Beyond the Prompt: Deploying Medical Foundation Models on Diverse Chest X-ray Populations

Louisa Fay[1,2,3] LFAY@STANFORD.EDU
Jean-Benoit Delbrouck[1] JBDEL@STANFORD.EDU
Thomas Küstner[2] THOMAS.KUESTNER@MED.UNI-TUEBINGEN.DE
Bin Yang[3] BIN.YANG@ISS.UNI-STUTTGART.DE
Noel C. F. Codella[4] NCODELLA@MICROSOFT.COM
Matthew P. Lungren[4] MLUNGREN@MICROSOFT.COM
Curtis P. Langlotz[1] LANGLOTZ@STANFORD.EDU
Sergios Gatidis[1] SGATIDIS@STANFORD.EDU
[1] *Department of Radiology, School of Medicine, Stanford University, CA, USA*
[2] *Medical Image and Data Analysis, University Hospital of Tübingen, Germany*
[3] *Institute of Signal Processing and System Theory, University of Stuttgart, Germany*
[4] *Microsoft Health and Life Sciences, Redmond, WA, USA*

Editors: Under Review for MIDL 2025

Abstract

Foundation models (FMs) have shown impressive performance in medical image analysis tasks, but their deployment in real-world clinical settings, especially across diverse patient populations such as adult and pediatric cases, remains challenging. Key open questions include optimal prompting techniques and strategies for model adaptation or fine-tuning for clinical use. In this study, we evaluated different approaches for deploying FMs in clinical scenarios for diverse patient populations. We use the lightweight, embedding-based vision-language FM *MedImageInsight* to predict pneumonia from chest X-rays, a condition common in both adult and pediatric patients. We observed large variation in model predictive performance depending on the chosen prompt design, highlighting the importance of text prompt design for successful zero-shot (ZS) application. On in-domain datasets, we found performance differences of up to 46% in Matthews correlation coefficient (MCC) and 56% in true positive rates across different text prompts. By introducing text and vision embedding ensembles, we achieved substantial ZS improvements, outperforming training-based methods (fine-tuning, Linear Probe) in low-data scenarios by up to 43% for adults and 35% for pediatric populations (MCC). This ensembling strategy also promotes resource-efficient equitable clinical use by supporting diverse demographic subgroups, achieving MCC improvements of 6% by sex, 17% by age, and 10% by race compared to Linear Probe.

Keywords: multimodal foundation model, bias, zero-shot, pneumonia, ensembles

1. Introduction

Foundation Models (FMs) that have been trained on extensive web-based datasets have demonstrated great promise and remarkable generalizability across a variety of tasks in different domains, including natural language processing, computer vision, and text and image generation (Brown et al., 2020; Radford et al., 2021). Similarly, their medical counterparts, trained on domain-specific datasets such as PubMed, electronic health records, and medical imaging, have shown significant potential to advance healthcare applications (Zhang

et al., 2022; Singhal et al., 2023). However, their reliable implementation in clinical settings without further adjustments remains challenging due to severe consequences of incorrect diagnoses or treatment plans (Huang et al., 2023). While an increasing number of FMs are developed using medical data (Blankemeier et al., 2024; Chen et al., 2024; Zhang et al., 2023), their clinical application often experiences performance drops on out-of-distribution data, such as in new patient populations (e.g., transitioning from adult to pediatric cases) (Huang et al., 2024).

Moreover, since many FMs are trained to derive predictions from vision-language similarities, their effective training-free, zero-shot (ZS) application depends not only on the input image but also on the given text prompt. Determining an optimal text prompt to achieve the best ZS performance in various environments, particularly in new distributions, still poses a major challenge.

A common strategy for applying vision-language FMs in clinical settings relies on adapting and fine-tuning image encoders (Chambon et al., 2022; Hu et al., 2021). However, this approach requires additional diverse and labeled data, which is expensive and difficult to acquire. Furthermore, most FMs are based on large transformer models, which require significant computational resources to fine-tune. As many healthcare facilities lack the necessary infrastructure, these approaches are unsuitable for integration into clinical workflows.

Our study aims to address these challenges by identifying effective strategies for the successful clinical application of FMs, focusing on the state-of-the-art, open-source, lightweight, embedding-based vision-language FM, *MedImageInsight* (Codella et al., 2024) which showed superior performance and suitability across multiple tasks and domains. Since *MedImageInsight* was predominantly trained on adult data, this study examines the prediction of pneumonia from chest X-rays in adult (in-domain) and pediatric (out-of-domain) cases using training-free ZS and training-based approaches.

The key contributions of our work include:

- Identification of effective deployment strategies in clinical settings by exploring ZS prediction, lightweight adapters (Linear Probe, kNN), and image encoder fine-tuning.
- Enhancing ZS prediction by introducing text and vision ensembles for medical tasks.
- A multi-site evaluation using MIMIC-CXR (in-domain, adults, part of training data), CheXpert (external data, adults) and VinDr-PCXR (out-of-domain, pediatric) datasets.
- Bias assessment of ZS and training-based methods across sex, age, and race subgroups.

An overview of related works is provided in Appendix Section A.

2. Methods

We evaluated the FM *MedImageInsight* for pneumonia detection in three different domains by assessing its ZS capability using nine different prompt types and enhanced prompt and vision ensembles (Figure 1) as well as by exploring training-based methods.

MedImageInsight (Codella et al., 2024) comprises three parts: an image encoder (360M parameters), a text encoder (250M parameters), and an optional text decoder (70M parameters). We excluded the text decoder, resulting in a lightweight FM with 610M parameters. The model was trained on image-text pairs from 14 modalities, including adult chest X-rays, along with radiology reports from MIMIC-CXR database (Johnson et al., 2019), and image-label pairs from NIH-CXR-LT (Holste et al., 2022) and Mass General Brigham database.

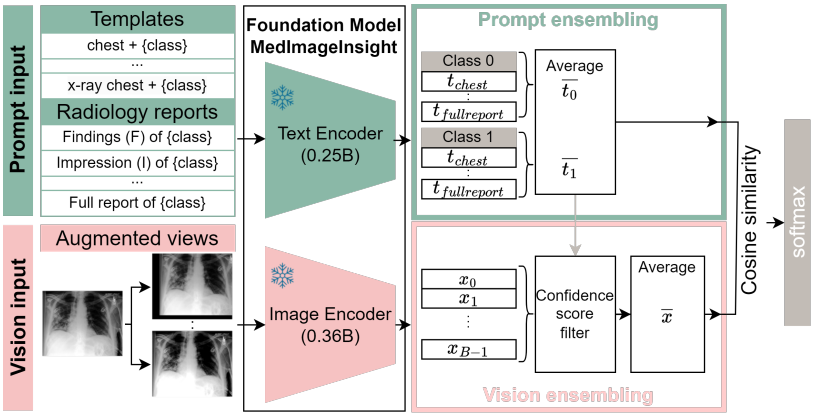

Figure 1: Overview of our Beyond-the-prompt pipeline using a zero-shot ensemble method. (top) Text embeddings are averaged over prompt templates and radiology reports and (bottom) compared to image ensembles generated by augmented views. A confidence-score filter is applied to select the most reliable image embeddings. Cosine similarity between text and image ensembles defines predicted classes.

2.1. Zero-shot Evaluation

Embedding-based contrastive vision-language FMs compute the cosine-similarity between an image embedding and multiple text embeddings to perform training-free ZS classification.

2.1.1. Prompt input

Templates as prompt. A common approach of constructing a prompt in ZS classification involves a combination of a *text-prompt template* plus a placeholder, {*class*}, that represents the class names, (here: *No Finding* or *Pneumonia*). The image is assigned to the class of the closest text prompt in the embedding space. We evaluated the FM's performance in predicting pneumonia using the following text prompts: 1) {*class*}, 2) *chest*+{class}, 3) *X-ray*+{class}, 4) *chest X-ray*+{class}, 5) *X-ray chest anteroposterior*+{class}.

Radiology reports as prompt. Since *MedImageInsight* is trained on radiology reports, we additionally evaluate our model by generating text embeddings using textual information extracted from the following parts of the radiology reports: 6) findings section, 7) impression section, 8) both, findings and impression sections, or 9) full radiology reports. We randomly sample ten reports from each class in the training set and compute the distance between the text embeddings generated from these reports and an image embedding from the test set. The predicted class is determined based on the majority of the five closest text embeddings.

Template and radiology report ensembles as prompt. As introduced by (Radford et al., 2021), generating averaged text embeddings can enhance ZS results and reduce computational complexity, as a general text ensemble is created once and reused during inference. An averaged text embedding $\overline{t_c} = \frac{1}{P}\sum_{p=1}^{P} t_{cp}$ is computed for each class c using all P prompt embeddings t. Most medical FMs are trained on template prompts and radiology

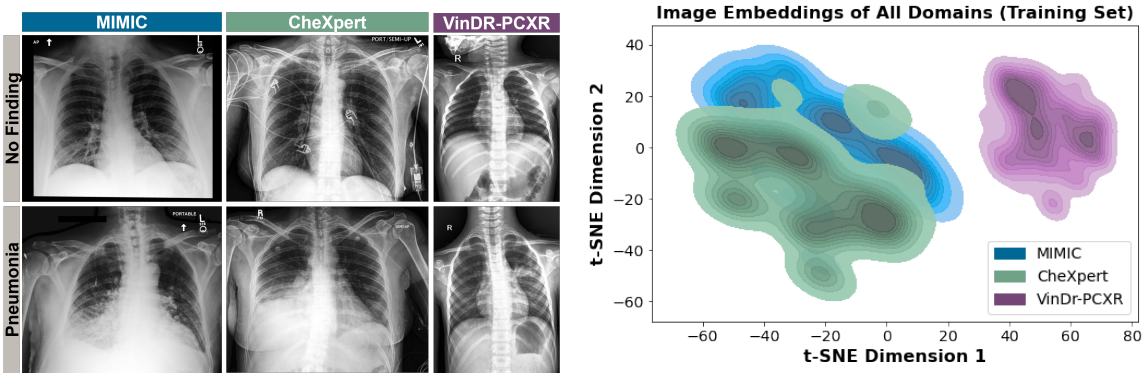

Figure 2: (left) Chest X-ray examples of 'No Finding' and 'Pneumonia' cases across adult (MIMIC, CheXpert) and pediatric (VinDr-PCXR) domains (right) Comparison of X-ray embeddings of *MedImageInsight* across the three domains using t-SNE. Embeddings of the test set and separated by classes are attached in Appendix C.

reports. Hence, merely averaging the template prompts may not be sufficient. Therefore, we propose an extension by generating: 10) averaged template-based embeddings using the templates (1-5), 11) report-based embeddings using ten reports per class of each report type (6-9), and 12) embeddings that incorporates both template- and report-embeddings (1-9).

2.1.2. Vision input

Original X-ray as vision input. The most common approach to generate an input embedding is to use the original image, which, in our case, represents a chest X-ray.

Vision ensemble as vision input. By augmenting an input image B times and creating B image embeddings, a single, representative embedding can be generated by averaging these B embeddings. This embedding is compared to a given text embedding (Shu et al., 2022; Döbler et al., 2024). This method aims to enhance the robustness and diversity of the image embeddings, potentially improving the alignment with text representations. We chose $B = 64$ as in (Shu et al., 2022) and applied a random selection of the following augmentation techniques: random rotation within a range of $\pm 10°$, random affine transformations with translation up to 10% of the image dimensions, color jittering (brightness=0.2, contrast=0.2), Gaussian blurring (kernel size = 5), and automatic contrast enhancement.

Confidence score filtered (CF) vision ensemble as vision input. In addition to averaging all augmented views, we implement a CF method as in (Shu et al., 2022). This approach identifies the N most confident augmented images using entropy-based confidence filtering by determining the 10% of samples with the lowest entropy.

2.2. Training-based adaption

To compare the training-free ZS methods, we investigate training-based adaptation strategies that leverage the image encoder by adding lightweight adapters as well as performing

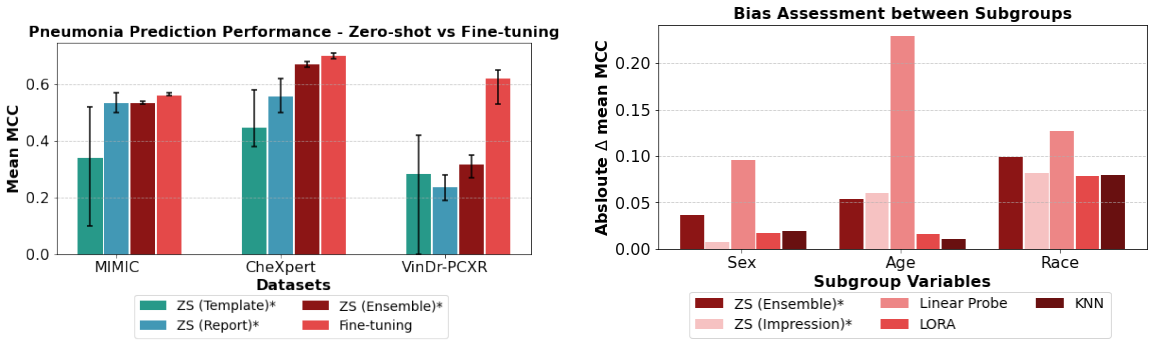

Figure 3: (left) Mean MCC for ZS experiments across different prompt types: (1-5) Template, (6-9) Report, and most effective Ensemble using (12) Template and Report (MIMIC, CheXpert); (10) Templates (VinDr-PCXR) compared to fine-tuning. (right) Bias Assessment: Absolute mean MCC across subgroups: sex (Male, Female), age (≤62,>62 years), race (White, Asian, Black). (*ZS: Zero-shot)

full fine-tuning of the image encoder. Unless stated otherwise, we used cross entropy loss and AdamW (Loshchilov, 2017) optimizer with a learning rate of 3×10^{-4}. If validation performance did not improve for five epochs, the learning rate was reduced by factor of 0.1.

Lightweight adapter training. In adapter training, the image encoder remains frozen while the image embeddings are further processed using subsequent lightweight adapter heads. We applied Linear Probing ($R^{1024\times 2}$) and k-nearest neighbor (kNN) ($k = 5$).

Fine-tuning using Low-Rank Adaptation (LoRA). The parameters of the image encoder are adapted using LoRA (Hu et al., 2021), which modifies the parameters with a low intrinsic rank updates (rank= 8).

Baseline Model. We used the DenseNet-121, which is based on a convolutional neural network architecture as a baseline model (Huang et al., 2017). With 8M parameters, DenseNet-121 is almost 98% smaller than the image encoder of *MedImageInsight*, but showed competitive results in various medical applications (Singh et al., 2024).

2.3. Datasets

We used three publicly available chest X-ray datasets to evaluate pneumonia prediction in three different environments. Figure 2 (left) shows representative examples for each dataset. To ensure consistency and fairness during training under varying amounts of training data, we balanced all training datasets using 744 samples. We also balanced all test datasets.

- **MIMIC-CXR (in-domain, adults)** (Johnson et al., 2019) was part of the FM training. Our test set included 8,186 X-rays with labels and radiology reports.
- **CheXpert (external validation, adults)** (Irvin et al., 2019) was not part of the training set for *MedImageInsight*, but comprises adult subjects, similar to MIMIC. Our balanced test set contains 2,508 samples. We also generated balanced test sets for the

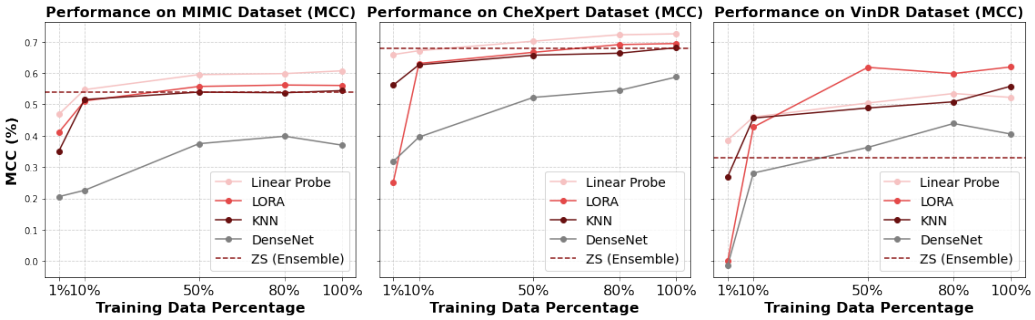

Figure 4: Comparison of ZS (Ensemble) and training-based methods across (left) MIMIC, (middle) CheXpert, (right) VinDr-PCXR using varying amount of training data. ZS (Ensemble) shows competitive performance, especially in low-data regimes.

demographic subgroups: sex (male/female - 1151 samples per group), age (young:<62 years/old: >62 years - 1128 samples per group), and race (White/Asian/Black - 171 samples per group) to assess biased prediction differences.

- **VinDr-PCXR (new domain, pediatrics)** (Pham et al., 2022) represents a new domain, as it exclusively contains pediatric cases, which were not part of the training of *MedImageInsight*. The balanced test set contains 178 samples.

3. Results and Discussion

3.1. Zero-shot Evaluation

As highlighted in Figure 3 (left), the ZS performance of *MedImageInsight* highly depends on the specific text prompt to which a given X-ray is compared. The detailed evaluation of all prompt types and metrics is depicted in Table 1.

Template as prompt. Among our five types of prompt templates, we obtained the highest Acc on all adult and pediatric datasets by using the prompt template (2) *chest*+{class}. Most other templates resulted even in either TPR or TNR below 50%.

Radiology reports as prompt. Comparing the image embeddings to parts of the radiology reports revealed notable performance boosts for the adult datasets. MIMIC performed best using (5) *Findings* as prompt (Acc = 78.0%, MCC = 0.57). CheXpert performed slightly better using (6) *Impression* (Acc = 80.5%, MCC = 0.62). While MIMIC and CheXpert achieved comparable results using either (5) *Findings* or (6) *Impression*, operating on full reports led to performance drops of up to 7% in Acc. In contrast, when comparing pediatric cases, VinDr-PCXR, to any part of MIMIC reports, TPR were consistently $< 20\%$.

Prompt Ensembles. Using prompt ensembles further improved ZS performance. For adult datasets, using prompt ensembles of (11) *Reports* or (12) *Templates + Reports* was most effective. Although, for MIMIC, using the simple (6) *Findings* prompt yielded slightly higher Acc and MCC, using prompt ensembles enhanced the TPR by up to 5.4% while maintaining a strong TNR. On VinDr-PCXR, significant improvements were observed using

Table 1: ZS performance for different text and vision prompts across adult (MIMIC, CheXpert), and pediatric (VinDr-PCXR) datasets. (Acc, TNR, TPR in [%].)

Prompt			Vision Ens.$^{+}$	MIMIC				CheXpert				VinDr-PCXR			
				Acc	**TNR**	**TPR**	**MCC**	**Acc**	**TNR**	**TPR**	**MCC**	**Acc**	**TNR**	**TPR**	**MCC**
Template	1		-	59.9	98.5	21.4	0.31	62.9	99.3	26.6	0.38	60.1	80.9	39.3	0.22
	2	*chest*		74.8	88.9	60.7	0.52	76.4	97.2	55.6	0.58	69.1	85.4	52.8	0.40
	3	*X-ray*		67.8	96.5	39.1	0.43	65.5	99.1	31.8	0.42	68.0	94.4	41.6	0.42
	4	*x. c.**		61.3	98.4	24.1	0.34	66.2	99.0	33.5	0.43	61.8	100	23.6	0.37
	5	*x. c. a.**		51.2	3.0	99.7	0.1	70.3	84.7	55.8	0.42	50.0	0	100	0
Report	6	Findings (F)	-	78.0	83.8	72.2	0.57	80.1	87.9	72.4	0.61	58.0	98.7	17.2	0.27
	7	Impression (I)		77.7	83.7	71.8	0.56	80.5	88.0	72.9	0.62	58.1	98.9	17.3	0.28
	8	F+I		74.6	84.7	64.5	0.50	73.6	90.3	56.9	0.50	53.3	100	7.0	0.19
	9	Full Report		74.7	84.8	64.5	0.50	73.5	90.3	56.6	0.50	53.9	100	8.0	0.2
Prompt Ensemble	10	Templates	-	64.2	97.5	30.8	0.38	73.7	98.3	48.5	0.54	63.5	66.9	60.1	0.27
	11	Reports		74.4	71.2	77.6	0.49	82.4	85.4	79.5	0.65	55.9	25.8	86.0	0.15
	12	Both		76.3	80.4	72.2	0.53	82.8	89.0	76.6	0.66	56.2	28.1	84.3	0.15
	10	Templates	All	64.3	97.6	31.0	0.38	70.1	99.2	40.1	0.49	66.9	77.5	56.2	0.35
	11	Reports	All	75.0	73.2	76.7	0.5	82.9	83.8	82.0	0.66	53.9	18.0	90.0	0.11
	12	Both	All	76.5	81.8	71.3	0.53	83.7	89.0	78.3	0.68	54.8	21.9	87.6	0.12
	10	Templates	CF^{++}	65.5	97.3	33.6	0.40	71.5	99.0	44.1	0.51	66.3	74.2	58.4	0.33
	11	Report	CF^{++}	75.1	74.3	76.0	0.50	82.7	83.9	81.6	0.65	55.6	18.0	93.2	0.17
	12	Both	CF^{++}	76.8	82.3	71.1	0.54	83.4	88.8	78.0	0.67	54.5	20.2	88.8	0.12

*x.=X-ray, c.=chest, a.=anteroposterior; $^{+}$Ens.= Ensemble; $^{++}$CF=confidence-score filter

(10) *Template* ensembles, resulting in a TPR increase of 7.3% compared to the best valid TPR of (2) *chest*+{*class*}.

Prompt and Vision Ensembles. By additionally generating vision ensembles with augmented views, slight improvements in Acc were achieved for MIMIC and CheXpert. Specifically, for the new adult domain, CheXpert, TPR was further improved by up to 2.5%. For the pediatric dataset, TPR slightly dropped when using vision ensembles.

Discussion. We found that using prompt ensembles is highly valuable and improves performance when applying FMs in ZS settings. Overall, we achieved best performance on CheXpert, followed by MIMIC, and the pediatric dataset (VinDr-PCXR). Our results showed that if a given X-ray image belongs to a distribution similar to training (i.e. MIMIC and CheXpert; Figure 2), prompt ensembles that include radiology reports enhance ZS performance. In this case, using vision ensembles further improved performance as more variability was added, better reflecting the known distribution. For new domains (e.g., VinDr-PCXR), where X-ray embeddings deviated from the learned distribution (Figure 2), and X-rays do not align with known radiology reports of adults, results were more reliable when using *Template* ensembles without report information. Similarly, image ensembles did not improve the results, as the augmented views failed to align with the known distribution.

3.2. Comparing training-based methods to ZS ensemble method

In Figure 4, we compare the MCC of the best ZS ensemble method against various training-based methods across all three datasets. For a detailed comparison of TPR and TNR, please refer to Figure 6 in Appendix E.

MIMIC. In low-data regimes, with only 1% of training data available, ZS ensembling performed best. With more training data, Linear Probe and LORA achieved higher MCC.

CheXpert. The ZS ensemble method performed best when less than 10% training data was available. If 50% of training data was available, Linear Probe outperformed the ZS approach by 2% in MCC. Using LoRA, 80% of training data was necessary to outperform the ZS approach by 1% in MCC. DenseNet and kNN performed worse than the ZS ensemble method regardless of the amount of training data.

VinDr-PCXR. While Linear Probe achieved the highest MCC with 1% of training data, its TPR remained $< 40\%$, indicating that it failed to reliably predict pneumonia. Hence, with only 1% training data, the ZS ensemble method still performed better than other methods. In general, LoRA fine-tuning achieved highest Acc with 86.6% using 50% data.

Discussion. In low data regimes, the training-free ZS ensemble method led to best performance in all domains. It remained competitive even with increasing amount of training data, especially on the adult datasets. With >10% annotated training data, Linear Probe further improved performance, highlighting the strength of *MedImageInsight* in capturing clinically meaningful features. On adult data, fine-tuning the image encoder caused catastrophic forgetting in low-data regimes and yielded only marginal improvements compared to the ZS ensemble method with more training data. In contrast, for pediatric cases, which are from an entirely new domain, fine-tuning with $> 50\%$ of training data captured the distribution shift from adults to pediatrics and outperformed other methods. The baseline model, DenseNet-121, failed to reliably learn pneumonia-related features from the adult datasets, as it consistently performed below the ZS ensemble. Only on VinDr, with $> 80\%$ training data, DenseNet-121 achieved a TPR and TNR $> 50\%$, and MCC exceeding the results of ZS ensemble. However, fine-tuning with LoRA still performed better than DenseNet-121.

3.3. Bias Assessment

Figure 3 (right) illustrates the absolute MCC differences across the demographic subgroups, sex, age, and race, in the CheXpert dataset. All evaluated methods exhibited varying levels of bias, reflected in performance differences across subgroups. However, ZS ensembling demonstrated a notable bias reduction compared to Linear Probe across all variables. For Linear Probe, differences ranged from 10% for sex, 14% for race, and up to 24% for age. In contrast, the ZS ensemble method achieved considerably lower bias levels, with the highest observed difference of 10% for race, while bias for sex and age remained below 5%. Although the ZS performance method demonstrated improved fairness compared to the Linear Probe, methods such as LORA and KNN exhibited even smaller differences across all variables.

4. Conclusion

In this work, we evaluated strategies for an effective application of a state-of-the-art vision-language FM in clinical settings based on pneumonia prediction across diverse populations. By applying text and vision ensembles for ZS prediction, we achieved up to 43% improvement in MCC in adults and 35% in pediatrics compared to training-based methods in low-data scenarios. Also, we reduced biases related to sex, age, and race by 6%-17% (MCC) compared to Linear Probe. This work supports the effective and resource-efficient integration of FMs into clinical workflows and enables equitable and accessible solutions for diverse patient populations.

Acknowledgments

LF is funded by the Global Glimpse Program of the University of Stuttgart, which is supported by the Deutsche Forschungsgemeinschaft (DFG) as part of the Excellence Strategy of the Federal and State Governments as well as by the Carl-Duisburg-Fellowship of the Bayer Foundation.

References

Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020.

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar Gardezi, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Cesar Truyts, et al. Merlin: A vision language foundation model for 3d computed tomography. *Research Square*, pages rs–3, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Pierre Chambon, Christian Bluethgen, Curtis P Langlotz, and Akshay Chaudhari. Adapting pretrained vision-language foundational models to medical imaging domains. *arXiv preprint arXiv:2210.04133*, 2022.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024.

Noel CF Codella, Ying Jin, Shrey Jain, Yu Gu, Ho Hin Lee, Asma Ben Abacha, Alberto Santamaria-Pang, Will Guyman, Naiteek Sangani, Sheng Zhang, et al. Medimageinsight: An open-source embedding model for general domain medical imaging. *arXiv preprint arXiv:2410.06542*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Mario Döbler, Robert A Marsden, Tobias Raichle, and Bin Yang. A lost opportunity for vision-language models: A comparative study of online test-time adaptation for vision-language models. *arXiv preprint arXiv:2405.14977*, 2024.

Louisa Fay, Erick Cobos, Bin Yang, Sergios Gatidis, and Thomas Küstner. Avoiding shortcut-learning by mutual information minimization in deep learning-based image processing. *IEEE Access*, 11:64070–64086, 2023.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Ben Glocker, Charles Jones, Mélanie Roschewitz, and Stefan Winzeck. Risk of bias in chest radiography deep learning foundation models. *Radiology: Artificial Intelligence*, 5(6): e230060, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Gregory Holste, Song Wang, Ziyu Jiang, Thomas C Shen, George Shih, Ronald M Summers, Yifan Peng, and Zhangyang Wang. Long-tailed classification of thorax diseases on chest x-ray: A new benchmark study. In *MICCAI Workshop on Data Augmentation, Labelling, and Imperfections*, pages 22–32. Springer, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Shih-Cheng Huang, Malte Engmann Kjeldskov Jensen, Serena Yeung-Levy, Matthew P Lungren, Hoifung Poon, and Akshay Chaudhari. Multimodal foundation models for medical imaging-a systematic review and implementation guidelines. *medRxiv*, pages 2024–10, 2024.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023.

Stephanie L Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, et al. Maira-1: A specialised large multimodal model for radiology report generation. *arXiv preprint arXiv:2311.13668*, 2023.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23): 12592–12594, 2020.

Ilya Loshchilov. Decoupled weight decay regularization. *arXiv e-prints*, pages arXiv–1711, 2017.

Hieu H Pham, Thanh T Tran, and Ha Quy Nguyen. Vindr-pcxr: An open, large-scale pediatric chest x-ray dataset for interpretation of common thoracic diseases. *PhysioNet (version 1.0. 0)*, 10:2, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

Samantha M Santomartino, John R Zech, Kent Hall, Jean Jeudy, Vishwa Parekh, and Paul H Yi. Evaluating the performance and bias of natural language processing tools in labeling chest radiograph reports. *Radiology*, 313(1):e232746, 2024.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

Sukhendra Singh, Manoj Kumar, Abhay Kumar, Birendra Kumar Verma, Kumar Abhishek, and Shitharth Selvarajan. Efficient pneumonia detection using vision transformers on chest x-rays. *Scientific Reports*, 14(1):2487, 2024.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138, 2024.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pages 962–970. PMLR, 2017.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022.

Appendix A. Related Work

The applied FM, *MedImageInsight*, builds upon large-scale contrastive multimodal pretraining, which includes 14 different domains, including chest X-rays. Compared to other FMs such as (Zhang et al., 2023; Hyland et al., 2023; Bannur et al., 2024), *MedImageInsight* is trained on images, text, and labels, enabling adaptation to diverse distributions, such as adult and pediatric cases. While some Large Language Models (LLMs), like Med-Gemini (Saab et al., 2024) or Med-PaLM-M (Tu et al., 2024), are trained on text and labels, they are more than 10 times larger compared to *MedImageInsight*.

Foundation models trained in a contrastive manner generalize well on zero-shot (ZS) tasks by aligning image and text embeddings without task-specific training. However, the success of ZS performance depends on the quality of the text prompt. Radford et al. (Radford et al., 2021) highlighted the sensitivity of ZS performance to text prompts in general natural tasks and, therefore, introduced the idea of text prompt ensembles using up to 80 different templates and averaging them over the embedding space. They demonstrated improvements of almost 5% on the natural image dataset ImageNet (Deng et al., 2009).

To this end, Shu et al. (Shu et al., 2022) introduced Test-time Prompt Tuning (TPT), which generates image embedding ensembles on the fly based on multiple augmented versions of one input image. To exclude noisy augmentations, they added a confidence-based filter. Döbler et al. (Döbler et al., 2024) combined both approaches, creating text ensembles from templates and vision ensembles from augmented images, and tested them on general domains. However, their effectiveness in medical contexts, particularly for diverse patient populations, remains unexplored.

The application of FMs in clinical studies has revealed significant biases in their feature embeddings (Glocker et al., 2023; Santomartino et al., 2024). Glocker et al. (Glocker et al., 2023) found that FMs often encode demographic factors, which might lead to performance differences across subgroups. Mitigation strategies include adversarial training (Ganin et al., 2016), fairness-aware loss functions (Zafar et al., 2017), or the reduction of shortcut learning

(Fay et al., 2023). In medical tasks, such biases are particularly concerning and need to be addressed to provide fair healthcare (Larrazabal et al., 2020).

Pneumonia detection from chest X-rays was studied for various model architectures in (Singh et al., 2024). They presented that different types of convolutional-based models, such as VGGs (Simonyan and Zisserman, 2014), ResNets (He et al., 2016), InceptionV3 (Szegedy et al., 2016), and DenseNets (Huang et al., 2017), perform worse than Vision Transformers (ViTs) (Alexey, 2020). However, in comparison, a Vision Transformer (ViT) has more than 85B trainable parameters, while the convolutional-based models have fewer than 200M. Likewise, the image encoder of *MedImageInsight* operates on 360M trainable parameters, offering a balance between efficiency and accuracy.

Appendix B. Limitations

While our study demonstrates promising approaches for deploying embedding-based vision-language FMs in clinical settings, this study does not explore other types of FMs, such as instruction-tuned FMs (e.g., (Chen et al., 2024)), which may require alternative strategies for adaptation and performance improvement. Moreover, we evaluated a variety of text prompts and ensemble strategies; however, it is possible that more effective templates or more complex prompt designs might further improve performance. Especially, the lack of pediatric-specific radiology reports likely constrained the performance on pediatric cases. Incorporating such domain-specific reports might yield better alignment and performance in pediatric populations. We limited our work to the prediction of pneumonia from chest X-rays, as this condition appears in both adult and pediatric patients. However, exploring further diseases as well as modalities could provide broader insights and reliability regarding the application of FMs in clinical settings. Finally, although our work explores the FM in three different environments, this may not fully reflect the heterogeneity of real-world clinical populations. In our upcoming work, we aim to address these limitations, especially by exploring more complex prompting strategies, as well as different domains and modalities, to enable a reliable and fair application of FMs in clinical workflows.

Appendix C. Extended Dataset

In Figure 5 (top), we show in addition to Figure 2 (right) the image embeddings of our balanced test datasets. Moreover, in Figure 5 (middle/bottom), the distributions of training/test dataset separated by the investigated classes *No Finding* and *Pneumonia* are shown as t-SNE plots across the adult datasets, MIMIC (left) and CheXpert (middle), as well as in the pediatric dataset, VinDr-PCXR (right).

Appendix D. Evaluation Metrics

In our experiments, all test datasets are balanced to ensure fair evaluation of pneumonia prediction. We evaluate the performance focusing on accuracy (Acc), true-negative rate (TNR), true-positive rate (TPR), and Matthews correlation coefficient (MCC).

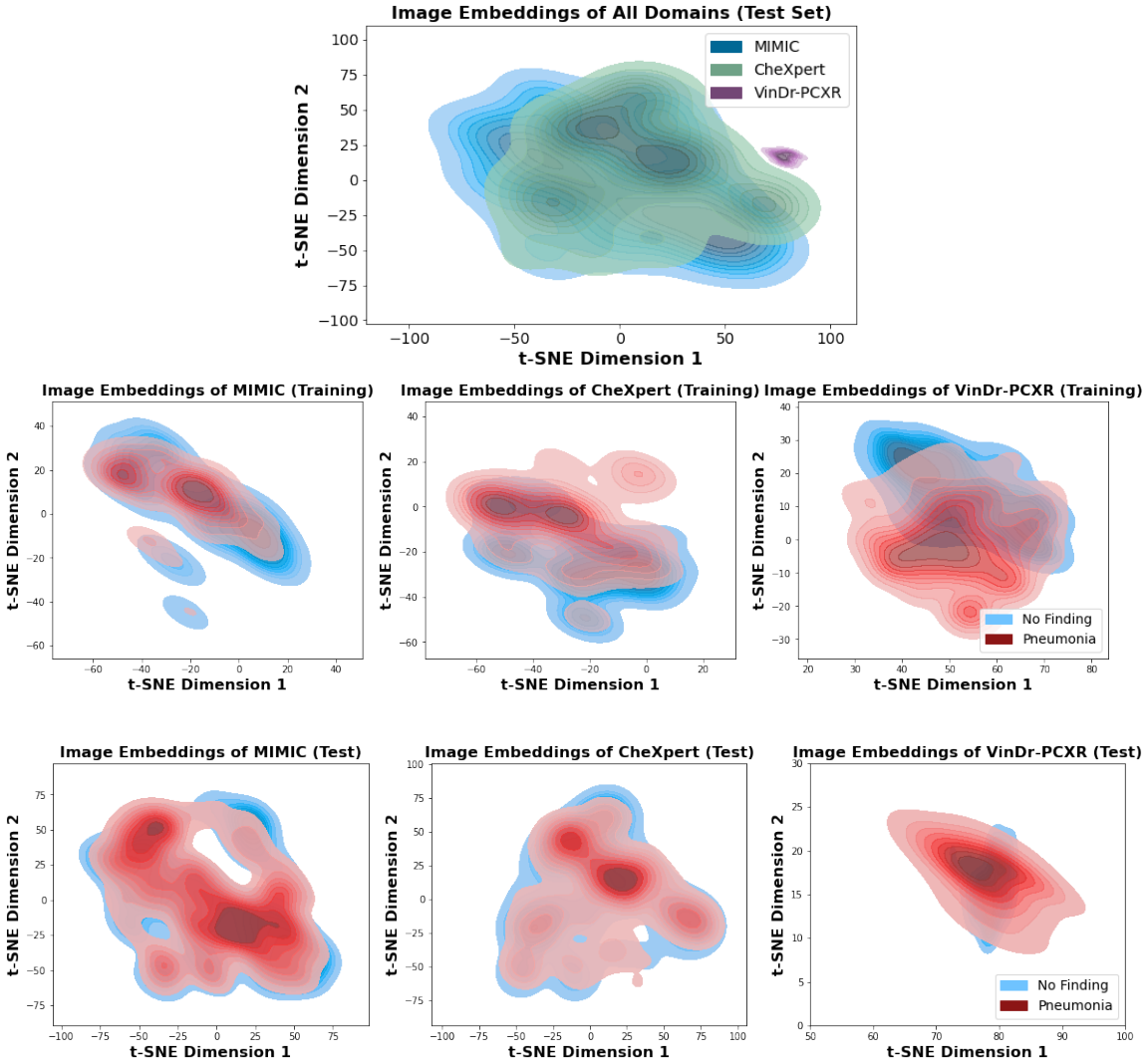

Figure 5: (Top) Comparison of X-ray embeddings of *MedImageInsight* across the test datasets of MIMIC (adults), CheXpert (adults), VinDr-PCXR (pediatrics) using t-SNE. (Middle) Comparison of the X-ray embeddings distribution of *No Finding* and *Pneumonia* across the training datasets of MIMIC (left), CheXpert (middle), VinDr-PCXR (right). (Bottom) Comparison for all test datasets.

Appendix E. Extended Results

In this section, we provide additional results of the experiment shown in Section 3.2. In Figure 6, we show Acc, TNR, and TPR of the training-based methods using a varying amount of training data compared to our ZS (Ensemble) Method.

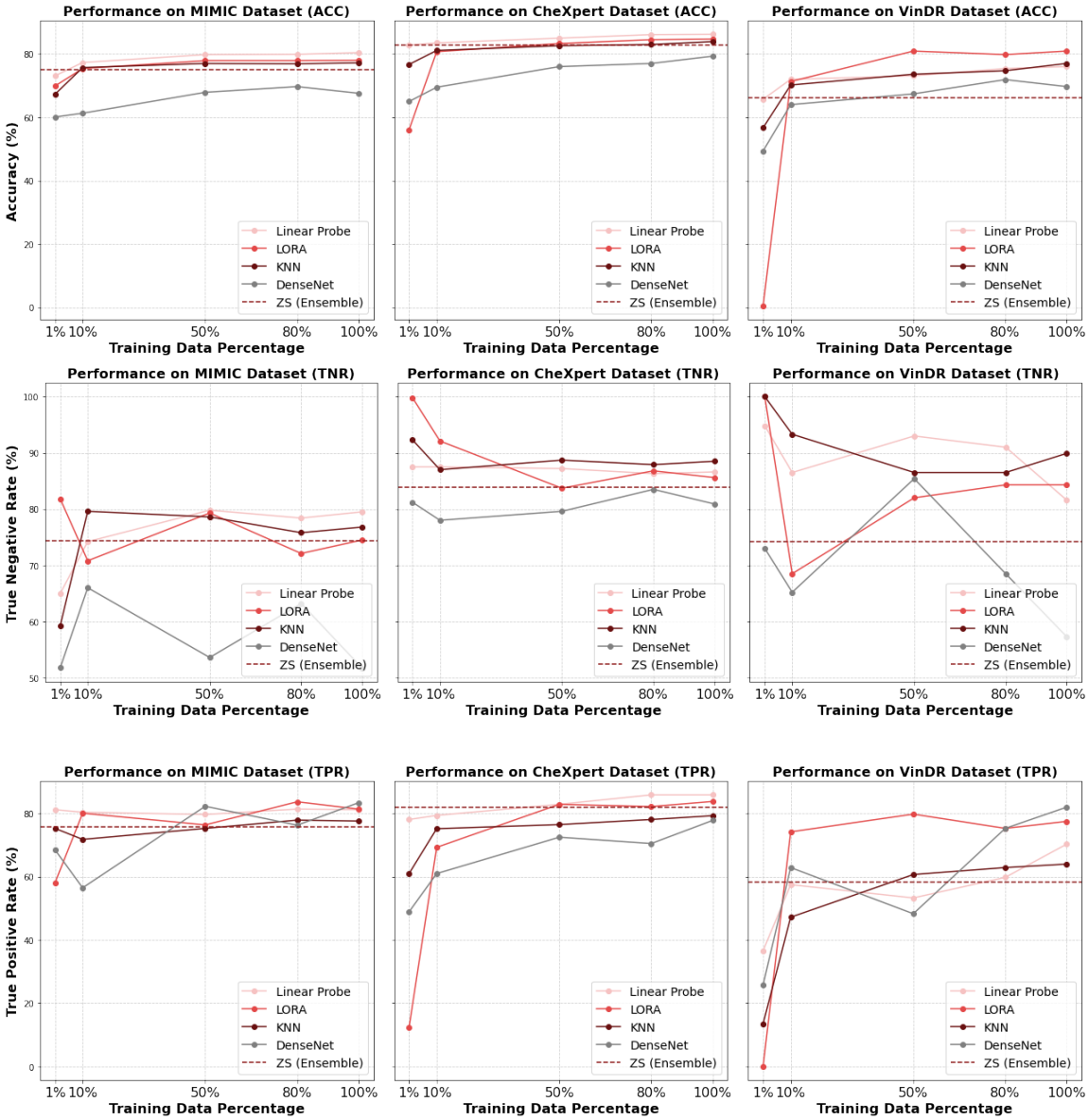

Figure 6: Comparison of Acc, TNR, TPR of ZS ensemble and training-based method across MIMIC, CheXpert, and VinDr-PCXR dataset.