

ZERO-SHOT PEDIATRIC TUBERCULOSIS DETECTION IN CHEST X-RAYS USING SELF-SUPERVISED LEARNING

Daniel Capellán-Martín^{1,2}, Abhijeet Parida², Juan J. Gómez-Valverde¹, Ramon Sanchez-Jacob³,
Pooneh Roshanitabrizi², Marius G. Lingurar^{2,4}, María J. Ledesma-Carbayo¹, Syed M. Anwar^{2,4}

¹ Biomedical Image Technologies, ETSI Telecomunicación, Universidad Politécnica de Madrid & CIBER-BBN, Madrid, Spain

² Sheikh Zayed Institute for Pediatric Surgical Innovation, Children's National Hospital, Washington, DC, USA

³ Division of Pediatric Radiology, Children's National Hospital, Washington, DC, USA

⁴ Departments of Radiology and Pediatrics, GWU School of Medicine, Washington, DC, USA

ABSTRACT

Tuberculosis (TB) remains a significant global health challenge, with pediatric cases posing a major concern. The World Health Organization (WHO) advocates for chest X-rays (CXRs) for TB screening. However, visual interpretation by radiologists can be subjective, time-consuming and prone to error, especially in pediatric TB. Artificial intelligence (AI)-driven computer-aided detection (CAD) tools, especially those utilizing deep learning, show promise in enhancing lung disease detection. However, challenges include data scarcity and lack of generalizability. In this context, we propose a novel self-supervised paradigm leveraging Vision Transformers (ViT) for improved TB detection in CXR, enabling zero-shot pediatric TB detection. We demonstrate improvements in TB detection performance ($\sim 12.7\%$ and $\sim 13.4\%$ top AUC/AUPR gains in adults and children, respectively) when conducting self-supervised pre-training when compared to fully-supervised (i.e., non pre-trained) ViT models, achieving top performances of 0.959 AUC and 0.962 AUPR in adult TB detection, and 0.697 AUC and 0.607 AUPR in zero-shot pediatric TB detection. As a result, this work demonstrates that self-supervised learning on adult CXRs effectively extends to challenging downstream tasks such as pediatric TB detection, where data are scarce.

Index Terms— Self-supervised, Vision Transformers, Zero-shot Classification, Chest X-Ray, Pediatric, Tuberculosis.

1. INTRODUCTION

Tuberculosis (TB), caused by the bacterium *Mycobacterium Tuberculosis*, remains a major global health problem, especially in developing regions, causing substantial mortality. Historically, TB outranked HIV/AIDS as the deadliest infectious disease until the emergence of COVID-19. However, the pandemic led to a decline in TB notifications, experiencing an increase in deaths due to reduced access to TB diagnosis and treatment [1]. Thus, timely detection and treatment are paramount to enhance patient outcomes and mitigate TB transmission. Moreover, the diagnosis and treatment of pediatric TB pose significant challenges, with a notable gap in the identification and treatment of affected children. Annually, 7.5 million children contract TB and 1.1 million develop the disease, contributing to 12% of the global TB burden [2]. Effective screening is therefore essential to identify potential cases and facilitate referral for testing and preventive treatment among children [3].

The World Health Organization (WHO) advocates the use of chest X-rays (CXR) as a TB screening tool due to its widespread availability, cost-effectiveness, speedy results, and reduced risk of cross-infection [4]. Moreover, in clinical practice, CXRs are the preferred imaging method for pediatric TB screening and diagnosis [5]. However, visual

interpretation by radiologists in pediatric TB can be subjective, time-consuming, and error-prone. The challenges arise from non-specific symptoms and relatively fewer radiological manifestations in children compared to adults, leading to a low inter-radiologist agreement and potentially uncertain diagnoses [6, 7].

AI-driven computer-aided detection (CAD) tools, especially those using deep learning, hold promise for enhancing healthcare outcomes by improving radiologists' interpretation accuracy, reducing errors, and enabling patient-centric care. These tools show potential in addressing the diagnostic expertise gap in pediatric TB and facilitating large-scale TB screening through CXR image analysis, particularly in high TB burden areas [8, 9, 10]. However, challenges include the scarcity of well-annotated medical data, especially in pediatric tasks, and a lack of generalizability across different clinical settings and devices. To overcome data scarcity (both in imaging data and in high-quality ground-truth), supervised and self-supervised pre-training strategies are employed. Self-supervised training allows models to learn meaningful features from unlabeled data, eliminating the need for costly and time-consuming annotations. Foundation models, such as those based on transformer architectures, have demonstrated remarkable performance across a wide range of natural language processing and computer vision tasks. These models, pre-trained on massive datasets, capture a broad understanding of the visual world, and fine-tuning them for specific medical imaging tasks presents an opportunity to leverage their robust generalization capabilities and tailor them for pediatric TB detection in CXRs.

Recently, multiple self-supervised learning strategies have demonstrated outstanding performance in natural image-based computer vision tasks, such as DINO [11], a knowledge distillation method without the need for labeled data, MoCo [12], a momentum-based contrastive unsupervised learning method, and MAE [13], based on masked autoencoders performing reconstruction tasks to learn meaningful representations. However, the application of these methods is still limited in the medical imaging domain, even more notably in the field of TB diagnosis.

In this context, we propose a novel self-supervised paradigm to address TB detection in CXRs for both adult and child patients. The main objective of our work is to propose a new self-supervised strategy with state-of-the-art performance on adult and pediatric TB detection requiring no supervised pre-training and minimal fine-tuning. We outline our contributions as follows:

1. We propose a novel self-supervised pre-training paradigm aimed at addressing pediatric TB detection in chest X-ray imaging using a ViT backbone.
2. We conduct comprehensive domain-specific downstream tasks on CXR images, using multiple self-supervised methods, both on adult and pediatric TB data, from public and private sources, to demonstrate the importance of pre-training on domain data for improved

performance.

3. To the best of our knowledge, we are the first to (i) demonstrate that self-supervised pre-training on adult chest X-rays (CXRs) contributes to learning valuable representations for TB detection in both adult and pediatric CXRs, with the latter being underrepresented in most datasets and more difficult to detect, and (ii) explore zero-shot classification in pediatric TB.

2. RELATED WORK

Several studies have explored deep learning (DL) techniques for TB detection in CXRs. Some approaches include CNNs and transfer learning techniques [14, 15]. Additional methods have explored ResNet-based solutions, emphasizing not only TB detection performance, but also efficiency and low computational requirements [16, 17]). Some studies have also examined DenseNet-121 models for improving diagnosis in HIV-infected TB patients [18], while others have studied self-attention mechanisms [19]. However, current approaches exhibit limited performance when tested on pediatric patients, with the domain of research in this specific age group remaining considerably underexplored. This underscores the urgent need to create AI systems tailored for infants and children. Recently, Palmer et al. [20] reported, for the first-time ever, results of a CAD software (CAD4TB) applied on pediatric CXRs to identify TB. Notably, they reported an AUC of 0.58 with a fully-supervised non-fine-tuned model, and an AUC of 0.72 with a fully-supervised fine-tuned model. Finally, although limited, some recent studies [21, 22, 23, 24, 25] verse about the application of self-supervised techniques to improve the detection of diverse lung diseases on CXR images, moving towards the concept of foundation medical image analysis models.

3. METHODS

3.1. Datasets

During the process of self-supervised pre-training, a large-scale dataset of 357,286 frontal (antero-posterior (AP) and postero-anterior (PA)) CXR images was used, sourced from diverse public CXR datasets:

- CheXpert [26] (191,229 frontal CXR images from 64,740 patients were used) and ChestXray-14 [27] (112,120 frontal CXR images from 30,805 patients were used), which conformed the majority of the pre-training data. Both publicly available datasets are weakly-annotated using natural language processing (NLP) techniques, yielding >90% labeling accuracy. They encompass annotations for the presence of 14 observations as positive, negative, or uncertain. Examples of radiological observations present in both datasets include lung opacity, atelectasis, pneumonia, cardiomegaly and pneumothorax, among others.
- The COVIDx CXR-3 dataset¹ [28, 29]. It is the most extensive and diverse COVID-19 CXR dataset in open access form, comprising 30,386 frontal CXR images taken from 17,026 patients.
- The SPR X-Ray Gender Prediction Challenge², intended to determine gender and age using CXRs. A total of 22,449 images were incorporated from this source.
- Shenzhen Hospital X-ray Set (SZ) [30], Belarus (BL)³, and Montgomery County X-ray Set (MC) [30] datasets, intended for supporting TB research. A total of 1,102 frontal CXR images were incorporated from these three datasets (662, 302 and 138 from SZ, BL and MC, respectively).

¹<https://www.kaggle.com/datasets/andyczhao/covidx-cxr2>

²<https://kaggle.com/competitions/spr-x-ray-gender>

³<https://github.com/frapa/tbcnn>

During the fine-tuning and evaluation phases, we used both adult and pediatric cohorts. For the adult experiments, SZ and MC were used (details specified above). Adult data for fine-tuning were split into training (80%, 640 frontal CXR images) and independent test (20%, 160 frontal CXR images, no patient overlap). Stratification was applied based on cohort, label, sex and age. For pediatrics, we used an in-house collected dataset from Children’s National Hospital (CNH), Washington, DC, USA, featuring 85 patients aged 0-18 years (32.94% 0-3 years, 29.41% 4-12 years, 37.65% 13-18 years). A total of 85 frontal CXR images were used (54 negative, 31 positive), configuring an unseen, out-of-domain (ODD), independent test set to evaluate, in a zero-shot fashion, our method.

MC and SZ standard TB references were either confirmed microbiologically or, when unfeasible, based on consistent clinical symptoms and imaging findings. On the other hand, CNH reference TB standards were confirmed microbiologically, except for one unconfirmed TB case. Among the 85 patients, three had a latent TB infection.

3.2. Architecture and self-supervised strategy

Vision transformers (ViT) [31] have recently shown great success in computer vision tasks, and their use in the medical domain is growing at an unprecedented rate. The ViT processes input images by tokenizing them into a sequence of patches. Each patch, obtained from the image’s height, width, and channels ($H \times W \times C$), is flattened into n two-dimensional patches of size $p_1 \times p_2 \times C$ pixels. A linear layer or a convolutional layer with specific parameters projects these patches to D hidden dimensions. To preserve spatial relationships, fixed or learnable position embeddings are added to patch embeddings. The Transformer encoder comprises L consecutive Multi-head Self-Attention (MSA) and Multi-Layer Perceptron (MLP) blocks. The MSA layer employs a trainable triplet of (query (Q), key (K), value (V)), producing output sequences that are linearly projected to $n \times D$ dimensions. The self-attention mechanism is based on the equation 1, where d_k represents the dimensionality of the key vectors in the attention mechanism. For classification, a trainable class token is added to the patch token sequence, processed through the Transformer encoder, and connected to a classification head. However, in unsupervised scenarios, such as self-supervised (SSL) learning, no classification token is required.

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (1)$$

For all the experiments, we used a vision transformer (ViT) [31] (ViT-S/16) as the backbone. During pre-training, we incorporated diverse SSL methods, including MAE [13], MOCO-v3 [12, 32] and DINO [11]. Also, fully-supervised, i.e., non pre-trained, ViT models were also used for baseline comparison. In contrastive SSL methods, instead of using classification tokens, we use a contrastive token and compute contrastive losses. On the other hand, in image reconstruction methods, we usually partially mask patches of the input (i.e., data tokens) and perform image reconstruction tasks. After pre-training, we conducted both linear probing and fine-tuning. During linear probing, the pre-trained model’s weights are frozen, except for the weights of the classification head, which is then trained or ”probed” to adapt the model to the specifics of the new task. During fine-tuning, instead, the entire pre-trained model layers are unfrozen, and the model is trained on the specifics of the new task. Fine-tuning is more intensive computationally compared to linear probing, but allows the model to adjust more parameters to better fit the new task.

Consequently, the self-supervised learning strategy proposed in this work is illustrated in Figure 1. It encompasses multiple steps described in this work, spanning from pre-training to TB prediction.

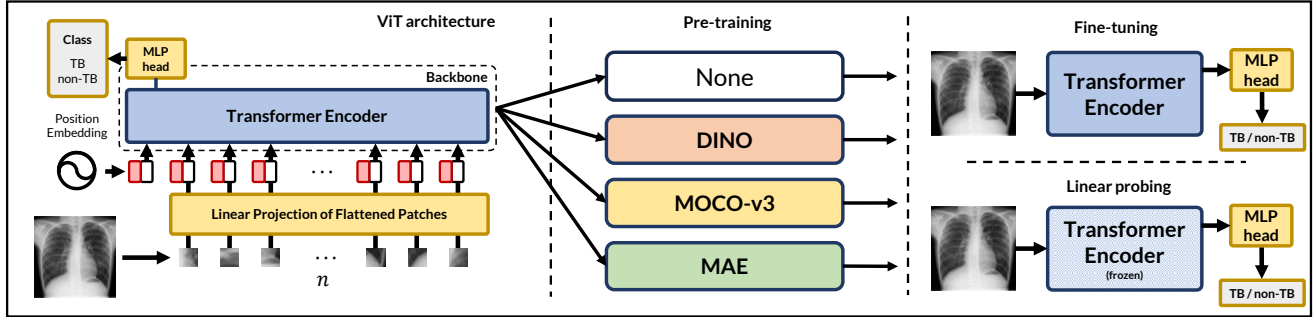


Fig. 1: We propose a novel self-supervised strategy leveraging ViTs for improved TB detection in CXR, enabling for zero-shot pediatric TB detection.

4. RESULTS

4.1. Experimental Setup

We conducted several downstream classification experiments to evaluate our approach on both adult (seen distribution) and pediatric (unseen, out-of-domain (OOD) and shifted distribution) TB datasets, employing diverse self-supervised learning methods (see Section 3.2). We executed two downstream binary (non-TB vs. TB) classification tasks. One task utilized an independent subset of adult CXR data (MC & SZ), with no distribution shift, i.e., similar data distribution as the one used for fine-tuning. The second task, following a zero-shot approach, involved an independent set of pediatric CXR data (CNH). This dataset presented distribution shifts due to distinct acquisition devices and setups, along with differences in age with respect to the fine-tuning datasets.

During the pre-training phase, regardless of the self-supervised approach, we maintained the following set-up: batch size of 64, 200 epochs, cosine learning rate (LR) scheduler ($LR_{min} = 1e^{-6}$, $LR_{initial} = 5e^{-4}$, 10 warm-up epochs), cosine weight decay scheduler (weight decay of 0.04) and AdamW optimizer [33]. Subsequently, we fine-tuned the pre-trained models (also non-pre-trained, for the sake of comparison) using a batch size of 48, 200 epochs and the same LR scheduler settings as in the pre-training. Pre-training was performed across multiple decentralized workstations utilizing multi-GPU settings. The fine-tuning step, however, was executed on the Magerit-3 supercomputer from Universidad Politécnic de Madrid, Spain, dedicating 1 NVIDIA V100/A100 GPU per experiment.

In all experiments, we employed a vision transformer [31] (ViT-S/16) as the backbone, and an image size of 256×256 . During evaluation, confidence intervals (CI) were calculated using DeLong’s approach [34, 35] for area under the receiver operating characteristic (ROC) curve (AUC) metrics and bootstrap method [36] (2000 repetitions) for accuracy (ACC) and area under the precision-recall curve (AUPR).

4.2. TB classification results

4.2.1. Evaluation on adult data

We conducted linear probing and fine-tuning experiments on adult TB data, i.e., using the independent test subsets of the Shenzhen Hospital X-ray Set (SZ) and the Montgomery County X-ray Set (MC). The results, as reported in Table 1, demonstrated that MAE achieved the best performance, achieving the highest AUC and AUPR scores: 0.926 and 0.932, respectively. This occurred during linear probing on the SZ & MC training subsets. DINO achieved the second-best AUPR and AUC scores and the best ACC and sensitivity (SN) values. On the other hand, when conducting fine-tuning on the same training subset, MOCO-v3 exhibited the highest AUC score (0.959 AUC), and DINO exhibited the highest AUPR score (0.962 AUPR). DINO also scored the highest ACC and

SN values. We can also notice a top performance gain of $\sim 12.7\%$ and $\sim 11.3\%$ in AUC and AUPR, respectively, for linear probing, and $\sim 10.7\%$ and $\sim 11.3\%$ in AUC and AUPR, respectively, for fine-tuning, when compared to a fully-supervised, i.e., non pre-trained, model performance. Additionally, the highest AUC and AUPR scores were attained through fine-tuning rather than linear probing. The ROC curves for the fine-tuned models when tested on MC & SZ are displayed in Figure 2a.

4.2.2. Evaluation on unseen pediatric data

With the models obtained in the previous experiments, i.e., those resulting from conducting linear probing and fine-tuning on adults (MC & SZ), we performed zero-shot classification on an unseen, out-of-domain, i.e., distribution-shifted, pediatric TB dataset collected in-house at Children’s National Hospital (CNH). This enables for assessing transferability of local and global features and patterns learnt from CXRs during pre-training and transfer learning. The results are reported in Table 2. In this other scenario, when conducting linear probing on adults (MC & SZ) and zero-shot testing on children, DINO demonstrated a superior performance (0.622 AUC, 0.560 AUPR) when compared to alternative self-supervised methods and a fully-supervised approach. On the other hand, when instead conducting fine-tuning on adults (MC & SZ) and zero-shot testing on children, MAE achieved a higher AUC score (0.697 AUC), and DINO a higher AUPR score (0.607), when compared to alternative self-supervised methods and a fully-supervised model. The highest ACC, SN and specificity (SP) values were achieved with MOCO-v3. Additionally, we were able to achieve a top performance improvement of $\sim 10\%$ and $\sim 13.4\%$ in AUC and AUPR, respectively, for linear probing, and $\sim 11.5\%$ and $\sim 12.4\%$ in AUC and AUPR, respectively, for fine-tuning, when compared to a fully-supervised model. As in the previous experiments, the highest AUC and AUPR scores were obtained through fine-tuning rather than linear probing. As a result, the ROC curves for the fine-tuned models during zero-shot pediatric TB testing are depicted in Figure 2b.

5. DISCUSSION

The conclusions drawn from the downstream TB classification experiments reveal that self-supervised pre-training on a ViT model leads to superior TB detection performance compared to fully-supervised alternatives. Depending on our goals, a range of self-supervised methods can be employed to enhance TB detection. Method comparisons can be conducted focusing on diverse metrics, although caution should be exercised when comparing methods using metrics such as accuracy due to potential bias from class imbalance. Fine-tuning/linear probing on children may increase SN and SP values, requiring more data. SN and SP values may vary with different prediction thresholds, making AUC and AUPR better for unbiased model performance assessment and results comparison, with AUPR being particularly advantageous

SSL Technique	Linear Probing					Fine-tuning				
	ACC (95% CI)	AUPR (95% CI)	AUC (95% CI)	TPR (SN)	TNR (SP)	ACC (95% CI)	AUPR (95% CI)	AUC (95% CI)	TPR (SN)	TNR (SP)
No Pretrain	.744 (.675, .812)	.819 (.739, .892)	.799 (.729, .869)	.620	.864	.794 (.725, .856)	.849 (.762, .926)	.852 (.792, .913)	.696	.889
MOCO-v3	.731 (.656, .8)	.823 (.734, .895)	.81 (.743, .877)	.671	.790	.881 (.825, .931)	.96 (.928, .984)	.959 (.933, .985)	.823	.938
DINO	.862 (.806, .912)	<u>.925 (.875, .962)</u>	<u>.919 (.878, .961)</u>	.848	<u>.877</u>	.9 (.85, .944)	.962 (.931, .986)	<u>.955 (.923, .987)</u>	.873	<u>.926</u>
MAE	<u>.85 (.794, .906)</u>	.932 (.887, .966)	.926 (.886, .966)	<u>.810</u>	.889	<u>.888 (.838, .931)</u>	<u>.935 (.879, .975)</u>	<u>.941 (.903, .979)</u>	<u>.848</u>	<u>.926</u>

ACC, accuracy; AUPR, Area Under the Precision-Recall (PR) Curve; AUC, Area Under the Receiver Operating Characteristic (ROC) Curve; TPR, True Positive Rate; SN, Sensitivity; TNR, True Negative Rate; SP, Specificity; CI: Confidence Interval.

Table 1: Adult TB classification results obtained from linear probing and fine-tuning on an independent adult test set (seen distribution) from Montgomery (MC) & Shenzhen (SZ) datasets. The highest scores are highlighted in bold, and the second best ones are underlined.

SSL Technique	Linear Probing					Fine-tuning				
	ACC (95% CI)	AUPR (95% CI)	AUC (95% CI)	TPR (SN)	TNR (SP)	ACC (95% CI)	AUPR (95% CI)	AUC (95% CI)	TPR (SN)	TNR (SP)
No Pretrain	.588 (.482, .694)	.426 (.296, .609)	.522 (.384, .659)	.484	.648	.659 (.541, .753)	.483 (.337, .658)	.582 (.451, .712)	.355	.833
MOCO-v3	.529 (.424, .635)	.351 (.255, .516)	.478 (.351, .604)	.258	.685	.718 (.624, .812)	<u>.604 (.439, .767)</u>	.66 (.534, .786)	.355	.926
DINO	.682 (.576, .776)	.56 (.404, .732)	.622 (.491, .753)	.194	.963	.659 (.553, .765)	.607 (.446, .765)	<u>.69 (.567, .812)</u>	.226	<u>.907</u>
MAE	.612 (.506, .718)	.439 (.311, .614)	.55 (.42, .68)	.129	.889	<u>.682 (.576, .788)</u>	.578 (.424, .763)	.697 (.577, .817)	.323	.889

ACC, accuracy; AUPR, Area Under the Precision-Recall (PR) Curve; AUC, Area Under the Receiver Operating Characteristic (ROC) Curve; TPR, True Positive Rate; SN, Sensitivity; TNR, True Negative Rate; SP, Specificity; CI: Confidence Interval.

Table 2: Zero-shot pediatric TB classification results obtained from linear probing and fine-tuning on adults (MC & SZ) and tested on CNH dataset (independent, unseen, ODD, shifted distribution). The highest scores are highlighted in bold, and the second best ones are underlined.

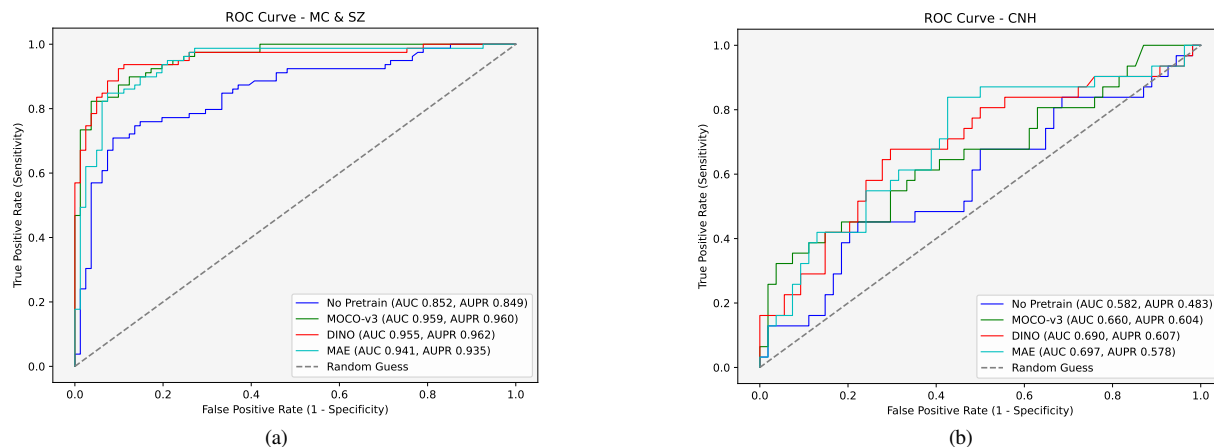


Fig. 2: Receiver operating characteristic (ROC) curves when conducting fine-tuning on adult TB data (MC & SZ training subsets) and: (a) testing on an independent adult TB dataset; (b) performing zero-shot inference on an independent, out-of-domain pediatric TB dataset.

for tasks involving class imbalance. As a result, when focusing on AUPR, DINO achieved better pediatric (out-of-distribution, unseen data) TB detection performance, while MAE scored a top AUC of 0.697, aligning closely with Palmer et al.’s [20] pediatric TB performance (0.72 AUC) and surpassing previous zero-shot results (0.58 AUC) in the same study. These findings demonstrate efficient knowledge transfer to an out-of-distribution scenario, i.e., pediatric TB, where radiological signs and disease detection pose unique challenges in clinical practice.

It’s important to note that using MC and SZ datasets for the fine-tuning and linear probing steps, as part of a seen distribution experiment, does not imply overfitting during inference. These self-supervised tasks induce learning of general features from given CXR distribution, rather than task-specific patterns. In addition, experiments were conducted on an unseen OOD dataset to assess the generalizability of the model to different scenarios.

These findings demonstrate the efficacy of self-supervised pre-training with a ViT backbone in enhancing TB detection performance compared to fully-supervised approaches. This emphasizes the effectiveness of foundational models in addressing challenges like data scarcity,

especially in tasks such as pediatric TB.

6. CONCLUSION

Given the increasing emphasis on CAD systems for TB, particularly in pediatric contexts, we have proposed a novel self-supervised paradigm to tackle complex downstream tasks like TB detection. We demonstrated how the transfer of rich learning representations learned from CXRs enhances TB detection performance, not only in adult TB but also when conducting zero-shot classification on pediatric TB. This research contributes to advancing the application of foundational models in challenging pediatric diagnostic tasks with limited data, aligning with the goal of improving clinical care and patient outcomes in TB diagnosis.

Compliance with ethical standards This research study was conducted using retrospective data from both open access and non-public sources (details in text). Ethical approval was granted from the Institutional Review Board (IRB) from the corresponding institutions.

Acknowledgements This work was supported by H2020-MSCA-RISE-2018 INNOVA4TB (EU) project (ID 823854), ADVANCE-TB Cost

Action (CA21164 (EU)) and by Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación grants PDC2022-133865-I00 and PID2022-141493OB-I0 (MCIN/AEI/10.13039/501100011033). The authors gratefully acknowledge the Universidad Politécnica de Madrid for providing computing resources on Magerit Supercomputer. We acknowledge Taylor Broudy for her participation in clinical data extraction.

7. REFERENCES

- [1] World Health Organization, “Global Tuberculosis Report,” 2023.
- [2] D. Jaganath, J. Beaudry, and N. Salazar-Austin, “Tuberculosis in Children,” *Infect Dis Clin N Am*, vol. 36, no. 1, pp. 49–71, Mar 2022.
- [3] B. Vonasek, T. Ness, Y. Takwoingi, et al., “Screening tests for active pulmonary tuberculosis in children,” *Cochrane Database of Systematic Reviews*, vol. 2021, no. 6, Jul 2021.
- [4] World Health Organization, “Chest Radiography in Tuberculosis Detection: Summary of current WHO recommendations and guidance on programmatic approaches,” 2016.
- [5] B. A. A. Ubaidi, “The Radiological Diagnosis of Pulmonary Tuberculosis (TB) in Primary Care,” *J Fam Med Dis Prev*, vol. 4, no. 1, Mar 2018.
- [6] A. P. Brady, “Error and discrepancy in radiology: inevitable or avoidable?,” *Insights into imaging*, vol. 8, no. 1, pp. 171–182, Feb 2017.
- [7] J. J. Gómez-Valverde, R. Sánchez-Jacob, J. L. Ribó, et al., “Chest x-ray-based telemedicine platform for pediatric tb diagnosis in low resource settings: Development and validation,” *JMIR Preprints*, Nov 2023.
- [8] S. Kulkarni and S. Jha, “Artificial Intelligence, Radiology, and Tuberculosis: A Review,” *Acad Radiol*, vol. 27, no. 1, pp. 71–75, Jan 2020.
- [9] Z. Z. Qin, T. Naheyan, M. Ruhwald, et al., “A new resource on artificial intelligence powered computer automated detection software products for tuberculosis programmes and implementers,” *Tuberculosis*, vol. 127, pp. 102049, Mar 2021.
- [10] M. P. McBee, O. A. Awan, A. T. Colucci, et al., “Deep Learning in Radiology,” *Acad Radiol*, vol. 25, pp. 1472–1480, Nov 2018.
- [11] M. Caron, H. Touvron, I. Misra, et al., “Emerging properties in self-supervised vision transformers,” in *IEEE/CVF ICCV*, 2021, pp. 9650–9660.
- [12] K. He, H. Fan, Y. Wu, et al., “Momentum Contrast for Unsupervised Visual Representation Learning,” *IEEE/CVF CVPR*, 2020.
- [13] K. He, X. Chen, S. Xie, et al., “Masked Autoencoders Are Scalable Vision Learners,” *IEEE/CVF CVPR*, pp. 16000–16009, 2022.
- [14] S. Hwang, H.-E. Kim, J. Jeong, and H.-J. Kim, “A novel approach for tuberculosis screening based on deep convolutional neural networks,” in *SPIE MI*, 2016, vol. 9785, pp. 750–757.
- [15] M. T. Islam, M. A. Aowal, A. T. Minhaz, and K. Ashraf, “Abnormality Detection and Localization in Chest X-Rays using Deep Convolutional Neural Networks,” *arXiv preprint arXiv:1705.09850*, May 2017.
- [16] F. Pasa, V. Golkov, F. Pfeiffer, et al., “Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization,” *Sci Rep*, vol. 9, no. 1, pp. 1–9, Apr 2019.
- [17] D. Capellán-Martín, J. J. Gómez-Valverde, D. Bermejo-Peláez, and M. J. Ledesma-Carbayo, “A Lightweight, Rapid and Efficient Deep Convolutional Network for Chest X-Ray Tuberculosis Detection,” in *IEEE ISBI*, 2023, pp. 1–5.
- [18] P. Rajpurkar, C. O’Connell, A. Schechter, et al., “CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV,” *npj Digit Med*, vol. 3, no. 1, pp. 1–8, Sep 2020.
- [19] A. Wong, J. R. H. Lee, H. Rahmat-Khah, et al., “TB-Net: A Tailored, Self-Attention Deep Convolutional Neural Network Design for Detection of Tuberculosis Cases From Chest X-Ray Images,” *Fr Art Int*, vol. 5, pp. 827299, Apr 2022.
- [20] M. Palmer, J. A. Seddon, M. M. van der Zalm, et al., “Optimising computer aided detection to identify intra-thoracic tuberculosis on chest x-ray in South African children,” *PLOS global public health*, vol. 3, no. 5, pp. e0001799, May 2023.
- [21] H. Sowrirajan, J. Yang, A. Y. Ng, and P. Rajpurkar, “MoCo Pretraining Improves Representation and Transferability of Chest X-ray Models,” in *MIDL*. PMLR, 2021, pp. 728–744.
- [22] M. Gazda, J. Plavka, J. Gazda, and P. Drotár, “Self-supervised deep convolutional neural network for chest x-ray classification,” *IEEE Access*, vol. 9, pp. 151972–151982, 2021.
- [23] S. Park, G. Kim, Y. Oh, et al., “Self-evolving vision transformer for chest X-ray diagnosis through knowledge distillation,” *Nat Commun*, vol. 13, no. 1, pp. 1–11, Jul 2022.
- [24] E. Tiu, E. Talius, P. Patel, et al., “Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning,” *Nat Biomed Eng*, vol. 6, no. 12, pp. 1399–1406, Sep 2022.
- [25] S. M. Anwar, A. Parida, S. Atito, et al., “SPCXR: Self-supervised Pretraining using Chest X-rays Towards a Domain Specific Foundation Model,” *arXiv preprint arXiv: 2211.12944*, 2023.
- [26] J. Irvin, P. Rajpurkar, M. Ko, et al., “Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison,” *AAAI*, vol. 33, no. 01, pp. 590–597, Jul. 2019.
- [27] X. Wang, Y. Peng, L. Lu, et al., “ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases,” in *IEEE/CVF CVPR*, 2017, pp. 2097–2106.
- [28] L. Wang, Z. Q. Lin, and A. Wong, “COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images,” *Sci Rep*, vol. 10, no. 1, pp. 1–12, Nov 2020.
- [29] M. Pavlova, T. Tuinstra, H. Aboutaleb, et al., “COVIDx CXR-3: A Large-Scale, Open-Source Benchmark Dataset of Chest X-ray Images for Computer-Aided COVID-19 Diagnostics,” *arXiv preprint arXiv:2206.03671*, 2022.
- [30] S. Jaeger, S. Candemir, S. Antani, et al., “Two public chest X-ray datasets for computer-aided screening of pulmonary diseases,” *Quant Imaging Med Surg*, vol. 4, no. 6, pp. 475, Dec 2014.
- [31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” *CIBCB*, 2020.
- [32] X. Chen, S. Xie, and K. He, “An Empirical Study of Training Self-Supervised Vision Transformers,” *IEEE/CVF ICCV*, pp. 9620–9629, Apr 2021.
- [33] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” *7th International Conference on Learning Representations, ICLR 2019*, Nov 2017.
- [34] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, “Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach,” *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.
- [35] X. Sun and W. Xu, “Fast implementation of delong’s algorithm for comparing the areas under correlated receiver operating characteristic curves,” *IEEE Signal Processing Letters*, vol. 21, no. 11, pp. 1389–1393, 2014.
- [36] J. Carpenter and J. Bithell, “Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians,” *Statistics in Medicine*, vol. 19, no. 9, pp. 1141–1164, 2000.