Supplementary Material : Chest X-Ray Feature Pyramid Sum Model with Diseased Area Data Augmentation Method

Changhyun Kim*
SK Telecom
Seoul, South-Korea
changhyk@sktmedai.com

Giyeol Kim†
Gacheon University
Gyeonggido, South-Korea
rlduf422@gachon.ac.kr

Sooyoung Yang‡
ChungAng University
Seoul, South-Korea
jimmy1016@cau.ac.kr

Hyunsu Kim§
SungKyunKwan University
Gyeonggido, South-Korea
hyunsu517@g.skku.edu

Sangyool Lee¶
SK Telecom
Seoul, South-Korea
sangyoollee@sktmedai.com

Hansu Cho‖
SK Telecom
Seoul, South-Korea
hansu.cho@sktmedai.com

Used Feature Pyramid	F1	mAP	Recall	Precision	AUC
4^{th}	27.23	32.66	76.39	20.54	82.69
4^{th} + 2^{nd}	26.97	30.92	74.32	20.33	81.26
4^{th} + 3^{rd} + 2^{nd}	27.12	31.96	74.91	20.42	81.75
4^{th} + 3^{rd} + 2^{nd} + 1^{st}	26.88	31.61	75.57	20.22	81.61

Table 1: Performance of Feature Pyramid Classifier, use_binary_enc=1, use_concatenate=1

Used Feature Pyramid	F1	mAP	Recall	Precision	AUC
4^{th}	27.18	32.76	76.74	20.48	82.58
4^{th} + 2^{nd}	26.84	30.99	74.41	20.37	81.12
4^{th} + 3^{rd} + 2^{nd}	26.98	32.21	74.93	20.31	81.86
4^{th} + 3^{rd} + 2^{nd} + 1^{st}	27.01	31.87	74.83	20.35	81.62

Table 2: Performance of Feature Pyramid Classifier, use_binary_enc=0, use_concatenate=1

Appendix A. MIMIC validation result using FPSM based on Tresnet [6]

When concatenate is used in the feature merger, overall performance deteriorates compared to when sum is used. Therefore, it can be seen that the advantage of used feature pyramid is no longer a subject of discussion.

Appendix B. Ablation studies for different dataset : Mura MSK (musculoskeletal) [5]

MURA MSK (Musculoskeletal) is a data set in which Train/Test consists of 32k images and 8k images and has normal/abnormal binary class labels for 7 parts of the musculoskeletal system (elbow, finger, forearm, hand, humerus, shoulder, wrist). We conducted a 5 fold-CV experiment and measured the CK score along with other metrics (Acc, AUPRC, AUC) to compare with other SOTA papers. As shown in Table 3, the Tresnet baseline compared to the existing SOTA [1] performance improved by 1.4% in the ck score, and the model using the feature pyramid 4 (4th+3rd+2nd+1st) achieved the best performance with 69.4%. If the proposed 1 model had the best performance in CK score and accuracy, the proposed 2 model had the best performance in AUPRC and AUC. This is a model using feature pyramid (4th+3rd+2nd) 3. Comparing Tables 3 and 4 with 5 and 6 can explain the same performance improvement as comparing Tables 2 and 3 in main paper with Tables 1 and 2. In other words, it was experimentally proved that sum feature merger is a more suitable process for multi-label classifier loss than concatenate feature merger.

Appendix C. Ablation studies for different models

To ascertain the efficacy of the TResNet model and the impact of data augmentation, a series of comprehensive ablation studies were conducted. The computations were carried out on A100 and V100 GPUs, employing the Adam op-

*First and Corresponding author
†Second author contributed equally
‡Second author contributed equally
§Second author contributed equally
¶Third author
‖Third author

Models	CK score	Acc	AUPRC	AUC
[1] Proposed 1	**0.694**	**0.848**	0.898	0.903
[2] Proposed 2	0.680	0.842	**0.903**	**0.906**
[Base line]TResNet [6]	0.674	0.844	0.894	0.899
CNN Ensemble [1]	0.66	0.797	n/a	n/a
VGG16 [7]	0.532	0.767	n/a	n/a

Table 3: Binary normal/abnormal classification performance in MURA [5] dataset ([1]Feature Pyramid=4, 7 part label embedding=0 use_concatenate=0, [2]Feature Pyramid=3, 7 part label embedding=1 use_concatenate=0)

Used Feature Pyramid	CK Score	Acc	AUPRC	AUC
4^{th} (Tresnet baseline)	67.42	84.42	89.46	89.89
4^{th} + 2^{nd}	67.62	83.92	90.08	90.35
4^{th} + 3^{rd} + 2^{nd}	67.26	83.77	89.47	89.76
4^{th} + 3^{rd} + 2^{nd} + 1^{st}	**69.41**	**84.83**	89.79	90.29

Table 4: Mura MSK [5] Performance of Feature Pyramid Classifier, use_binary_enc=0, use_concatenate=0

Used Feature Pyramid	CK Score	Acc	AUPRC	AUC
4^{th} (Tresnet baseline)	67.89	84.24	89.57	90.13
4^{th} + 2^{nd}	67.91	83.95	89.98	90.11
4^{th} + 3^{rd} + 2^{nd}	67.99	84.24	**90.27**	**90.57**
4^{th} + 3^{rd} + 2^{nd} + 1^{st}	67.92	84.08	89.77	90.08

Table 5: Mura MSK [5] Performance of Feature Pyramid Classifier, use_binary_enc=1, use_concatenate=0

Used Feature Pyramid	CK Score	Acc	AUPRC	AUC
4^{th} (Tresnet baseline)	68.74	84.39	89.01	90.08
4^{th} + 2^{nd}	68.88	84.61	89.62	90.10
4^{th} + 3^{rd} + 2^{nd}	66.60	83.86	89.25	89.62
4^{th} + 3^{rd} + 2^{nd} + 1^{st}	67.52	84.27	88.91	89.85

Table 6: Mura MSK [5] Performance of Feature Pyramid Classifier, use_binary_enc=1, use_concatenate=1

timizer in conjunction with the Cosine Annealing LR scheduler.

Initially, we employed the ResNet50 as a feature extractor, utilizing the model weights provided by the Chest X-ray Self-Supervised model called CheSS [3] paper. CheSS

Used Feature Pyramid	CK Score	Acc	AUPRC	AUC
4^{th} (Tresnet baseline)	68.25	84.42	88.96	89.93
4^{th} + 2^{nd}	67.55	84.05	88.57	90.03
4^{th} + 3^{rd} + 2^{nd}	66.58	83.77	89.22	89.60
4^{th} + 3^{rd} + 2^{nd} + 1^{st}	67.17	83.80	89.01	89.58

Table 7: Mura MSK [5] Performance of Feature Pyramid Classifier, use_binary_enc=0, use_concatenate=1

is a model pretrained using the MoCo-v2 [2] methodology on an X-ray dataset, with the ResNet50 serving as the base network. When the model was trained on the original MIMIC dataset, the validation accuracy achieved was 71%, with a validation mAP of 25.7%. Subsequently, under the same conditions, when experimenting with the ImageNet pretrained ViT[4] model, the validation accuracy reached 78%, accompanied by a validation mAP of 29.5%.

Moving on, experiments were conducted on the ViT model using augmented MIMIC data. This led to two distinct scenarios: training the augmented MIMIC data from scratch, or first pretraining on the original MIMIC dataset and then fine-tuning on the augmented MIMIC data. In the former case, a validation accuracy of 85% and a validation mAP of 28.1% were achieved. In contrast, the latter scenario resulted in a validation accuracy of 70% and a validation mAP of 26.1%. Notably, both scenarios exhibited slightly lower performance compared to using the original MIMIC dataset exclusively.

Data	Image Size	Model	Val Acc	Val mAP
Orig. MIMIC	512	ResNet50	71	25.7
Orig. MIMIC	384	ViT	78	29.5
Aug. MIMIC Fine-tuning	384	ViT	85	28.1
Aug. MIMIC Scratch	384	ViT	70	26.1
Orig. MIMIC	512	FPSM (Proposed)	N/A	33.1
Aug. MIMIC Fine-tunning	512	FPSM (Proposed)	N/A	33.37

Table 8: Performance of Various Experiments

References

[1] Dennis Banga and Peter Waiganjo. Abnormality detection in musculoskeletal radiographs with convolutional neural networks (ensembles) and performance optimization. *arXiv preprint arXiv:1908.02170*, 2019.

[2] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020.

[3] Kyungjin Cho, Ki Kim, Yujin Nam, Jiheon Jeong, Jeeyoung Kim, Changyong Choi, Soyoung Lee, Jun Lee, Seoyeon Woo, Gil-Sun Hong, Joon Beom Seo, and Namkug Kim. Chess: Chest x-ray pre-trained model via self-supervised contrastive learning. *Journal of Digital Imaging*, 36, 01 2023.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[5] Aarti Bagul Daisy Ding Tony Duan Hershel Mehta Brandon Yang Kaylie Zhu Dillon Laird Robyn L. Ball Curtis Langlotz Katie Shpanskaya Matthew P. Lungren Andrew Y. Ng Pranav Rajpurkar, Jeremy Irvin. Mura: Large dataset for abnormality detection in musculoskeletal radiographs, 2018.

[6] Tal Ridnik, Hussam Lawen, Asaf Noy, Emanuel Ben Baruch, Gilad Sharir, and Itamar Friedman. Tresnet: High performance gpu-dedicated architecture, 2020.

[7] AFM Saif, Celia Shahnaz, Wei-Ping Zhu, and M Omair Ahmad. Abnormality detection in musculoskeletal radiographs using capsule network. *IEEE Access*, 7:81494–81503, 2019.