Fig. 1(a) presents five HiResCAMs visualizations with the lowest Proportional Energy scores obtained by the unbalanced model M_U. These samples, all correctly identified as active TB, exhibit saliency maps with notable focuses on the center of the CXRs and outside the lung areas, *e. g.*, in the armpit or the bottom of the
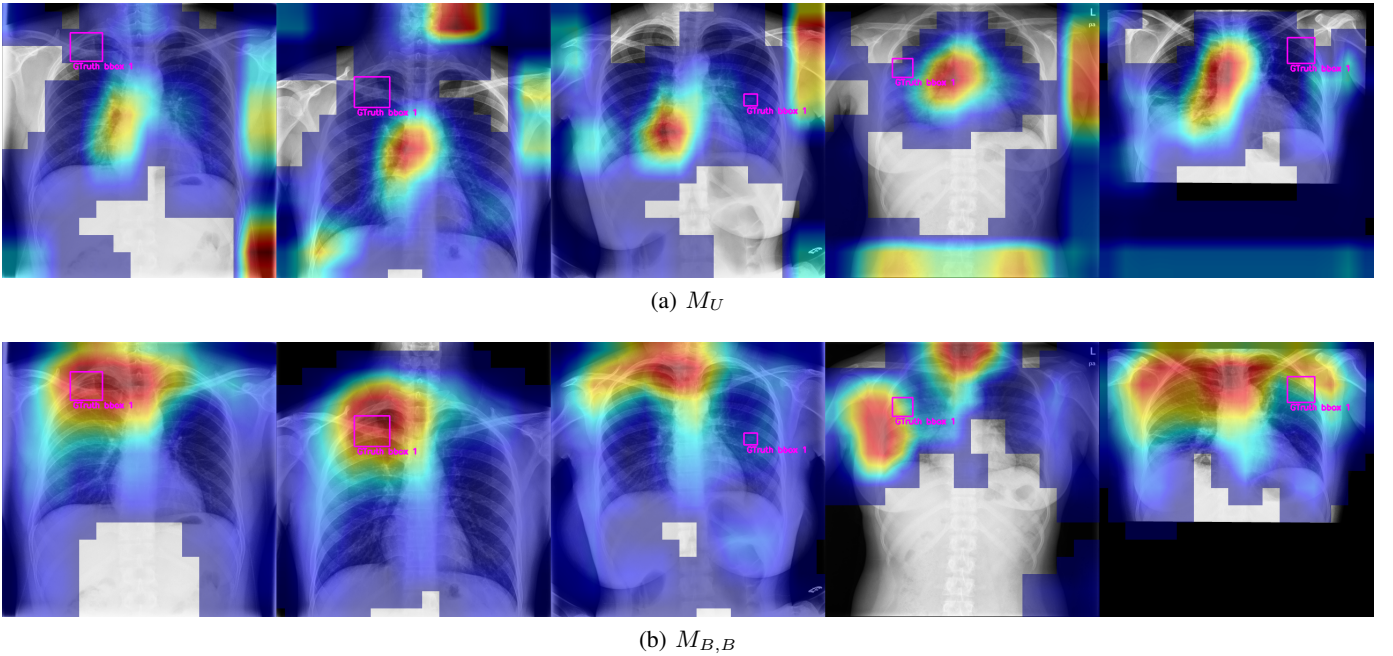
(a) M_U

(b) $M_{B,B}$

Fig. 1: SALIENCY MAPS. In this figure, we show (a) Saliency maps for the unbalanced model M_U featuring the five cases from the TBX11K test set with the lowest Proportional Energy scores; (b) respective predictions of our best balanced model $M_{B,B}$. Human-annotated ground-truth regions including radiological signs are indicated by bright magenta bounding boxes. The heatmaps (ranging from red to blue) indicate the contribution of different regions to the models' decision-making, with non-colored areas having no significant contribution.

image where typically diagnostic information is absent. In Fig. 1(b), the same five samples are visualized with the balanced model $M_{B,B}$. In the left two images, the saliency maps now correspond to human (ground-truth) annotations, and results for the other three cases at least exhibit a closer alignment with the patients' bodies. Hence, the $M_{B,B}$ model not only maintains accurate classification but also demonstrates a marked refinement in CAM localization.

V. CONCLUSION

Training deep neural networks (DNNs) in a naive manner often results in models with undesirable biases. Contrary to previous experiences with large-scale data [15], we found that training a balanced classifier on target data decreased interpretability. However, pre-training the classifier on a larger, related multi-objective classification task significantly mitigated this issue and improved model transferability to an *external* dataset. Additionally, balancing classes during pre-training and fine-tuning enhanced alignment with human expectations without compromising utility on the *target* and *external* datasets.

Our study has notable findings, but also some limitations. While our models demonstrate utility on a cross-dataset evaluation, this did not fully confirm their generalization, suggesting the presence of residual biases. Specifically, models pre-trained on the NIH-CXR14 dataset exhibit better generalization capabilities. Despite observed improvements in Proportional Energy, it is important to acknowledge that the values remain significantly lower than the ideal score of 1. This discrepancy highlights the necessity for further testing, particularly with more precisely annotated radiological signs, and more diverse datasets, to better understand the limits of this metric. Moreover, while our study demonstrates potential, it does not fully address how these methods could be integrated into clinical settings and how they would perform under these conditions. Finally, the human interpretability testing methodology proposed in this study shows promise for debiasing other models while preserving their utility.

REFERENCES

[1] Z. Z. Qin, S. Ahmed, M. S. Sarker, K. Paul, A. S. S. Adel, T. Naheyan, R. Barrett, S. Banu, and J. Creswell, "Tuberculosis detection from chest X-rays for triaging in a high tuberculosis-burden setting: An evaluation of five artificial intelligence

algorithms," *The Lancet Digital Health*, vol. 3, no. 9, pp. e543–e554, 2021.

[2] World Health Organization, *WHO operational handbook on tuberculosis: Module 2: Screening: Systematic screening for tuberculosis disease*. Geneva: World Health Organization, 2021. [Online]. Available: https://apps.who.int/iris/handle/10665/340256

[3] C. Geric, Z. Z. Qin, C. M. Denkinger, S. V. Kik, B. Marais, A. Anjos, P.-M. David, F. A. Khan, and A. Trajman, "The rise of artificial intelligence reading of chest X-rays for enhanced TB diagnosis and elimination," *International Journal of Tuberculosis and Lung Diseases*, vol. 27, no. 5, pp. 367–372, 2023.

[4] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining Explanations: An Overview of Interpretability of Machine Learning," in *International Conference on Data Science and Advanced Analytics (DSAA)*, 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8631448

[5] Z. C. Lipton, "The mythos of model interpretability," *Communications of the ACM*, vol. 61, no. 10, p. 36–43, 2018.

[6] H. L. Kundel, C. F. Nodine, and E. A. Krupinski, "Computer-displayed eye position as a visual aid to pulmonary nodule interpretation," *Investigative Radiology*, vol. 25, no. 8, pp. 890–896, 1990.

[7] P. Rajpurkar, C. O'Connell, A. Schechter, N. Asnani, J. Li, A. Kiani, R. L. Ball, M. Mendelson, G. Maartens, D. J. van Hoving, R. Griesel, A. Y. Ng, T. H. Boyles, and M. P. Lungren, "CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest X-Rays in patients with HIV," *npj Digital Medicine*, vol. 3, no. 1, p. 115, 2020.

[8] Y. Liu, Y.-H. Wu, Y. Ban, H. Wang, and M.-M. Cheng, "Rethinking computer-aided tuberculosis diagnosis," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2643–2652.

[9] World Health Organization, "Global Tuberculosis Report 2023," 2023. [Online]. Available: https://www.who.int/teams/global-tuberculosis-programme/tb-reports

[10] Y. Liu, Y.-H. Wu, S.-C. Zhang, L. Liu, M. Wu, and M.-M. Cheng, "Revisiting computer-aided tuberculosis diagnosis," *arXiv preprint arXiv:2307.02848*, 2023.

[11] World Health Organization, *Chest radiography in tuberculosis detection: summary of current WHO recommendations and guidance on programmatic approaches*. World Health Organization, 2016. [Online]. Available: https://apps.who.int/iris/handle/10665/252424

[12] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-CAM: Score-weighted visual explanations for convolutional neural networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 111–119.

[13] P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz, B. N. Patel, K. W. Yeom, K. Shpanskaya, F. G. Blankenberg, J. Seekins, T. J. Amrhein, D. A. Mong, S. S. Halabi, E. J. Zucker, A. Y. Ng, and M. P. Lungren, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLOS Medicine*, vol. 15, no. 11, p. e1002686, 2018.

[14] E. M. Rudd, M. Günther, and T. E. Boult, "MOON: A mixed objective optimization network for the recognition of facial attributes," in *European Conference on Computer Vision (ECCV)*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer, 2016, pp. 19–35.

[15] X. Zhang, J. S. Bieri, and M. Günther, "Biased binary attribute classifiers ignore the majority classes," in *Swiss Conference on Data Science (SDS)*, 2024.

[16] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-Ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3462–3471.

[17] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative Imaging in Medicine and Surgery*, vol. 4, no. 6, pp. 475–477, 2014.

[18] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.

[19] F. Pasa, V. Golkov, F. Pfeiffer, D. Cremers, and D. Pfeiffer, "Efficient deep network architectures for fast chest X-Ray tuberculosis screening and visualization," *Scientific Reports*, vol. 9, no. 1, p. 6268, Dec. 2019. [Online]. Available: http://www.nature.com/articles/s41598-019-42557-4

[20] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *International Conference on Document Analysis and Recognition*, vol. 3, 2003, pp. 958–958.

[21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3462–3471.

[22] R. L. Draelos and L. Carin, "Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks," *arXiv preprint arXiv:2011.08891*, 2021.

[23] Ö. A. Güler, "Explaining CNN-based active tuberculosis detection in chest X-rays through saliency mapping techniques," Master's thesis, University of Zurich, 2023.