Self-feedback Transformer: A Multi-label Diagnostic Model for Real-World Pancreatic Neuroendocrine Neoplasms Data

Mingyu Wang[1,2], Yi Li[1,2], Bin Huang[1,2], Chenglang Yuan[1,2], Yangdi Wang[3], Yanji Luo[3](✉), and Bingsheng Huang[1,2](✉)

[1] Medical AI Lab, School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, Shenzhen, China
huangb@szu.edu.cn

[2] Marshall Laboratory of Biomedical Engineering, Shenzhen University, Shenzhen, China

[3] Department of Radiology, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China
luoyj26@mail.sysu.edu.cn

Abstract. CAD is an emerging field, but most models are not equipped to handle missing and noisy data in real-world medical scenarios, particularly in the case of rare tumors like pancreatic neuroendocrine neoplasms (pNENs). Multi-label models meet the needs of real-world study, but current methods do not consider the issue of missing and noisy labels. This study introduces a multi-label model called Self-feedback Transformer (SFT) that utilizes a transformer to model the relationships between labels and images, and uses a ingenious self-feedback strategy to improve label utilization. We evaluated SFT on 11 clinical tasks using a real-world dataset of pNENs and achieved higher performance than other state-of-the-art multi-label models with mAUCs of 0.68 and 0.76 on internal and external datasets, respectively. Our model has four inference modes that utilize self-feedback and expert assistance to further increase mAUCs to 0.72 and 0.82 on internal and external datasets, respectively, while maintaining good performance even with input label noise ratios up to 40% in expert-assisted mode.

Keywords: Computer Aided Diagnosis · Real-world · Multi-label · Self-feedback Transformer

1 Introduction

Computer-aided Diagnosis (CAD) systems have achieved success in many clinical tasks [5,6,12,17]. Most CAD studies were developed on regular and selected

M. Wang, Y. Li and B. Huang—Contributed equally to this work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43990-2_49.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14226, pp. 521–530, 2023.
https://doi.org/10.1007/978-3-031-43990-2_49

datasets in the laboratory environment, which avoided the problems (data noise, missing data, etc.) in the clinical scenarios [3,6,9,13,18]. In a real clinical scenario, the clinicians generally synthesize all aspects of information, and conduct consultations with Multidisciplinary Team (MDT), to accurately diagnose and plan the treatment [9,10,13]. Real-world studies have received increasing attention [11,16], and it is challenging for the CAD in the real-world scenarios as: 1) Consistent with the clinical workflow, CAD needs to consider multidisciplinary information to obtain multidimensional diagnosis; 2) Due to information collection, storage and manual evaluation, there are missing and noisy medical data. This phenomenon is especially common in rare tumors like pancreatic neuroendocrine neoplasms (pNENs).

In order to overcome above challenges, some studies [3,9,13,18] used multi-label method because of the following advantages: 1) The input of the model is only a single modality such as images, which is easy to apply clinically; 2) The model learns multi-label and multi-disciplinary knowledge, which is consistent with clinical logic; 3) Multi-label simultaneous prediction, which meets the need of clinical multi-dimensional description of patients. For the above advantages, multi-label technology is suitable for real-world CAD. The previous multi-label CAD studies were designed based on simple parameter sharing methods [9,15,20] or Graph Neural Network (GNN) method [2]. The former implicitly interacts with multi-label information, making it difficult to fully utilize the correlation among labels; And the latter requires the use of word embeddings pre-trained on public databases, which is not friendly to many medical domain proper nouns. The generalizability of previous multi-label CAD studies is poor due to these disadvantages. In addition, none of the current multi-label CAD studies have considered the problem of missing labels and noisy labels.

Considering these real-world challenges, we propose a multi-label model named Self-feedback Transformer (SFT), and validate our method on a real-world pNENs dataset. The main contributions of this work are listed: 1) A transformer multi-label model based on self-feedback mechanism was proposed, which provided a novel method for multi-label tasks in real-world medical application; 2) The structure is flexibility and interactivity to meet the needs of real-world clinical application by using four inference modes, such as expert-machine combination mode, etc.; 3) SFT has good noise resistance, and can maintain good performance under noisy label input in expert-assisted mode.

2 Method

Transformer has achieved success in many fields [4,19]. Inspired by DETR [1] and C-Tran [8], we propose a multi-label model based on transformer and self-feedback mechanism. As shown in Fig. 1, 1) image is embedded by Convolutional Neural Network (CNN) firstly; 2) then all labels are embedded and combined with their state embeddings; 3) finally, all embeddings are fed into a transformer, and the output label tokens are fed into Fully Connection (FC) layers for final predictions. Based on this network, we further introduce a self-feedback strategy,

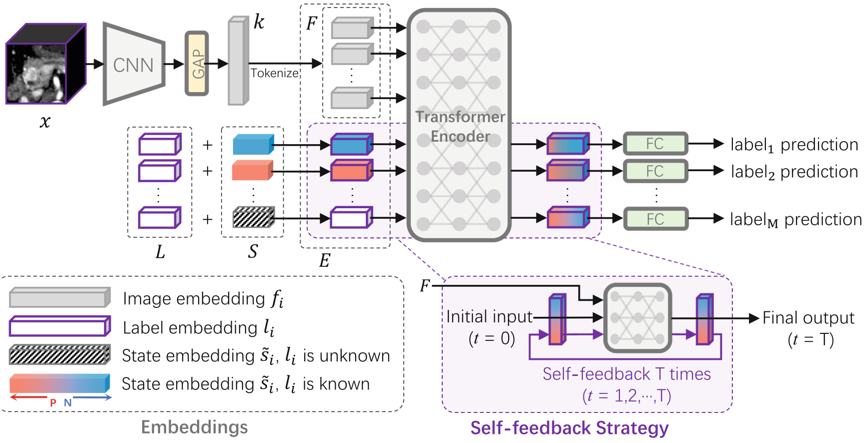

Fig. 1. SFT architecture and illustration of self-feedback strategy.

which allows the label information (including the missing labels) to be reused iteratively for enhancing the utilization of labels.

2.1 Transformer-Based Multi-label Model

Image Embeddings *F*. Given input image $x \in \mathbb{R}^{L \times W \times H}$, the feature vector $k \in \mathbb{R}^C$ is extracted by a CNN after Global Average Pooling (GAP), where the output channel $C = 256$. Then k is split along the channel dimension into N ($N = 8$) sub-feature vectors $F = \{f_1, f_2, \ldots, f_N\}$, $f_i \in \mathbb{R}^d$, $d = C/N$ for tokenization. We choose 3D VGG8, a simple CNN with 8 convolution layers.

Label Embeddings *L*. In order to realize the information interaction among labels, and between labels and image features, we embed labels by an embedding layer. Each image x has M labels, and all labels are embedded into a vector set $L = \{l_1, l_2, \ldots, l_M\}$, $l_i \in \mathbb{R}^d$ by the learnable embedding layer of size $d \times M$.

Soft State Embeddings *S*. There is a correlation between labels, e.g. the lesions with indistinct borders tend to be malignant. Therefore, we hypothesize that the states (GT values) of some labels can be a context for helping predict the remaining labels. We use a soft state embedding method. Specifically, we first embed the positive and negative states into s^p and s^n, both $\in \mathbb{R}^d$, and then the final state embedding $\widetilde{s}_i$ is the weighted sum of s^p and s^n as shown in Equation (1). The state weight w_i^p and w_i^n is the true label value (eg. $w_i^p = 1.0$ when label is positive), where $w_i^p + w_i^n = 1$. For labels with continuous values such as age, the value normalized to $0 \sim 1$ is w_i^p. The $\widetilde{s}_i$ is set as a zero vector for unknown label. $\widetilde{l_i} = \widetilde{s}_i + l_i$ is the final label embedding.

$$\widetilde{s}_i = \begin{cases} w_i^p s^p + w_i^n s^n, & \text{if } l_i \text{ is known} \\ 0, \quad \text{if } l_i \text{ is unknown} & \end{cases}. \tag{1}$$

Multi-label Inference with Transformer Encoder. In a transformer, each output token is the integration of all input tokens. Taking advantage of this structure, we use a transformer encoder to integrate image embeddings and label embeddings, and used the output label tokens to predict label value. Specifically, embedding set $E = \{f_1, f_2, \cdots, f_N, \widetilde{l_1}, \widetilde{l_2}, \cdots, \widetilde{l_M}\}$ are the input tokens, the attention value α and output token e' are computed as follows:

$$\alpha_{ij} = softmax((W^q e_i)^{\mathsf{T}} (W^k e_j)/\sqrt{d}), \tag{2}$$

$$\bar{e}_i = \sum_{j=1}^{M} \alpha_{ij} W^v e_j, \tag{3}$$

$$e_i' = ReLU\left(\bar{e}_i W^r + b_1\right) W^o + \; b_2, \tag{4}$$

where e_i is from E, W^q, W^k and W^v are weight matrices of query, key and value, respectively, W^r and W^o are transformation matrices, and b_1 and b_2 are bias vectors. This update procedure is repeated for L layers, where the e_i' are fed to the successive transformer layer. Finally, all e_i' which are label output tokens are fed into M independent FC layers for predicting value of each label.

2.2 Self-feedback Strategy

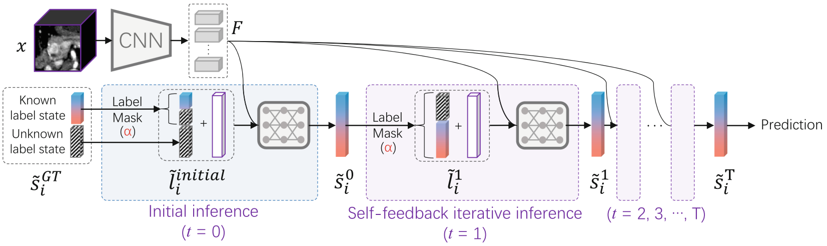

Fig. 2. Illustration of the self-feedback strategy.

The states of unknown labels cannot provide context, thus, the information interaction between known labels and unknown labels may be weaken. To overcome this problem, we propose a Self-feedback Strategy (SFS) inspired by Recurrent Neural Networks (RNN) to enhance the interaction of labels.

Training Progress and Loss Function. As shown in Fig. 2, at time point $t = 0$, the state embedding is initialized to $\widetilde{s}_i^{GT}$ by Ground Truth (GT) value, and the initial label embedding $\widetilde{l}_i^{initial}$ is computed by $\widetilde{l}_i^{initial} = \widetilde{s}_i^{GT} + l_i$. The $\widetilde{l}_i^{initial}$ is combined with f_i as the initial input, and then the output predicted value is converted into state embedding $\widetilde{s}_i^0$ by Equation (1). When t>0, the label embedding $\widetilde{l}_i^t$ is updated iteratively by $\widetilde{l}_i^t = \widetilde{s}_i^{t-1} + l_i$, and then fed into the transformer T times. For classification and regression labels, we use focal cross-entropy loss and L2 loss respectively, and use the method in [7] to auto-weight the loss value of each label. The backpropagation of gradients and parameter updates are performed immediately after calculating the loss at each time point t. In the regular inference phase, the state of all labels is initialized as unknown.

Label Mask Strategy. To avoid predicting with labels' own input state, we use a Label Mask Strategy (LMS) during training phase to randomly mask a certain proportion a of known labels, which causes the labels' states to be embedded as zero vectors. Meanwhile, only the loss of the masked known label is calculated.

3 Experiments and Results

3.1 Dataset and Evaluation

Real-World pNENs Dataset. We validated our method on a real-world pNENs dataset from two centers. All patients with arterial phase Computed Tomography (CT) images were included. The dataset contained 264 and 28 patients in center 1 and center 2, and a senior radiologist annotated the bounding boxes for all 408 and 28 lesions. We extracted 37 labels from clinical reports, including survival, immunohistochemical (IHC), CT findings, etc. Among them, 1)RECIST drug response (RS), 2)tumor shrink (TS), 3)durable clinical benefit (DCB), 4)progression-free survival (PFS), 5)overall survival (OS), 6)grade (GD), 7)somatostatin receptor subtype 2(SSTR2), 8)Vascular Endothelial Growth Factor Receptor 2 (VEFGR2), 9)O6-methylguanine methyltransferase (MGMT), 10)metastatic foci (MTF), and 11)surgical recurrence (RT) are main tasks, and the remaining are auxiliary tasks. 143 and 28 lesions were segmented by radiologists, and the radiomics features of them were extracted, of which 162 features were selected and binarized as auxiliary tasks because of its statistically significant correlation with the main labels. The label distribution and the overlap ratio (Jaccard index) of lesions between pairs of labels are shown in Fig. 3. It is obvious that the real-world dataset has a large number of labels with randomly missing data, thus, we used an adjusted 5-fold cross-validation. Taking a patient as a sample, we chose the dataset from center 1 as the internal dataset, of which the samples with most of the main labels were used as Dataset 1 (219 lesions) and was split into 5 folds, and the remaining samples are randomly divided into the training set Dataset 2 (138 lesions) and the validation set Dataset 3 (51 lesions), the training set and the validation set of the corresponding folds were added during cross-validation, respectively. All samples in Center 2 left as external test set. Details of each dataset are in the Supplementary Material.

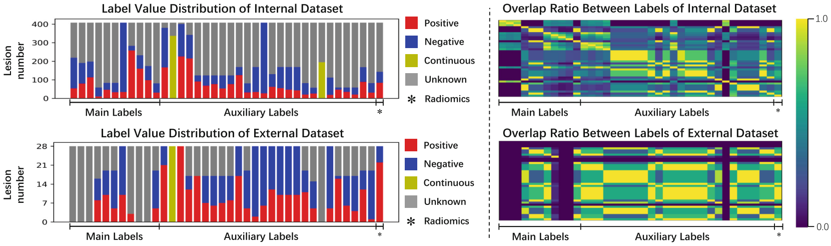

Fig. 3. Label value distribution and overlap ratio of lesions between pairs of labels.

Dataset Evaluation Metrics. We evaluate the performance of our method on the 10 main tasks for internal dataset, and due to missing labels and too few SSTR2 labels, only the performance of predicting RT, PFS, OS, GD, MTF are evaluated for external dataset. We employ accuracy (ACC), sensitivity (SEN), specificity (SPC), F1-score (F1) and area under the receiver operating characteristic (AUC) for each task, and compute the mean value of them (e.g. mAUC).

3.2 Implementation Details

The CT window width and window level were adjusted to 310 and 130 HU refer to [17], and the image inside the bounding box was cropped and scaled to 128×128×64 pixels. The numbers of convolutional kernels of VGG8 are [3, 32, 32, 64, 64, 128, 128, 256]. Transformer encoder contained 2 layers and 8 heads. Layer normalization was used in transformer. The LMS a was set as 0.5, and the training feedback times T was 4. We used Adam optimiser, and used cosine annealing to reduce the learning rate from 1e–4 to 1e–12 over all 200 epochs. Our method was implemented in Pytorch, using an NVIDIA RTX TITAN GPU.

3.3 Comparison and Ablation Experiments

We compared our method with Single-task(ST), Parameters Sharing (PS), ML-Decoder [14] and C-tran [8]. Specifically, a ST model is trained using single label, and PS model uses a FC layer followed by the CNN to predict all labels. It should be noted that the CNN backbone of each method was replaced as 3D VGG8 to ensure fair comparison. In the ablation experiment, we removed the LMS and SFS to analyze their impact. The AUC of each main label is shown in Fig. 4, and the average performance is shown in Table 1. It can be seen that multi-label models is better than that of ST due to using the relationship among labels, and SFT outperform other methods on most tasks and the average performance. The ablation experiments results showed that removing the LMS and SFS components causes performance degradation, indicating the necessity of them.

Table 1. Results of the comparison experiments and the ablation experiments (%). SFT** means SFT w/o SFS and LMS, and SFT* means SFT w/o SFS.

Method	Internal Dataset (219+51 lesions)					External Dataset (28 lesions)				
	mAUC	mACC	mSEN	mSPC	mF1	mAUC	mACC	mSEN	mSPC	mF1
ST	62.56	63.76	64.36	70.09	54.77	64.47	66.98	70.00	70.68	59.57
PS	62.04	64.27	68.45	65.97	60.71	69.85	71.00	**86.50**	64.09	68.40
ML-Decoder	59.36	59.13	**73.81**	60.23	57.26	67.66	65.02	79.50	61.94	60.86
C-tran	62.53	63.05	60.23	73.11	56.14	70.96	**80.05**	63.50	**86.59**	66.54
SFT**	61.01	64.90	60.98	72.77	54.46	62.27	74.35	67.50	76.83	65.46
SFT*	64.39	64.41	71.75	64.48	60.82	74.19	76.12	73.50	79.21	68.53
SFT	**67.76**	**66.53**	69.51	**73.62**	**61.51**	**75.50**	78.08	73.00	80.71	**71.33**

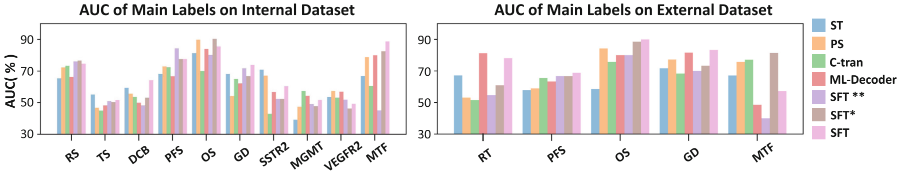

Fig. 4. Predictive performance of different methods on the main tasks.

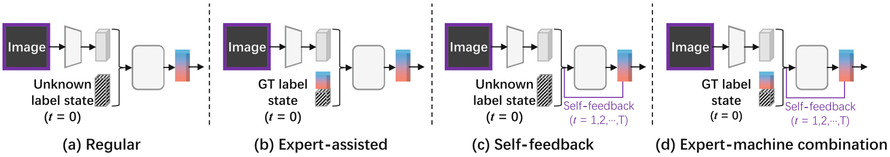

Fig. 5. Different inference modes of SFT.

3.4 Performance of Different Inference Modes

As shown in the Fig. 5, we designed 4 inference modes: 1) Regular, only input images; 2) Expert-assisted (EA), certain information is provided by clinicians; 3) Self-feedback (SF), iteratively inference T times by using prediction; 4) Expert-machine Combination (EMC), expert-assisted and self-feedback inference are both performed. Only the auxiliary labels states were input in EA mode. The results is shown in Table 2. Both SF and EA perform better than regular mode, and EMC outperforms other modes with a mAUC of 0.72 (0.82 on external dataset). We tested the SFT with and without SFS training under different feedback times T in SF mode, and results (Fig. 6.) showed that the performance of SFT with SFS training increases gradually with the increase of T, while the SFT without SFS training has a general trend of decreasing performance after continuous iteration.

Table 2. Performance of different inference modes by using SFT (%).

Mode	Internal Dataset (219+51 lesions)					External Dataset (28 lesions)				
	mAUC	mACC	mSEN	mSPC	mF1	mAUC	mACC	mSEN	mSPC	mF1
Regular	67.76	66.53	69.51	73.62	61.51	75.50	78.08	73.00	80.71	71.33
SF ($T = 1$)	69.94	67.24	**72.24**	72.24	62.29	77.33	81.95	**84.00**	74.13	**78.35**
EA	71.26	68.40	68.42	77.58	62.88	81.30	82.29	66.00	**94.44**	74.44
EMC	**72.22**	**71.26**	66.87	**81.77**	**63.65**	**82.45**	**83.01**	70.00	93.81	75.24

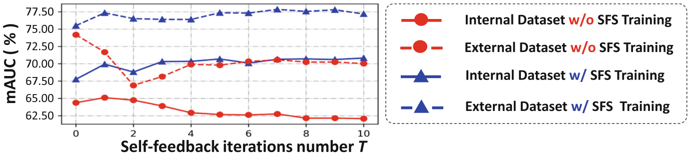

Fig. 6. The relationship between mAUC and the iterations number T in SF mode.

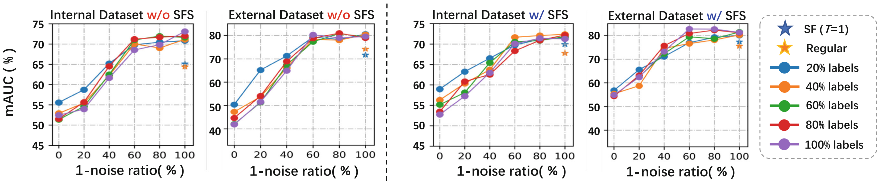

Fig. 7. Performance of SFT under different number of labels and different noise ratios.

3.5 Analysis of Noise Resistance

To explore the noise resistance of SFT in EA mode, we selected 20, 40, 60, 80, and 100 percent of the known labels respectively, and further negated 0, 20, 40, 60, 80, 100 percent of the selected labels to simulate noisy labels. The way to negate a label is to change the label value x to $1 - x$. As shown in Fig. 7, as the noise ratio increases, the performance shows a decreasing trend, and the performance decreases slightly when the noise ratio ≤ 40 %. Finally, it can be observed that the SFT using the SFS training strategy is relatively less affected by noise. When using 100 percent labels, the mAUC of the internal dataset decreased from 0.71 (noise ratio $= 0.0$) to 0.53 (noise ratio $= 1.0$), a decrease of 0.19 (0.26 on external dataset); the corresponding internal mAUC of the model without SFS training decreased by 0.21 (0.38 on external dataset). So SFS training can improve a certain anti-noise ability.

4 Conclusion

We proposed a novel model SFT for multi-label prediction on real-world pNENs data. The model integrates label and image informations based on a transformer encoder, and iteratively uses its own prediction based on a self-feedback mechanism to improve the utilization of missing labels and correlation among labels. Experiment results demonstrated our proposed model outperformed other multi-label models, showed flexibility by multiple inference modes, and had a certain ability to maintain performance when the input context noise was less than 40%.

Acknowledgements. This work was supported by National Natural Science Foundation of China (81971684), Marshall Lab of Biomedical Engineering open fund: Medical-Engineering Project.

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
2. Chen, B., Li, J., Lu, G., Yu, H., Zhang, D.: Label co-occurrence learning with graph convolutional networks for multi-label chest x-ray image classification. IEEE J. Biomed. Health Inform. **24**(8), 2292–2302 (2020)
3. Choi, H., Ha, S., Kang, H., Lee, H., Lee, D.S., Initiative, A.D.N., et al.: Deep learning only by normal brain pet identify unheralded brain anomalies. EBioMedicine **43**, 447–453 (2019)
4. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Eweje, F.R., et al.: Deep learning for classification of bone lesions on routine MRI. EBioMedicine **68**, 103402 (2021)
6. Jiang, Y., et al.: Development and validation of a deep learning CT signature to predict survival and chemotherapy benefit in gastric cancer: a multicenter, retrospective study. Ann. Surg. **274**(6), e1153–e1161 (2021)
7. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7482–7491 (2018)
8. Lanchantin, J., Wang, T., Ordonez, V., Qi, Y.: General multi-label image classification with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16478–16488 (2021)
9. Lin, D., et al.: Application of comprehensive artificial intelligence retinal expert (care) system: a national real-world evidence study. Lancet Digit. Health **3**(8), e486–e495 (2021)
10. Partelli, S., et al.: European cancer organisation essential requirements for quality cancer care (erqcc): pancreatic cancer. Cancer Treat. Rev. **99**, 102208 (2021)
11. Penberthy, L.T., Rivera, D.R., Lund, J.L., Bruno, M.A., Meyer, A.M.: An overview of real-world data sources for oncology and considerations for research. CA: Cancer J. Clin. **72**, 287–300 (2021)
12. Peng, S., et al.: Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: a multicentre diagnostic study. Lancet Digit. Health **3**(4), e250–e259 (2021)
13. Ravizza, S., et al.: Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. Nat. Med. **25**(1), 57–59 (2019)

14. Ridnik, T., Sharir, G., Ben-Cohen, A., Ben-Baruch, E., Noy, A.: ML-Decoder: scalable and versatile classification head. arXiv preprint arXiv:2111.12933 (2021)
15. Shen, S., Han, S.X., Aberle, D.R., Bui, A.A., Hsu, W.: An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. Expert Syst. Appl. **128**, 84–95 (2019)
16. Sherman, R.E., et al.: Real-world evidence—what is it and what can it tell us? (2016)
17. Song, C., et al.: Predicting the recurrence risk of pancreatic neuroendocrine neoplasms after radical resection using deep learning radiomics with preoperative computed tomography images. Ann. Transl. Med. **9**(10), 833 (2021)
18. Tai, Y., Gao, B., Li, Q., Yu, Z., Zhu, C., Chang, V.: Trustworthy and intelligent COVID-19 diagnostic IoMT through XR and deep-learning-based clinic data access. IEEE Internet Things J. **8**(21), 15965–15976 (2021)
19. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems 30 (2017)
20. Zhang, S., et al.: A novel interpretable computer-aided diagnosis system of thyroid nodules on ultrasound based on clinical experience. IEEE Access **8**, 53223–53231 (2020)