IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

NEPTUNE- CXR: A Neuro- symbolic, Counterfactual, Uncertainty- Aware Model for Multi- Label Thoracic Disease Screening on Chest X- rays

Mrs. K Harshita Lakshmi, Dr. B Prasad Babu, Mr. Ch Venkatesh, Mr. Y Nagendra Kumar, Mr. K Gopala Reddy, Mrs.
D Naga Malika

Department of Computer Science and Engineering Ramachandra College of Engineering, Eluru, AP, India

Doi: https://doi.org/10.56975/ijcrt.v14i7.309987

Abstract—Automated thoracic disease screening from frontal chest X-rays is a multi-label learning problem where findings co-occur, supervision is often weak or uncertain, and deployment commonly involves domain shift and high interpretability requirements. NEPTUNE-CXR is a unified framework that couples anatomy-aware representation learning with uncertainty-aware decision outputs and clinically meaningful explanation artifacts. A frontal CXR is processed by APZ to estimate lung and heart masks, partition the lungs into six zones, and extract morphometrics; Z-ViT encodes a global token and zone tokens using mask-guided attention; CDN generates a closest-healthy counterfactual and residual evidence maps together with a distance-to-healthy plausibility signal; NSCG performs concept-graph reasoning to enforce cross-label consistency and produce concept activations and label logits; Uncertainty integrates evidential and Gaussian-process last-layer uncertainty with the plausibility signal to output calibrated multi-label probabilities. Outputs include per-label uncertainty scores, an abstention gate for selective prediction, conformal set-valued predictions, and explanation artifacts from zone attention, counterfactual/residual evidence, and concept activations. Evaluation is conducted on CheXpert (frontal-only) with comparisons against CuSeCXR and DEED, and optional transfer testing on one additional open-access corpus. The experimental study indicates improved discrimination, better calibration, stronger risk–coverage behavior, and more informative conformal sets, while providing evidence grounded in anatomy, counterfactual residuals, and concepts.

Keywords—Chest X-Ray Screening; Multi-Label Classification; Zone-Aware Vision Transformer; Counterfactual Normalization; Neuro-Symbolic Concept Graph; Uncertainty Quantification; Selective Prediction; Conformal Prediction.

I. INTRODUCTION

Chest X-ray imaging remains a primary tool for first-line assessment of thoracic disease due to low cost, fast acquisition, and broad availability across care settings. Automated screening from frontal chest X-rays supports triage, prioritization, and quality assurance in high-volume workflows, especially under limited access to specialist review. Many findings co-occur, including opacity patterns, pleural changes, and cardiac enlargement. Multi-label prediction with observations provides a natural formulation, and modeling label dependencies supports clinically meaningful decision support, Pham et al., [1]; Chen et al., [2]. Open-access

CXR corpora commonly use report-derived supervision with uncertainty. Weak supervision can be represented as $\tilde{} \in \{-1,0,1\}$, where $\tilde{} = 1$ denotes a positive label, $\tilde{} = 0$ denotes a negative label, and $\tilde{} = -1$ denotes an uncertain label. Hard binary targets can hide ambiguity and can encourage overconfident predictions under limited evidence. Distribution shift across institutions, scanners, protocols, and cohorts further degrades reliability. Interpretability is required for clinical trust, but post-hoc heatmaps can be unstable and can highlight non-pathologic regions. Transformer-based encoders support multi-disease CXR prediction by capturing long-range visual context, Du et al., [3]. Multi-relationship graph learning captures richer label relations, including spatial and semantic associations, Zhang et al., [4]. Severity- and uncertainty-aware learning supports dependency modeling under ambiguous supervision, Zhang et al., [5]. Practical screening also requires calibrated uncertainty and explanations aligned with clinical reasoning. Uncertainty quantification is central to trustworthy medical AI. Evidential learning represents uncertainty through distributions over class probabilities and reduces overconfidence under ambiguous supervision, Ashfaq et al., [6]. Bayesian approaches provide an additional route to epistemic uncertainty, including scalable sparse Bayesian modeling for medical imaging, Abboud et al., [7]. Uncertainty scores alone remain insufficient when uncertainty is not connected to interpretable evidence suitable for screening workflows. Interpretability has expanded beyond heatmaps to counterfactual explanations that generate plausible alternatives. Counterfactual generation for chest X-rays has been explored using generative models, Atad et al., [8], and calibration-aware counterfactual training has been studied to reflect decision uncertainty, Thiagarajan et al., [9]. Latent space smoothing improves plausibility and stability, Li et al., [10]. Class- specific counterfactuals support inherently interpretable multi-label prediction, Sun et al., [11].

Screening also requires consistency across multiple labels and alignment with clinical logic. Logic-guided learning incorporates background knowledge to reduce contradictions in classification outputs, Ledaguenel et al., [12]. Concept bottleneck models expose intermediate concept states for structured reasoning, and counterfactual concept bottleneck formulations connect concept interventions to explanation, Dominici et al., [13]. A unified design remains needed to combine anatomy-grounded evidence, cross-label consistency, and calibrated uncertainty outputs.

A gap remains for a single screening framework that links anatomy-aware representation learning, healthy-reference counterfactual evidence, neuro-symbolic concept-graph consistency, and uncertainty-aware decision outputs such as abstention and conservative label sets. NEPTUNE-CXR addresses this need through a unified pipeline that provides calibrated probabilities, uncertainty scores, conformal label sets, and evidence grounded in anatomy, counterfactual residuals, and concepts.

Contributions:

- A unified architecture for multi-label thoracic screening that combines anatomy-aware encoding, healthy-reference counterfactual evidence extraction, and concept-graph reasoning within one pipeline.
- Zone-aware representation learning that supports region-structured evidence aggregation and zone-level attribution aligned with thoracic anatomy.
- Weak -label handling using soft targets and reliability weights for uncertain supervision, compatible with a fixed evaluation mapping.
- Healthy -reference counterfactual normalization that derives localized residual evidence and a global plausibility signal for explanation and uncertainty-aware decisions.
- Neuro -symbolic concept graph integration that promotes clinically consistent outputs and exposes intermediate concept activations as rationales.
- Uncertainty -aware outputs that combine complementary uncertainty signals and support abstention-based selective prediction and calibrated set-valued predictions for conservative reporting.

Related research on dependency modeling, uncertainty estimation, counterfactual explanation, and logic-guided consistency is summarized next.

II. RELATED RESEARCH

Multi-label thoracic screening and label dependency modelling: Multi-label screening requires joint prediction of findings with clinically meaningful dependencies. Graph-based formulations treat findings as nodes and encode relations such as co-occurrence and semantic similarity; label semantic enhancement with graph convolution aligns visual features with structured label relations, Cai et al., [14]. Multi- relationship graphs extend relation modeling with multiple edge types to capture richer clinical structure, Zhang et al., [4]. Severity- and uncertainty-aware learning supports dependency modeling under ambiguous supervision, Zhang et al., [5]. Transformer-based encoders model global context and improve representation learning for multi-disease CXR prediction, Du et al., [3]. Hybrid designs combine transformer encoders with graph components for label relations, Lu et al., [15]. Language-derived signals from radiology text can enrich label semantics under weak supervision, Efimovich et al., [16]. Dependency-focused methods remain sensitive to dataset-specific correlations under label noise, long-tail label frequencies, and incomplete relation graphs. Coupling dependency modeling with calibrated uncertainty and evidence-grounded explanations remains an open direction.

Uncertainty modeling under weak and ambiguous supervision: Weak labels and expert disagreement motivate uncertainty-aware outputs for screening. Weak supervision can be represented as $\tilde{} \in \{-1,0,1\}$ with an explicit uncertain state. Evidential deep learning represents uncertainty through learned evidence and uncertainty-aware probabilities without sampling-based Bayesian inference, Ashfaq et al., [6]. Bayesian and approximate Bayesian methods provide epistemic uncertainty signals that can respond to distribution shift; sparse Bayesian modeling has been explored for efficiency in medical imaging, Abboud et al., [7]. Benchmark datasets and protocols emphasize uncertainty and severity as part of supervision and evaluation, Zhang et al., [4]; Zhang et al., [5]. Remaining challenges include calibration stability under domain shift, alignment between uncertainty scores and clinically interpretable evidence, and compatibility with conservative decision interfaces such as selective prediction and set-valued reporting. Such gaps motivate designs that connect uncertainty estimation to evidence sources and decision rules.

Interpretability and evidence communication: Heatmap-based explanations such as CAM can be diffuse or inconsistent in multi-label settings. Confidence-aware heatmap interpretation and localization reliability have been studied to improve usability, Rocha et al., [17], and ensemble strategies have been used to stabilize saliency under weak supervision, Aasem et al., [18]. Counterfactual explanations generate plausible alternatives by modifying image content to change predicted outputs. Trustworthy counterfactuals with latent space smoothing improve plausibility and reduce artifacts, Li et al., [10]. Class-specific counterfactual mechanisms integrate explanation into multi-label prediction, Sun et al., [11]. Counterfactual-powered reasoning can expose decisive evidence patterns for knowledge discovery, Fang et al., [19]. Concept-based explanations expose intermediate concept states for structured reasoning, and counterfactual concept bottleneck models connect concept activations to interventions, Dominici et al., [13]. Open limitations include dependence on plausibility constraints, sensitivity to spurious cues, and difficulty maintaining concept consistency under weak supervision. Debiasing counterfactuals under spurious correlations targets spurious cue sensitivity, Kumar et al., [20]. Evidence-grounded rationales compatible with uncertainty communication remain important for screening.

Neuro-symbolic reasoning and logic-based consistency in medical AI: Multi-label screening can yield clinically inconsistent label sets without structured constraints. Logic-based regularization incorporates background knowledge as constraints during learning or inference and can improve consistency under noisy labels, Ledaguenel et al., [12]. Concept-centric representations provide an interface for knowledge constraints, including counterfactual concept bottlenecks, Dominici et al., [13], and graph-based relation learning over findings, Zhang et al., [4]. Severity- and uncertainty-aware relation learning suggests that consistency constraints should accommodate ambiguity rather than enforce reliability weights w^{lab} □ w□ y~ □□$[0,1]^C$. The objective is estimation of calibrated label probabilities and uncertainty-aware decision outputs under ambiguous report-derived supervision.

NEPTUNE-CXR links anatomy grounding, counterfactual evidence extraction, concept-level consistency, and uncertainty-aware decision outputs (Figure 1). APZ estimates anatomical masks, partitions the lung field into six zones (right/left □ upper/mid/lower), and extracts morphometric descriptors. Z-ViT encodes x into a structured latent representation with a global token g and zone tokens $\{z_i\}_{i=1}^{6}$ constrained by mask-guided attention. CDN generates an

rigid rules, Zhang et al., [5]. Key challenges include scalability, anatomy-conditioned closest-healthy counterfactual x^{cf} ,

dependence on clinically valid rule sets, and robustness when rules are incomplete or context dependent. Soft, differentiable consistency mechanisms integrated with uncertainty-aware learning and evidence-based explanations address these challenges.

Robustness, domain shift, open-set settings, and multimodal trends: Generalization across institutions, scanners, and patient cohorts remains a barrier for deployment. Causality-inspired methods reduce spurious correlations through disentanglement and contrastive learning, Carloni et al., [21]. Counterfactual contrastive learning uses causal image synthesis to encourage robust representations, Roschewitz et al., [22]. Debiasing approaches directly address spurious residual evidence $R = x - x^{cf}$, and a distance-to-healthy score $d(x)$. NSCG performs message passing over a concept

graph to integrate $\{g, z_i\}_{i=1}^{6}$, residual summaries, and

morphometrics, producing concept activations and label logits under soft consistency constraints. An uncertainty layer combines evidential uncertainty, GP last-layer variance, and the plausibility signal $d(x)$, and supports abstention-based

selective prediction (Eq. 21) and conformal set-valued predictions (Eq. 22).

correlations, Kumar et al., [20]. Open-set recognition introduces additional reliability requirements; difficulty-aware $(\hat{p}, U, \mathrm{E}) = f_\theta(x)$ (1)

multi-label diagnosis has been studied for unknown conditions while maintaining performance on known findings, Ko et al., Here in Eq 1. $\hat{p} \in [0,1]^C$ denotes predicted label

[23]. Multimodal trends combine image models with NLP signals to strengthen label semantics, Efimovich et al., [16], and vision-language approaches increasingly emphasize uncertainty-aware reporting, Sharma et al., [24]. Existing work often optimizes dependency modeling, uncertainty, explanations, consistency, and robustness as separate probabilities, $U \in \mathbb{R}^C$ denotesper-label uncertainty estimates, and E denotes explanation artifacts derived from

zone-token attention, counterfactual residual maps R , and concept activations.

Notation: $x \in \mathbb{R}^{H \times W}$; $\tilde{y} \in \{-1,0,1\}^C$; $y = \phi(\tilde{y}) \in [0,1]^C$;

objectives. Limited work unifies anatomy-grounded evidence, healthy-reference counterfactual residuals, soft consistency mechanisms, and calibrated uncertainty outputs within one $w^{lab} \in [0,1]^C$; C labels; anatomical masks

M_{lung}, M_{heart}, M_{med} and zone masks $\{M_i\}_{i=1}^{6}$; global token

screening architecture.

g and zone tokens $\{z_i\}_{i=1}^{6}$; count x^{cf} ; residual map

III. METHODS

A. *Problem Formulation & Overall Architecture*

Multi-label thoracic screening from frontal chest radiographs is defined as prediction of C clinically relevant observations $R \square x \square x^{cf}$; distance-to-healthy $d \square x \square$; morphometrics $m \square R^{dm}$. Figure 1 overview of the proposed NEPTUNE- CXR pipeline. from a single chest X-ray (CXR). Let x $\square R^{H \square W}$ A frontal chest X-ray x is processed by APZ to obtain anatomical masks (lung, heart, and six lung zones) and normalized frontal CXR. Let $y \sim \square \{\square 1, 0,1\}^{C}$

IV. CONCLUSION

Multi-label thoracic screening from frontal chest X rays requires accurate recognition of co-occurring findings while maintaining interpretability and calibrated uncertainty under weak labels and deployment shift. NEPTUNE-CXR integrates APZ anatomy-aware encoding, Z-ViT zone-aware representations, CDN healthy-reference counterfactual evidence, NSCG concept-graph consistency, and Uncertainty outputs for abstention and conformal set-valued reporting. Evaluation on CheXpert against CuSeCXR and DEED indicates stronger overall discrimination, improved probability calibration, improved selective prediction behavior with lower risk at matched coverage, and tighter conformal sets at comparable coverage. Explanation artifacts support case inspection through zone-level attribution, counterfactual-to-healthy visualization with localized residual evidence, and concept-level rationales.

Limitations include reliance on weak labels and uncertain-label handling assumptions, sensitivity to image quality, positioning, and device artifacts, limited counterfactual plausibility in rare or complex multi-pathology cases, dependence on the completeness and correctness of the concept graph and rule set, and limited external validity evidence beyond a single benchmark.

- Multi-institutional validation across diverse protocols and cohorts, with stronger domain-shift and subgroup analysis.
- Clinically curated, maintainable concept graphs with processes for updating relationships and constraints.
- Improved detection of out-of-distribution and unknown conditions, paired with principled equal to the median. Group sizes follow the median split of the abstention for safe escalation. CheXpert test set: $N_A \square 9556$ and $N_B \square 9557$. NEPTUNE-CXR reports Group A macro AUROC 0.914 with ECE 0.026 and Group B macro AUROC 0.908 with ECE Human-in-the-loop evaluation of explanations and uncertainty for trust, error detection, and triage. Prospective study designs assessing workflow impact and whether calibrated uncertainty reduces downstream risk.

Overall, NEPTUNE-CXR advances a unified direction for more trustworthy multi-label thoracic screening where transparency and uncertainty are core outputs.

CONFLICTS OF INTEREST

"The authors declare no conflict of interest."

ACKNOWLEDGMENT

The authors acknowledge the creators and maintainers of publicly available chest X-ray datasets and benchmark resources, particularly CheXpert, which enabled the evaluation reported in this study. The authors also recognize the broader medical imaging research community for foundational contributions in uncertainty-aware learning, counterfactual explanation, graph-based reasoning, and interpretable chest X-ray analysis that informed the development of NEPTUNE-CXR.

REFERENCES

[1] H. H. Pham, T. T. Le, D. Q. Tran, D. T. Ngo, and H. Q. Nguyen, "Interpreting chest X-rays via CNNs that exploit disease dependencies and uncertainty labels," Nov. 29, 2019. doi: 10.1101/19013342.

[2] B. Chen, J. Li, G. Lu, H. Yu, and D. Zhang, "Label Co-Occurrence Learning With Graph Convolutional Networks for Multi-Label Chest X-Ray Image Classification," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 8, pp. 2292–2302, Aug. 2020, doi: 10.1109/JBHI.2020.2967084.

[3] D. Songfeng and C. Xingwu, "Exploiting Uncertainty: A Transformer-Based Multi-Disease Diagnostic Approach to Chest X-RAYS," in *2023 20th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, Chengdu, China: IEEE, Dec. 2023, pp. 1–5. doi: 10.1109/ICCWAMTIP60502.2023.10387037.

[4] M. Zhang *et al.*, "A New Benchmark: Clinical Uncertainty and Severity Aware Labeled Chest X-Ray Images With Multi-Relationship Graph Learning," *IEEE Trans. Med. Imaging*, vol. 44, no. 1, pp. 338–347, Jan. 2025, doi: 10.1109/TMI.2024.3441494.

[5] M. Zhang *et al.*, "Expert Uncertainty and Severity Aware Chest X-Ray Classification by Multi-Relationship Graph Learning," 2023, *arXiv*. doi: 10.48550/ARXIV.2309.03331.

[6] A. Ashfaq, M. Lingman, M. Sensoy, and S. Nowaczyk, "DEED: DEep Evidential Doctor," *Artificial Intelligence*, vol. 325, p. 104019, Dec. 2023, doi: 10.1016/j.artint.2023.104019.

[7] Z. Abboud, H. Lombaert, and S. Kadoury, "Sparse Bayesian Networks: Efficient Uncertainty Quantification in Medical Image Analysis," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, vol. 15010, M. G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, and J. A. Schnabel, Eds., Cham: Springer Nature Switzerland, 2024, pp. 675–684. doi: 10.1007/978-3-031-72117-5_63.

[8] M. Atad *et al.*, "CheXplaining in Style: Counterfactual Explanations for Chest X-rays using StyleGAN," 2022, *arXiv*. doi: 10.48550/ARXIV.2207.07553.

[9] J. J. Thiagarajan, K. Thopalli, D. Rajan, and P. Turaga, "Training calibration-based counterfactual explainers for deep learning models in medical image analysis," *Sci Rep*, vol. 12, no. 1, p. 597, Jan. 2022, doi: 10.1038/s41598-021-04529-5. Y. Li, X. Cai, C. Wu, X. Lin, and G. Cao, "A Trustworthy Counterfactual Explanation Method With Latent Space Smoothing," *IEEE Trans. on Image Process.*, vol. 33, pp. 4584–4599, 2024, doi: 10.1109/TIP.2024.34426144

[10] S. Sun, S. Woerner, A. Maier, L. M. Koch, and C. F. Baumgartner, "Inherently Interpretable Multi-Label Classification Using Class-Specific Counterfactuals," 2023, *arXiv*. doi: 10.48550/ARXIV.2303.00500.

[11] A. Ledaguenel, C. Hudelot, and M. Khouadjia, "Improving Neural-based Classification with Logical Background Knowledge," 2024, *arXiv*. doi: 10.48550/ARXIV.2402.13019.

[12] G. Dominici, P. Barbiero, F. Giannini, M. Gjoreski, G. Marra, and M. Langheinrich, "Counterfactual Concept Bottleneck Models," 2024, *arXiv*. doi: 10.48550/ARXIV.2402.01408.

[13] D. Cai, H. Lu, Z. Chai, R. Wang, W. Zhu, and Y. Yao, "Label Semantic Improvement with Graph Convolutional Networks for Multi-Label Chest X-Ray Image Classification," in *2023 13th International Conference on Information Technology in Medicine and Education (ITME)*, Wuyishan, China: IEEE, Nov. 2023, pp. 711–717. doi: 10.1109/ITME60234.2023.00147.

[14] Y. Lu *et al.*, "CvTGNet: A Novel Framework for Chest X-Ray Multi-label Classification," in *Proceedings of the 21st ACM International Conference on Computing Frontiers*, Ischia Italy: ACM, May 2024, pp. 12–20. doi: 10.1145/3649153.3649216.

[15] M. Efimovich, J. Lim, V. Mehta, and E. Poon, "Multilabel Classification for Lung Disease Detection: Integrating Deep Learning and Natural Language Processing," 2024, *arXiv*. doi: 10.48550/ARXIV.2412.11452.

[16] J. Rocha, S. Cardoso Pereira, A. Campilho, and A. M. Mendonça, "Confident-CAM: Improving Heat Map Interpretation in Chest X-Ray Image Classification," in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Istanbul, Turkiye: IEEE, Dec. 2023, pp. 4116–4123. doi: 10.1109/BIBM58861.2023.10386065.

[17] M. Aasem and M. Javed Iqbal, "Toward explainable AI in radiology: Ensemble-CAM for effective thoracic disease localization in chest X-ray images using weak supervised learning," *Front. Big Data*, vol. 7, p. 1366415, May 2024, doi: 10.3389/fdata.2024.1366415.

[18] Y. Fang, Z. Jin, X. Xing, S. Walsh, and G. Yang, "Decoding Decision Reasoning: A

Counterfactual-Powered Model for Knowledge Discovery," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, Athens, Greece: IEEE, May 2024, pp. 1–5. doi: 10.1109/ISBI56570.2024.10635275.

[19] A. Kumar *et al.*, "Debiasing Counterfactuals in the Presence of Spurious Correlations," in *Clinical Image-Based Procedures, Fairness of AI in Medical Imaging, and Ethical and Philosophical Issues in Medical Imaging*, vol. 14242, S. Wesarg, E. PuyolAntón, J. S. H. Baxter, M. Erdt, K. Drechsler, C. Oyarzun Laura, M. Freiman, Y. Chen, I. Rekik, R. Eagleson, A. Feragen, A. P. King, V. Cheplygina, M. Ganz-Benjaminsen, E. Ferrante, B. Glocker, D. Moyer, and E. Petersen, Eds., Cham: Springer Nature Switzerland, 2023, pp. 276–286. doi: 10.1007/978-3-031-45249-9_27.

[20] G. Carloni, S. A. Tsaftaris, and S. Colantonio, "CROCODILE: Causality Aids RObustness via COntrastiveDIsentangledLEarning," in *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, vol. 15167, C. H. Sudre, R. Mehta, C. Ouyang, C. Qin, M. Rakic, and W. M. Wells, Eds., Cham: Springer Nature Switzerland, 2025, pp. 105–116. doi: 10.1007/978-3-031-73158-7_10.

[21] M. Roschewitz, F. De Sousa Ribeiro, T. Xia, G. Khara, and B. Glocker, "Counterfactual Contrastive Learning: Robust Representations via Causal Image Synthesis," in *Data Engineering in Medical Imaging*, vol. 15265, B. Bhattarai, S. Ali, A. Rau, R. Caramalau, A. Nguyen, P. Gyawali, A. Namburete, and D. Stoyanov, Eds., Cham: Springer Nature Switzerland, 2025, pp. 22–32. doi: 10.1007/978-3-031-73748-0_3.

[22] K. Ko, B. Lee, J. Hong, and H. Ko, "Open Set Medical Diagnosis via Difficulty-Aware Multi-Label Thorax Disease Classification," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Orlando, FL, USA: IEEE, Jul. 2024, pp. 1–4. doi: 10.1109/EMBC53108.2024.10782506.

[23] N. Sharma, "CXR-Agent: Vision-language models for chest X-ray interpretation with uncertainty aware radiology reporting," 2024, *arXiv*. doi: 10.48550/ARXIV.2407.08811. Dataset: Irvin, Jeremy, Pranav Rajpurkar, Michael Ko, Yifan Yu, SilvianaCiurea-Ilcus, Chris Chute, Henrik Marklund et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison." In *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, pp. 590-597. 2019.

[24] CheXpert. https://www.kaggle.com/datasets/ashery/chexpert.