CXR-CML: Improved zero-shot classification of long-tailed multi-label diseases in Chest X-Rays

Rajesh Madhipati[1*], Sheethal Bhat[1*], Lukas Buess[1], and Andreas Maier[1]

Friedrich-Alexander University Erlangen-Nuremberg, Erlangen 91054, Germany
{Rajesh.madhipati, Sheethal.bhat, lukas.buess, Andreas.maier}@fau.de

Abstract. Chest radiography (CXR) plays a crucial role in the diagnosis of various diseases. However, the inherent class imbalance in the distribution of clinical findings presents a significant challenge for current self-supervised deep learning models. These models often fail to accurately classify long-tailed classes. Current Vision-Language models such as Contrastive Language Image Pre-training (CLIP) models effectively model the manifold distribution of the latent space, enabling high zero-shot classification accuracies. Although CLIP performs well on most of the primary classes in the dataset, our work reveals that its effectiveness decreases significantly for classes with a long-tailed distribution. Our approach employs a class-weighting mechanism that directly aligns with the distribution of classes within the latent space. This method ensures a substantial improvement in overall classification performance, with particular emphasis on enhancing the recognition and accuracy of rarely observed classes. We accomplish this by applying Gaussian Mixture Model (GMM) clustering to the latent space. The subsequent clusters are further refined by Student t-distribution, followed by a metric loss that utilizes the altered embeddings. Our approach facilitates stable and adaptive clustering of the features. This results in a notable average improvement of 7% points in zero-shot AUC scores across 40 classes in the MIMIC-CXR-JPG dataset from previous SOTA models.
Our code is publicly available at: CXR-CML.

1 Introduction

Chest radiography (CXR) is one of the most widely utilized diagnostic tools in clinical practice, providing essential information on a variety of pulmonary and cardiothoracic conditions [27]. The availability of large public CXR datasets [9] with corresponding clinical reports has driven the development of vision-language (VL) models in CXR Artificial Intelligence (AI) research [1]. Moreover, the high cost of annotations, coupled with a shortage of radiologists, has prompted investigations on the effective use of self-supervised learning (SSL) methods [25,31,30,18,5]. While SSL methods rely on contrastive learning principles, CLIP [24] directly contrasts the extracted visual and language features,

[*] These authors contributed equally to this work.

thus achieving impressive zero-shot performance on inherently derived categories. Unlike other SSL methods, CLIP eliminates the need for further category based finetuning. Due to the prevalent data distribution in CXR datasets [20], CLIP based VL-SSL models deliver impressive performance [29,16,31,1] on commonly occuring diseases such as pneumonia or pleural effusion. However, many clinically relevant findings are underrepresented in training distributions [12], thus negatively impacting the robustness and clinical applicability of these models [33]. Early VL approaches like ConVIRT [34] and GLoRIA [11] introduced contrastive learning frameworks to align medical images and text, thereby improving multimodal representation learning. Building on these, CXR-BERT [3] specialized in pretraining on chest X-ray reports, while MedKLIP [31] enhanced performance by integrating structured clinical knowledge. Recent advancements [32,7,26], have further improved zero-shot pathology classification across multiple CXR datasets, demonstrating the reliability of VL models.

Although notable, these systems remain unsuitable for practical deployment due to inconsistent performance across all categories of pathologies [2]. Specifically, these methods often assume uniform Gaussian distributions in the latent space, which may not adequately represent the distribution heterogeneity present in medical datasets. This leads to suboptimal clustering and conflating feature representation for rare diseases [10].

Consequently, we present CXR-CML (Chest X-ray Contrastive Metric Learning) which seeks to model the latent distribution such that the long-tailed classes are appropriately clustered. For this purpose we first apply a Gaussian Mixture Model (GMM)[28] on the latent space derived from CLIP [24]. A Student t-distribution [15] further refines the GMM [28], enhancing the inherent clusters [24]. Subsequently, we utilize domain-specific metric learning to leverage these clusters and enhance the feature space, thereby ensuring that the model acquires distinct representations for both frequently and rarely seen classes. To evaluate the robustness of our model, we evaluated it in the 40 categories released by the "MICCAI challenge" [22]. This encompasses 12 rare and 28 common classes, offering a comprehensive model evaluation previously unavailable in literature.

Main contributions: 1) We introduce CXR-CML, a method to model the latent distribution manifold more effectively. This is accomplished through the application of GMM [28] and refined with a Student-t distribution. 2) We leverage the clustered distribution to apply a metric loss that yields robust improvement across a wide range of categories. 3) We conduct a robust evaluation using 5-fold cross-validation on 20% of the dataset across 12 long-tailed and 28 base classes in MIMIC-CXR-JPG dataset[12] . To our knowledge, this study covers the widest range of categories for CXR zero-shot classification evaluation.

2 Method

Given its remarkable zero-shot performance, our method is based on CLIP [23]. The model learns to correlate the images with their corresponding text captions

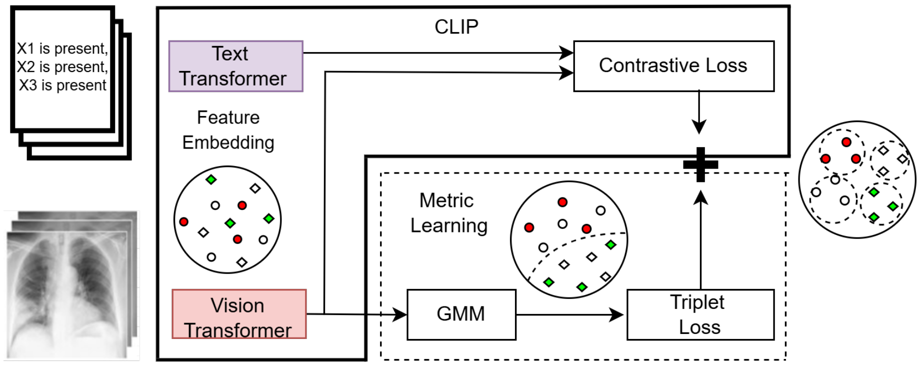


Fig. 1: A Text Transformer & Vision Transformer extract embeddings from textual and visual inputs. A Gaussian Mixture Model (GMM), enhanced by the Student's t-distribution, is used to form and refine initial clusters, where Label A and Label B represent semantically & visually similar conditions. Label C & Label D form another related cluster. We use contrastive and triplet loss, further refining the embedding space by improving intra-class compactness and inter-class separation. For example, Label A & Label B correspond to 'Atelectasis' and 'Lobar Atelectasis,' while Label C & Label D correspond to 'Pleural Effusion' and 'Pulmonary Edema,' respectively.

through a cosine similarity loss in a shared latent space. In **CXR-CML**, we apply GMM [28] on the CLIP extracted visual-language embeddings and refine them using a Student t-distribution. This is followed by a metric loss that further refines the feature space.

Modeling the latent space: GMM The GMM [28] is a probabilistic framework that models complex data distributions using multiple Gaussian components. The unsupervised algorithm effectively clusters high-dimensional data, without overfitting to the dominant classes. However, its clusters are soft assignments with overlapping boundaries, making it less distinct than more specialized clustering methods. This makes it especially suitable for multi-label CXR data where images contain multiple pathologies. Therefore, GMM's flexibility in approximating diverse distributions makes it a viable choice for modeling heterogeneous CXR data. The algorithm requires us to set the expected number of clusters, denoted as N. In CXR-CML, we apply GMM only on the visual features, as they exhibit greater variability within our dataset when compared to text.

Modeling the distribution: Student t-distribution The t-distribution's [15,21] heavy-tailed nature allows it to assign non-negligible probabilities to data points far from the mean, which is critical for capturing underrepresented classes in long-tailed distributions. This property is mathematically supported by its polynomial decay, as opposed to the exponential decay of the Gaussian distribution, enabling it to better model the rare but significant instances often found in medical data. Unlike traditional GMMs [28], which assume Gaussian-distributed data, the t-distribution is better suited for capturing the heavy-tailed nature of medical data, providing more robust covariance estimation and preventing overfitting to Gaussian assumptions. By integrating the t-distribution into our framework, we enhance the model's ability to handle long-tailed classes, ensuring stable and discriminative clustering.

Given a batch of feature embeddings $\mathbf{z} \in \mathbb{R}^d$, we model the data using a mixture of Student's t-distributions. This is mathematically described as:

$$p(\mathbf{z}) = \sum_{k=1}^{K} \pi_k \, \mathcal{T}(\mathbf{z} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \nu), \tag{1}$$

where π_k is the mixture weight, $\boldsymbol{\mu}_k$ is the mean vector, $\boldsymbol{\Sigma}_k$ is the covariance matrix, and ν is the degrees of freedom for the k-th t-distribution component. The degrees of freedom ν control the heaviness of the tails, allowing the model to better adapt to outliers and rare classes.

The Student's t-distribution is defined as:

$$\mathcal{T}(\mathbf{z} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \nu) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{d/2}|\boldsymbol{\Sigma}_k|^{1/2}} \left(1 + \frac{(\mathbf{z}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{z}-\boldsymbol{\mu}_k)}{\nu}\right)^{-\frac{\nu+d}{2}}, \tag{2}$$

where d is the dimensionality of the embeddings, and $\Gamma(\cdot)$ is the gamma function. As $\nu \to 1$, the distribution approaches the standard Cauchy distribution [13] with heavy tails, while as $\nu \to \infty$, it converges to the Gaussian distribution. This formulation allows the model to assign higher probabilities to outliers, making it more robust to rare classes. As a result, this algorithm enhances existing image-text correlation clusters.

Metric Learning: Triplet Loss In addition to the original contrastive loss $\mathcal{L}_c$ [24], we apply a metric loss on the GMM clusters. Metric learning includes various loss functions, such as Ranked List Loss and center loss, to enhance class discriminability. It is typically used to train a network to distinguish features that are hard to differentiate. In this context, given the data distribution and clusters formed by the GMM phase, we find triplet loss $\mathcal{L}_m$ to be suitable [4,17]. The GMM clustering assignments are used to generate pseudo-labels, which guide the selection of triplets needed for $\mathcal{L}_m$. The triplets $(\mathbf{a}, \mathbf{p}, \mathbf{n})$ are selected, where $\mathbf{a}$ is an anchor, $\mathbf{p}$ is a positive sample (from the same cluster as $\mathbf{a}$), and $\mathbf{n}$ is a negative sample (from a different cluster). The loss is mathematically defined as:

$$\mathcal{L}_m = \sum_{(\mathbf{a},\mathbf{p},\mathbf{n})} \max(0, d(\mathbf{a},\mathbf{p}) - d(\mathbf{a},\mathbf{n}) + \alpha), \tag{3}$$

where $d(\cdot, \cdot)$ is the standard Euclidean distance between embeddings $\mathbf{a}$ and $\mathbf{p}$, and α is a hyperparameter that defines the margin, ensuring sufficient separation between clusters. While $\mathcal{L}_c$ implicitly encourages intra-cluster compactness and inter-cluster separation, $\mathcal{L}_m$ provides explicit further guidance to the network. CXR-CML is trained with the complete loss which is given as,

$$\mathcal{L} = \mathcal{L}_c + \mathcal{L}_m. \tag{4}$$

Text Generation MIMIC-CXR-JPG [12] is a multi-label dataset. We generate "meta labels" using the data provided by MICCAI challenge "CXR-LT" [22]. The labels are derived from the clinical reports provided in the dataset. These labels indicate the absence or presence of specific diseases for each image. Textual descriptions are constructed for all classes that are marked as present in the groundtruth annotations. For example, for an image containing *adenopathy* and *pulmonary effusion*, the generated text is:

"adenopathy is present, pulmonary effusion is present"

By utilizing standard disease names, our method enables the text embeddings to act as weak supervisory signals to enhance training.

3 Experimental Setup

Dataset We evaluate our method on the MIMIC-CXR-JPG dataset [12,22], which consists of 234,800 frontal-view chest X-ray JPG images labeled with 39 disease classes. The original MIMIC-CXR-JPG dataset includes 14 disease classes, while the "CXR-LT MICCAI challenge" [22], introduced an additional 26 classes. [22] expands the scope of the dataset to include a much larger category of critical and underrepresented classes. Table 1 denotes the high data imbalance, showing the distribution of disease classes according to the sample count for each category. There are 11 disease classes containing more than 10,000 samples, 17 classes ranging between 1,000 and 10,000 samples, and 12 classes having fewer than 1,000 samples. For our experiments, we categorize the dataset into base and rare classes: classes with fewer than 1,000 samples are rare and the remaining are base. Consequently, rare classes constitute only 2% of the dataset.

Table 1: Distribution of disease classes based on the number of samples.

Class Category	**Number of Classes**
Common (>10,000 samples)	11
Medium (1,000–10,000 samples)	17
Rare (<1,000 samples)	12

Furthermore, we split the dataset into training and test sets using an 80:20 ratio, with no patient overlap between the sets. Additionally, we employ 5-fold cross-validation to robustly evaluate model performance. Text is generated from the annotation data provided by MICCAI challenge as in Sec. 2.

Implementation Details We use the ViT-B/32 backbone for CLIP [6] and conduct our experiments on a single node with 1 × NVIDIA RTX 2080 Ti GPU (11 GB memory). All images are scaled to 224 × 224 in keeping with the original CLIP architecture. The implementation is based on PyTorch version 2.4.0+cu118. We use a learning rate of $1e-6$ and a batch size bs of 32. The AUC scores for base and rare classes are reported as the average of 5 runs. For GMM, we use $N = 40$, as our dataset has labels for 40 pathologies. During the final loss calculation, both $\mathcal{L}_c$ and $\mathcal{L}_m$ are equally weighted, ensuring a balanced contribution to the optimization process. To optimize training, we employ a ReduceLROnPlateau [19] learning rate scheduler with a reduction factor of 0.1 and a patience of 2 epochs.

4 Results

Table 2 indicates the 5-fold average AUC scores for all, base and rare classes on the validation set [12]. Comparisons with other VL SOTA methods are also shown. All the comparison models are trained and evaluated in a similar way. MedClip [30], MedKLIP [31], and SLIP [18], achieve average AUCs of 0.476, 0.576, and 0.502, respectively. MedKLIP performs slightly better on rare classes (0.574 AUC) compared to SLIP (0.495 AUC) and MedClip (0.450 AUC). These models reflect considerable classification uncertainty, particularly for rare findings.

CheXzero [8], our baseline model, achieves a total AUC of 0.644, with 0.647 and 0.631 AUC for base and rare classes, respectively. On the other hand, our method, CXR-CML, achieves a macro AUC of 0.715, with 0.711 and 0.72 AUC for base and rare classes, respectively. CXR-CML delivers an impressive 7% gain over previous SOTA, indicating that CXR-CML successfully achieves SOTA when evaluated over a comprehensive list of categories. These results demonstrate how our method effectively leverages meta labels to learn the underlying data distribution manifold in the CXR dataset. Consequently, the model is able to discern even rarer classes that make up less than 2% of the dataset.

Ablation Study Table 3 shows the ablation study starting from the baseline [8] model, followed by CheXzero trained on our dataset, and then with different model configurations of CXR-CML. The impact of various batch sizes bs and degrees of freedom (ν) in Student's t-distribution is shown. The best performance is achieved with a batch size of 32 and $\nu = 4$, highlighting the importance of careful hyperparameter tuning for optimal results.

Degrees of Freedom (ν) We investigate the impact of the degrees of freedom parameter (ν) in Student's t-distribution. Our classification results improve with

Table 2: Zero-shot performance comparison of different methods depicted as average AUC for the 28 base classes and 12 rare classes. Results are averaged over 10 runs (std dev. < 0.04). Statistical significance ($p < 0.00001$) is indicated by * when compared to baseline.

Method	**Total AUC**	**Base AUC**	**Rare AUC**
MedClip [30]	0.476	0.488	0.450
MedKLIP [31]	0.576*	0.574*	0.574*
SLIP [18]	0.502	0.505	0.495
CheXzero (Baseline) [8]	0.644*	0.647*	0.631*
CXR-CML (Ours)	$\mathbf{0.715}^*$	$\mathbf{0.711}^*$	$\mathbf{0.720}^*$

Table 3: Ablation study of CXR-CML with different configurations compared to the CheXzero baseline [8]. Results are reported as average AUC for base and rare classes. Statistical significance ($p < 0.00001$) is indicated by *. Comparison of AUC scores for different model configurations. bs is batch size and ν is degree of freedom, $\nu = \infty$ represents Gaussian distribution

Method	**Total AUC**	**Base AUC**	**Rare AUC**
CheXzero (Baseline) [8]	0.644*	0.647*	0.631*
CheXzero + Meta labels 2	0.691*	0.681*	0.707*
CXR-CML ($bs = 32$, $\nu = 2$)	0.710	0.710	0.710
CXR-CML ($bs = 32$, $\nu = 4$)	$\mathbf{0.715}^*$	$\mathbf{0.711}^*$	$\mathbf{0.720}^*$
CXR-CML ($bs = 32$, $\nu = 6$)	**0.715**	0.710	0.718
CXR-CML ($bs = 32$, $\nu = \infty$)	**0.716**	0.713	0.720
CXR-CML ($bs = 16$, $\nu = 4$)	0.705	0.700	0.710
CXR-CML ($bs = 32$, $\nu = 4$)	$\mathbf{0.715}^*$	$\mathbf{0.711}^*$	$\mathbf{0.720}^*$
CXR-CML ($bs = 64$, $\nu = 4$)	0.711	0.707	0.716

increasing ν, while ν approaches values defining a Gaussian distribution [14]. The Gaussian distribution AUC scores are marginally higher, but without statistical significance when compared to ν=6. However, our experiments indicate that applying a student t-distribution results in greater model stability.

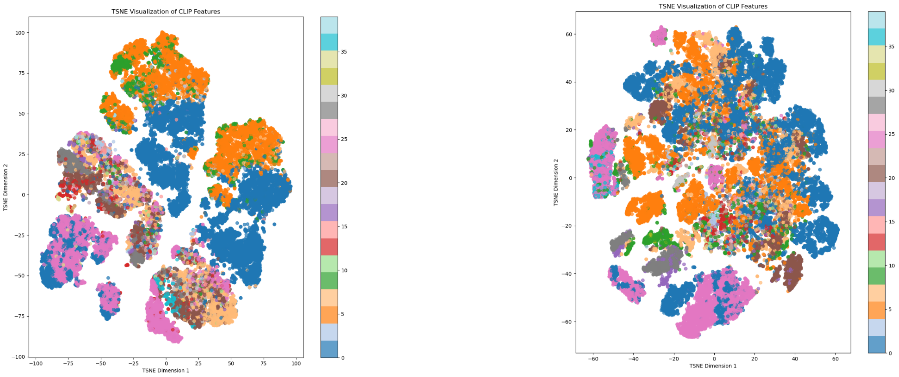
Batch Size We evaluate the effect of batch size on model performance. A batch size of 32 achieves the best performance, with an overall AUC of 0.715. Increasing the batch size to 64 results in a slight performance drop (0.715 to 0.711 AUC). Smaller batch sizes (e.g., 16 and 8) lead to more significant performance degradation (0.715 to 0.705 AUC), likely due to dependence of the contrastive loss $\mathcal{L}c$ on the batch size [24].

Feature visualization Fig. 2 illustrates the t-SNE plots of both the CheXzero baseline and our method, with a batch size of 32. Though the baseline model exhibits well defined clusters, the quantitative results show that our method performs better across the complete list of categories. Overall the results suggest

Table 4: Comparison of computational efficiency between CheXzero + Meta labels and CXR-CML.

Metric	**CheXzero + Meta labels**	**CXR-CML**	**Difference (%)**
FLOPs per second	1.87×10^7	2.26×10^7	+20.83%
Training time per step (s)	53.3882	66.2795	+24.15%

that CXR-CML can accurately capture the differentiating characteristics of the long-tailed classes.



(a) Latent space of CheXzero

(b) Latent space of CXR-CML

Fig. 2: t-SNE visualization of CLIP features for different models. The plot represents the latent space clustering of 40 disease classes using CheXzero and CXR-CML over full validation set.

Computational efficiency Table 4 tabulates the additional computational cost associated with CXR-CML compared to CheXzero + Meta labels. Specifically, CXR-CML incurs a 20.83% increase in FLOPs per second and a 24.15% increase in training time per step. This increase in computational overhead can be mainly attributed to the GMM component.

5 Conclusion

CXR-CML showcases a notable enhancement in zero-shot classification performance, by modeling the latent space with an emphasis on the long-tailed data distribution. In this study, we employ the Student t-distribution to provide a robust mathematical framework for clustering, enhancing the representation of underrepresented categories. This improved clustering further strengthens the subsequent metric learning stage, leading to enhanced classification performance. Future work will explore additional VL-SSL methods as comparative baselines and extend evaluation to other medical domains. A key limitation of CXR-CML is the additional computational cost introduced by the GMM stage to be addressed in future research.

Acknowledgments. This study was conducted as part of a Master's thesis. HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). The hardware is funded by the German Research Foundation (DFG).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., Ilse, M., Pérez-García, F., Salvatelli, V., Sharma, H., et al.: Maira-2: Grounded radiology report generation. arXiv preprint arXiv:2406.04449 (2024) 1, 2
2. Bhat, S., Panambur, A.B., Mansoor, A., Georgescu, B., Grbic, S., Maier, A.: Towards robust zero-shot chest x-ray classification: Exploring data distribution bias in chest x-ray datasets. BVM (2025) 2
3. Boecking, B., et al.: Cxr-bert: Pretraining chest x-ray reports for multimodal alignment. Journal of Biomedical Informatics (2022), https://huggingface.co/microsoft/BiomedVLP-CXR-BERT-general 2
4. Chen, K., Lei, W., Zhang, R., Zhao, S., shi Zheng, W., Wang, R.: Pcct: Progressive class-center triplet loss for imbalanced medical image classification (2022), https://arxiv.org/abs/2207.04793 4
5. Delitzas, A., Parelli, M., Hars, N., et al., G.V.: Multi-clip: Contrastive vision-language pre-training for question answering tasks in 3d scenes. In: 34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023. BMVA (2023), https://papers.bmvc2023.org/0748.pdf 1
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T.: An image is worth 16x16 words: Transformers for image recognition at scale (2021), https://arxiv.org/abs/2010.11929 6
7. Du, Y., Chang, B., Dvornek, N.C.: CLEFT. In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15012. Springer Nature Switzerland (October 2024) 2
8. E, T., Talius, Patel, et al., P.: Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. Nature Biomedical Engineering pp. 1399–1406 (2022) 6, 7
9. G, S.V., Ponraj, N., L, D.P.: Study on public chest x-ray data sets for lung disease classification. In: 2021 3rd International Conference on Signal Processing and Communication (ICPSC). pp. 54–58 (2021). https://doi.org/10.1109/ICSPC51351.2021.9451726 1
10. Holste, G., Wang, S., Jiang, Z., Shen, T.C., Shih, G., Summers, R.M., Peng, Y., Wang, Z.: Long-Tailed Classification of Thorax Diseases on Chest X-Ray: A New Benchmark Study, p. 22–32. Springer Nature Switzerland (2022). https://doi.org/10.1007/978-3-031-17027-0_3 2

11. Huang, Z., et al.: Gloria. Medical Image Analysis (2021), https://arxiv.org/html/2312.07353v3 2
12. Johnson, A.E.W., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., ying Deng, C., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs (2019), https://arxiv.org/abs/1901.07042 2, 5, 6
13. Lee, H.Y., Park, H.J., Kim, H.M.: A clarification of the cauchy distribution. Communications for Statistical Applications and Methods **21** (03 2014). https://doi.org/10.5351/CSAM.2014.21.2.183 4
14. Ley, C., Neven, A.: The value at the mode in multivariate *t* distributions: a curiosity or not? (2014), https://arxiv.org/abs/1211.1174 7
15. Li, R., Nadarajah, S.: A review of student's t distribution and its generalizations. Empirical Economics **58**, 1461–1490 (2020). https://doi.org/10.1007/s00181-018-1570-0 2, 3
16. Li, Y., et al.: Clip for cxr: Fine-tuning clip for chest x-ray disease classification. IEEE TMI (2023) 2
17. Messina, P., Vidal, R., Parra, D., Álvaro Soto, Araujo, V.: Extracting and encoding: Leveraging large language models and medical knowledge to enhance radiological text representation (2024), https://arxiv.org/abs/2407.01948 4
18. Mu, N., Kirillov, A., Wagner, D., Xie, S.: Slip: Self-supervision meets language-image pre-training (2021), https://arxiv.org/abs/2112.12750 1, 6, 7
19. Mukherjee, K., Khare, A., Verma, A.: A simple dynamic learning rate tuning algorithm for automated training of dnns (2019), https://arxiv.org/pdf/1910.11605 6
20. Park, S., Kim, G., Oh, Y.e.a.: Self-evolving vision transformer for chest x-ray diagnosis through knowledge distillation. Nature Communications **13**, 3848 (2022). https://doi.org/10.1038/s41467-022-31514-x 2
21. Peel, D., McLachlan, G.J.: Robust mixture modelling using the t distribution. Statistics and Computing **10**(4), 339–348 (2000) 3
22. Peng, Y., Lin, M., Holste, G., Wang, S., Zhou, e.a.: Cxr-lt 2024: Long-tailed, multi-label, and zero- shot classification on chest x-rays (Apr 2024). https://doi.org/10.5281/zenodo.10991413 2, 5
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A.: Learning transferable visual models from natural language supervision (2021), https://arxiv.org/abs/2103.00020 2
24. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., et al., G.G.: Learning transferable visual models from natural language supervision (2021), https://arxiv.org/abs/2103.00020 1, 2, 4, 7
25. Seputis, D., Mihailov, S., Chatterjee, S., Xiao, Z.: Multi-modal adapter for vision-language models (2024), https://arxiv.org/abs/2409.02958 1
26. Shentu, J., Al Moubayed, N.: Cxr-irgen: An integrated vision and language model for the generation of clinically accurate chest x-ray image-report pairs. In: WACV. pp. 5200–5209 (2024). https://doi.org/10.1109/WACV57701.2024.00513 2
27. Speets, A.M., van der Graaf, Y., Hoes, A.W., et al.: Chest radiography in general practice: indications, diagnostic yield and consequences for patient management (2006), https://pubmed.ncbi.nlm.nih.gov/16882374/ 1
28. Wan, H., Wang, H., Scotney, B., Liu, J.: A novel gaussian mixture model for classification. In: SMC. pp. 3298–3303 (2019). https://doi.org/10.1109/SMC.2019.8914215 2, 3, 4
29. Wang, X., et al.: Clip on medical twitter: Leveraging social media for disease detection. Journal of Medical Internet Research (2023) 2

30. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text (2022), https://arxiv.org/abs/2210.10163 1, 6, 7
31. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Medklip: Medical knowledge enhanced language-image pre-training in radiology (2023), https://arxiv.org/abs/2301.02228 1, 2, 6, 7
32. You, K.e.a.: Cxr-clip: Toward large scale chest x-ray language-image pre-training. In: Greenspan, H.e.a. (ed.) MICCAI 2023. Lecture Notes in Computer Science, vol. 14221, pp. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43895-0_10, https://doi.org/10.1007/978-3-031-43895-0_10 2
33. Zhang, X., Wu, C., Zhang, Y., Xie, W., Wang, Y.: Knowledge-enhanced visual-language pre-training on chest radiology images. Nature Communications **14**, 4542 (2023). https://doi.org/10.1038/s41467-023-40260-7 2
34. Zhang, Y., et al.: Convirt:. NeurIPS (2020), https://arxiv.org/pdf/2210.10163 2