Multi-modality Contrastive Learning for Sarcopenia Screening from Hip X-rays and Clinical Information

Qiangguo Jin[1], Changjiang Zou[1], Hui Cui[2], Changming Sun[3], Shu-Wei Huang[4], Yi-Jie Kuo[4,5], Ping Xuan[6], Leilei Cao[1], Ran Su[7], Leyi Wei[8], Henry B. L. Duh[2], and Yu-Pin Chen[4,5(✉)]

[1] School of Software, Northwestern Polytechnical University, Shaanxi, China
[2] Department of Computer Science and Information Technology, La Trobe University, Melbourne, Australia
[3] CSIRO Data61, Sydney, Australia
[4] Department of Orthopedics, Wan Fang Hospital, Taipei Medical University, Taipei, Taiwan
99231@w.tmu.edu.tw
[5] Department of Orthopedics, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan
[6] Department of Computer Science, School of Engineering, Shantou University, Guangdong, China
[7] School of Computer Software, College of Intelligence and Computing, Tianjin University, Tianjin, China
[8] School of Software, Shandong University, Shandong, China

Abstract. Sarcopenia is a condition of age-associated muscle degeneration that shortens the life expectancy in those it affects, compared to individuals with normal muscle strength. Accurate screening for sarcopenia is a key process of clinical diagnosis and therapy. In this work, we propose a novel multi-modality contrastive learning (MM-CL) based method that combines hip X-ray images and clinical parameters for sarcopenia screening. Our method captures the long-range information with Non-local CAM Enhancement, explores the correlations in visual-text features via Visual-text Feature Fusion, and improves the model's feature representation ability through Auxiliary contrastive representation. Furthermore, we establish a large in-house dataset with 1,176 patients to validate the effectiveness of multi-modality based methods. Significant performances with an AUC of 84.64%, ACC of 79.93%, F1 of 74.88%, SEN of 72.06%, SPC of 86.06%, and PRE of 78.44%, show that our method outperforms other single-modality and multi-modality based methods.

Keywords: Sarcopenia screening · Contrastive learning · Multi-modality feature fusion

1 Introduction

Sarcopenia is a progressive and skeletal muscle disorder associated with loss of muscle mass, strength, and function [1,5,8]. The presence of sarcopenia increases the risk of

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14225, pp. 85–94, 2023.
https://doi.org/10.1007/978-3-031-43987-2_9

hospitalization and the cost of care during hospitalization. A systematic analysis of the world's population showed that the prevalence of sarcopenia is approximately 10% in healthy adults over the age of 60 [5,13]. However, the development of sarcopenia is insidious, without overt symptoms in the early stages, which means that the potential number of patients at risk for adverse outcomes is very high. Thus, early identification, screening, and diagnosis are of great necessity to improve treatment outcomes, especially for elderly people.

The development of effective, reproducible, and cost-effective algorithms for reliable quantification of muscle mass is critical for diagnosing sarcopenia. However, automatically identifying sarcopenia is a challenging task due to several reasons. First, the subtle contrast between muscle and fat mass in the leg region makes it difficult to recognize sarcopenia from X-ray images. Second, although previous clinical studies [7,11] show that patient information, such as age, gender, education level, smoking and drinking status, physical activity (PA), and body mass index (BMI), is crucial for correct sarcopenia diagnosis, there is no generalizable standard. It is of great importance to develop a computerized predictive model that can fuse and mine diagnostic features from heterogeneous hip X-rays and tabular data containing patient information. Third, the number of previous works on sarcopenia diagnosis is limited, resulting in limited usable data.

Deep learning attracted intensive research interests in various medical diagnosis domains [17,18]. For instance, Zhang et al. [19] proposed an attention residual learning CNN model (ARLNet) for skin lesion classification to leverage multiple ARL blocks to tackle the challenge of data insufficiency, inter-class similarities, and intra-class variations. For multi-modality based deep learning, PathomicFusion (PF) [3] fused multimodal histology images and genomic (mutations, CNV, and RNA-Seq) features for survival outcome prediction in an end-to-end manner. Based on PF [3], Braman et al. [2] proposed a deep orthogonal fusion model to combine information from multiparametric MRI exams, biopsy-based modalities, and clinical variables into a comprehensive multimodal risk score. Despite the recent success in various medical imaging analysis tasks [2,3,19], sarcopenia diagnosis by deep learning based algorithms is still under study. To the best of our knowledge, recent work by Ryu et al. [12] is the most relevant to our proposed method. Ryu et al. [12] first used three ensembled deep learning models to test appendicular lean mass (ALM), handgrip strength (HGS), and chair rise test (CRT) performance using chest X-ray images. Then they built machine learning models to aggregate predicted ALM, HGS, and CRT performance values along with basic tabular features to diagnose sarcopenia. However, the major drawback of their work lies in the complex two-stage workflow and the tedious ensemble training. Besides, since sarcopenia is defined by low appendicular muscle mass, measuring muscle wasting through hip X-ray images, which have the greatest proportion of muscle mass, is much more appropriate for screening sarcopenia.

In this work, we propose a multi-modality contrastive learning (MM-CL)[1] model for sarcopenia diagnosis from hip X-rays and clinical information. Different from Ryu et al.'s model [12], our MM-CL can process multi-modality images and clinical data and screen sarcopenia in an end-to-end fashion. The overall framework is given in Fig. 1.

[1] Source code will be released at https://github.com/qgking/MM-CL.git.

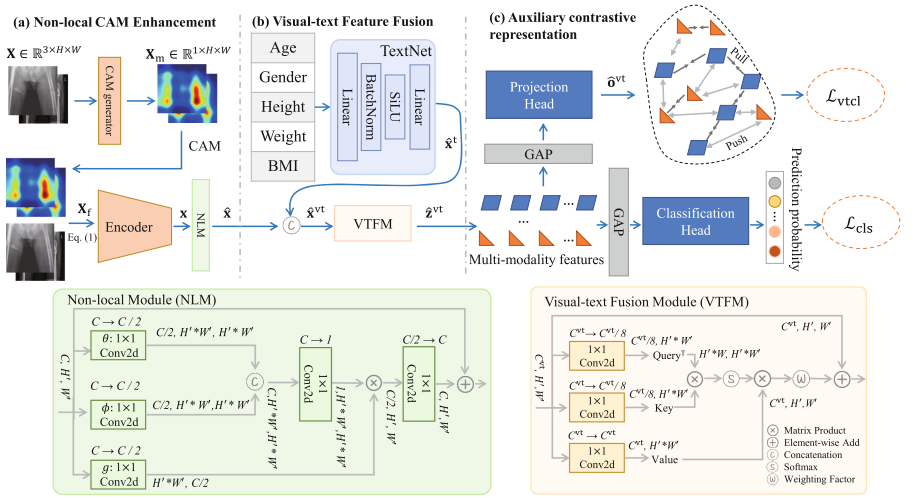

Fig. 1. The framework of our proposed MM-CL. MM-CL is composed of (a) Non-local CAM Enhancement, (b) Visual-text Feature Fusion, and (c) Auxiliary contrastive representation. The output feature size of each block is given in the channel size × height × width ($C \times H \times W$) format. GAP denotes the global average pooling, and CAM denotes the class activation map.

The major components include Non-local CAM Enhancement (NLC), Visual-text Feature Fusion (VFF), and Auxiliary contrastive representation (ACR) modules. Non-local CAM Enhancement enables the network to capture global long-range information and assists the network to concentrate on semantically important regions generated by class activation maps (CAM). Visual-text Feature Fusion encourages the network to improve the multi-modality feature representation ability. Auxiliary contrastive representation utilizes unsupervised learning and thus improves its ability for discriminative representation in the high-level latent space. The main contributions of this paper are summarized as follows. First, we propose a multi-modality contrastive learning model, which enhances the feature representation ability via integrating extra global knowledge, fusing multi-modality information, and joint unsupervised and supervised learning. Second, to address the absence of multi-modality datasets for sarcopenia screening, we select 1,176 patients from the Taipei Municipal Wanfang Hospital. To the best of our knowledge, our dataset is the largest for automated sarcopenia diagnosis from images and tabular information to date. Third, we experimentally show the superiority of the proposed method for predicting sarcopenia from hip X-rays and clinical information.

2 Data Collection

In this retrospective study, we collected anonymized data from patients who underwent sarcopenia examinations at the Taipei Municipal Wanfang Hospital. The data collection was approved by an institutional review board. The demographic and clinical characteristics of this dataset are shown in Table 1. 490 of 1,176 eligible patients who had

developed sarcopenia were annotated as positive, while the remaining 686 patients were labeled as negative. The pixel resolution of these images varies from 2266 × 2033 to 3408 × 3408. Each patient's information was collected from a standardized questionnaire, including age, gender, height, weight, BMI, appendicular skeletal muscle index (ASMI), total lean mass, total fat, leg lean mass, and leg fat. We use 5 numerical variables including age, gender, height, weight, and BMI as clinical information for boosting learning as suggested by the surgeon. To the best of our knowledge, this is the largest dataset for automated sarcopenia diagnosis from images and tabular information to date.

Table 1. Demographic and clinical characteristics of sarcopenia patients.

Characteristics	Type	Entire cohort (n = 1176)
Gender	Male	272 (23.12%)
	Female	904 (76.88%)
Age at diagnosis⋆		71 [63–81]
BMI⋆		22.8 [20.5–25.2]
Height (cm)⋆		155.9 [150.2–162]
Weight (kg)⋆		55 [49.5–63]

Note: ⋆ indicates the median values [interquartile range, 25th–75th percentile].

3 Methodology

As shown in Fig. 1, MM-CL consists of three major components. The Non-local CAM Enhancement module is proposed to force the network to learn from attentional spatial regions learned from class activation map (CAM) [20] to enhance the global feature representation ability. Then, we fuse the heterogeneous images and tabular data by integrating clinical variables through a Visual-text Feature Fusion module. Finally, we present an unsupervised contrastive representation learning strategy to assist the supervised screening by Auxiliary contrastive representation.

3.1 Non-local CAM Enhancement

Considering the large proportion of muscle regions in hip X-ray images, capturing long-range dependencies is of great importance for sarcopenia screening. In this work, we adopt the non-local module [16] (NLM) and propose using coarse CAM localization maps as extra information to accelerate learning. We have two hypotheses. First, the long-range dependency of the left and right legs should be well captured; Second, the CAM may highlight part of muscle regions, providing weak supervision to accelerate the convergence of the network. Figure 1(a) shows the overall structure of the Non-local CAM Enhancement.

CAM Enhancement: First, each training image $\mathbf{X} \in \mathbb{R}^{3\times H\times W}$ is sent to the CAM generator as shown in Fig. 1(a) to generate coarse localization map $\mathbf{X}_{\mathrm{m}} \in \mathbb{R}^{1\times H\times W}$. We use the Smooth Grad-CAM++ [10] technique to generate CAM via the ResNet18 [9] architecture. After the corresponding CAM is generated, the training image $\mathbf{X}$ is enhanced by its coarse localization map $\mathbf{X}_{\mathrm{m}}$ via smooth attention to the downstream precise prediction network. The output image $\mathbf{X}_{\mathrm{f}}$ is obtained as:

$$\mathbf{X}_{\mathrm{f}} = \mathbf{X} \cdot (1 + \mathrm{sigmoid}(\mathbf{X}_{\mathrm{m}})), \tag{1}$$

where $\mathrm{sigmoid}$ denotes the Sigmoid function. The downstream main encoder is identical to ResNet18.

Non-local Module: Given a hip X-ray image $\mathbf{X}$ and the corresponding CAM map $\mathbf{X}_{\mathrm{m}}$, we apply the backbone of ResNet18 to extract the high-level feature maps $\mathbf{x} \in \mathbb{R}^{C\times H'\times W'}$. The feature maps are then treated as inputs for the non-local module. For output $\hat{\mathbf{x}}_i$ from position index i, we have

$$\hat{\mathbf{x}}_i = \sum_{j=1}^{H'W'} a_{ij} g\left(\mathbf{x}_j\right) + \mathbf{x}_i, \quad a_{ij} = \mathrm{ReLU}\left(\mathbf{w}_f^T \,\mathrm{concat}(\theta\left(\mathbf{x}_i\right), \phi\left(\mathbf{x}_j\right))\right), \tag{2}$$

where concat denotes concatenation, $\mathbf{w}_f$ is a weight vector that projects the concatenated vector to a scalar, ReLU is the ReLU function, a_{ij} denotes the non-local feature attention that represents correlations between the features at two locations (i.e., $\mathbf{x}_i$ and $\mathbf{x}_j$), θ, ϕ, and g are mapping functions as shown in Fig. 1(a).

3.2 Visual-Text Feature Fusion

After capturing the global information, we aim to fuse the visual and text features in the high-level latent space. We hypothesize that the clinical data may have a positive effect to boost the visual prediction performance. The overall structure of this strategy is given in Fig. 1(b). We extract the clinical features using a simple network, termed as TextNet. Finally, we propose a visual-text fusion module inspired by self-attention [15] to fuse the concatenated visual-text features.

Visual-Text Fusion Module: In order to learn from clinical data, we first encode 5 numerical variables as a vector and send it to TextNet. As shown in Fig. 1(b), TextNet consists of two linear layers, a batch normalization layer, and a sigmoid linear unit (SiLU) layer. We then expand and reshape the output feature $\hat{\mathbf{x}}^{\mathrm{t}} \in \mathbb{R}^{C^{\mathrm{t}}}$ of TextNet to fit the size of $C \times H' \times W'$. The text and visual representations are then concatenated as $\hat{\mathbf{x}}^{\mathrm{vt}} = \mathrm{concat}(\hat{\mathbf{x}}, \mathrm{reshape}(\hat{\mathbf{x}}^{\mathrm{t}}))$ before sending it to the visual-text fusion module, where $\hat{\mathbf{x}} \in \mathbb{R}^{C\times H'\times W'}$ denotes the output features from Non-local CAM Enhancement. Feature vector $\hat{\mathbf{x}}_i^{\mathrm{vt}} \in \mathbb{R}^{C^{\mathrm{vt}}}$ encodes information about the combination of a specific location i in image and text features with $C^{\mathrm{vt}} = C + C^{\mathrm{t}}$. The visual-text self-attention module first produces a set of query, key, and value by 1×1 convolutional transformations as $\mathbf{q}_i = \mathbf{W}_{\mathrm{q}}\hat{\mathbf{x}}_i^{\mathrm{vt}}$, $\mathbf{k}_i = \mathbf{W}_{\mathrm{k}}\hat{\mathbf{x}}_i^{\mathrm{vt}}$, and $\mathbf{v}_i = \mathbf{W}_{\mathrm{v}}\hat{\mathbf{x}}_i^{\mathrm{vt}}$ at each spatial location i, where $\mathbf{W}_{\mathrm{q}}$, $\mathbf{W}_{\mathrm{k}}$,

and $\mathbf{W}_{\mathrm{v}}$ are part of the model parameters to be learned. We compute the visual-text self-attentive feature $\hat{\mathbf{z}}_i^{\mathrm{vt}}$ at position i as

$$\hat{\mathbf{z}}_i^{\mathrm{vt}} = \sum_{j=1}^{H'W'} s_{ij}\mathbf{v}_j + \mathbf{v}_i, \quad s_{ij} = \mathrm{Softmax}\left(\mathbf{q}_j^{\mathrm{T}} \cdot \mathbf{k}_i\right). \tag{3}$$

The softmax operation indicates the attention across each visual and text pair in the multi-modality feature.

3.3 Auxiliary Contrastive Representation

Inspired by unsupervised representation learning [4], we present a contrastive representation learning strategy that encourages the supervised model to pull similar data samples close to each other and push the different data samples away in the high-level embedding space. By such means, the feature representation ability in the embedding space could be further improved.

During the training stage, given N samples in a mini-batch, we obtain $2N$ samples by applying different augmentations (AutoAugment [6]) on each sample. Two augmented samples from the same sample are regarded as positive pairs, and others are treated as negative pairs. Thus, we have a positive sample and $2N-2$ negative samples for each patch. We apply global average pooling and linear transformations (Projection Head in Fig. 1(c)) to the visual-text embeddings $\hat{\mathbf{z}}^{\mathrm{vt}}$ in sequence, and obtain transformed features $\hat{\mathbf{o}}^{\mathrm{vt}}$. Let $\hat{\mathbf{o}}^{\mathrm{vt+}}$ and $\hat{\mathbf{o}}^{\mathrm{vt-}}$ denote the positive and negative embeddings of $\hat{\mathbf{o}}^{\mathrm{vt}}$, the formula of contrastive loss is defined as

$$\mathcal{L}_{\mathrm{vtcl}} = -\log \frac{\exp\left(sim\left(\hat{\mathbf{o}}^{\mathrm{vt}}, \hat{\mathbf{o}}^{\mathrm{vt+}}\right)/\tau\right)}{\exp\left(sim\left(\hat{\mathbf{o}}^{\mathrm{vt}}, \hat{\mathbf{o}}^{\mathrm{vt+}}\right)/\tau\right) + \sum_{\hat{\mathbf{o}}^{\mathrm{vt-}} \in \mathcal{N}} \exp\left(sim\left(\hat{\mathbf{o}}^{\mathrm{vt}}, \hat{\mathbf{o}}^{\mathrm{vt-}}\right)/\tau\right)}, \tag{4}$$

where $\mathcal{N}$ is the set of negative counterparts of $\hat{\mathbf{o}}^{\mathrm{vt}}$, the $sim(\cdot,\cdot)$ is the cosine similarity between two representations, and τ is the temperature scaling parameter. Note that all the visual-text embeddings in the loss function are ℓ_2-normalized.

Finally, we integrate the auxiliary contrastive learning branch into the main Classification Head as shown in Fig. 1(c), which is a set of linear layers. We use weighted cross-entropy loss $\mathcal{L}_{\mathrm{cls}}$ as our classification loss. The overall loss function is calculated as $\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{cls}} + \beta\mathcal{L}_{\mathrm{vtcl}}$, where β is a weight factor.

4 Experiments and Results

4.1 Implementation Details and Evaluation Measures

Our method is implemented in PyTorch using an NVIDIA RTX 3090 graphic card. We set the batch size to 32. Adam optimizer is used with a polynomial learning rate policy, where the initial learning rate 2.5×10^{-4} is multiplied by $\left(1 - \frac{epoch}{total\_epoch}\right)^{power}$ with $power$ as 0.9. The total number of training epochs is set to 100, and early stopping is adopted to avoid overfitting. Weight factor β is set to 0.01. The temperature constant

τ is set to 0.5. Visual images are cropped to 4/5 of the original height and resized to 224×224 after different online augmentation. The backbone is initialized with the weights pretrained on ImageNet.

Extensive 5-fold cross-validation is conducted for sarcopenia diagnosis. We report the diagnosis performance using comprehensive quantitative metrics including area under the receiver operating characteristic curve (AUC), F1 score (F1), accuracy (ACC), sensitivity (SEN), specificity (SPC), and precision (PRE).

4.2 Quantitative and Qualitative Comparison

We implement several state-of-the-art single-modality (ResNet, ARLNet, MaxNet [2], Support vector machine (SVM), and K-nearest neighbors(KNN)) and multi-modality methods (PF [3]) to demonstrate the effectiveness of our MM-CL. For a fair comparison, we use the same training settings.

Table 2. Sarcopenia diagnosis performance of recently proposed methods.

Method	Modality	AUC (%)	ACC (%)	F1 (%)	SEN (%)	SPC (%)	PRE (%)
ResNet18 [9]	Image	76.86	72.53	64.58	60.56	79.35	69.46
ARLNet [19]	Image	76.73	72.87	65.33	61.72	80.78	69.08
MaxNet [2]	Text	80.05	72.44	58.30	47.05	90.68	78.10
SVM	Text	82.72	73.63	65.47	60.43	83.26	72.08
KNN	Text	80.69	73.21	66.94	65.59	78.80	68.80
PF [3]	Multi	77.88	73.98	67.33	65.11	80.30	70.65
MM-CL	Multi	**84.64**	**79.93**	**74.88**	**72.06**	**86.06**	**78.44**

Table 2 outlines the performance of all methods. As shown, our model achieves the best AUC of 84.64% and ACC of 79.93% among all the methods in comparison.

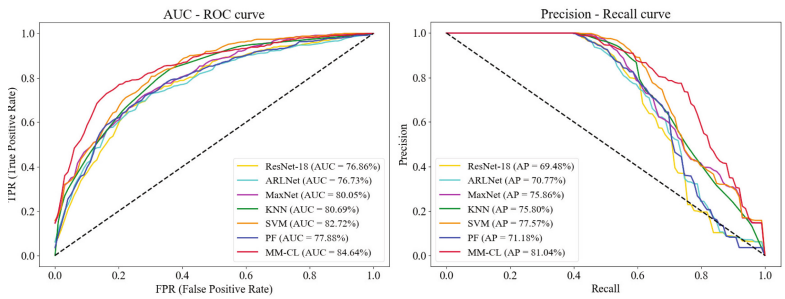

Fig. 2. AUC-ROC (a) and Precision-Recall (b) curves for comparison with state-of-the-art methods.

When compared to the single-modality models, MM-CL outperforms state-of-the-art approaches by at least 6% on ACC. Among all the single-modality models, MaxNet [2], SVM, and KNN gain better results than image-only models. When compared to the multi-modality models, MM-CL also performs better than these methods by a large margin, which proves the effectiveness of our proposed modules.

We further visualize the AUC-ROC and Precision-Recall curves to intuitively show the improved performance. As shown in Fig. 2, the MM-CL achieves the best AUC and average precision (AP), which demonstrates the effectiveness of the proposed MM-CL.

We have three observations: (1) Multi-modality based models outperform single-modality based methods, and we explain this finding that multiple modalities complement each other with useful information. (2) MaxNet [2] gains worse results than traditional machine learning methods. One primary reason is that MaxNet contains a large number of parameters to be learned, the tabular information only includes 5 factors, which could result in overfitting. (3) With the help of NLC, VFF, and ACR, our MM-CL achieves substantial improvement over all the other methods.

4.3 Ablation Study of the Proposed Method

Table 3. Sarcopenia diagnosis performance with ablation studies.

Modality		NLC		VFF	ACR	AUC (%)	ACC (%)	F1 (%)	SEN (%)	SPC (%)	PRE (%)
Image	Text	CAM	NLM								
√						76.86	72.53	64.58	60.56	79.35	69.46
√		√				77.09	73.46	65.55	60.83	82.55	71.53
√		√	√			77.86	73.80	66.23	62.85	81.22	70.93
√	√	√	√	√		84.21	79.16	75.13	76.63	80.69	74.03
√	√	√	√	√	√	84.64	79.93	74.88	72.06	86.06	78.44

We also conduct ablation studies to validate each proposed component i.e., NLC, VFF, and ACR. CAM/NLM of NLC denotes the CAM enhancement/non-local module. Results are shown in Table 3. Utilizing CAM in the network as an enhancement for optimization improves 0.93% for average ACC, when compared to the baseline model (ResNet18). Meanwhile, capturing long-range dependencies via NLM brings improvement on AUC, ACC, F1, and SEN. Equipped with the text information via VFF, our

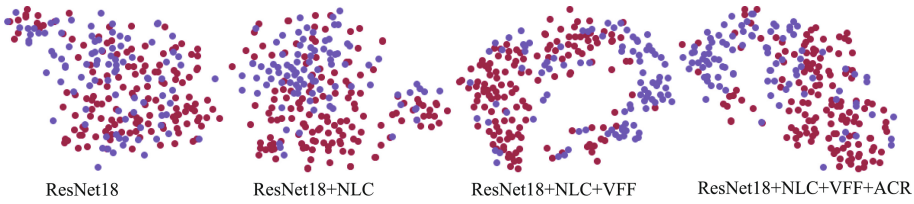

Fig. 3. Visual interpretation of high-level features using t-SNE. The red and blue circles are sarcopenia and non-sarcopenia instances respectively. (Color figure online)

method can lead to significant performance gains on ACC compared with image-only experiments, e.g., 79.16% vs. 73.80%. Lastly, applying ACR to the network improves the average ACC score from 79.16% to 79.93%.

We also visualize the ability of feature representation in the high-level semantic latent feature space before the final classification via t-SNE [14]. As can be seen in Fig. 3, by gradually adding the proposed modules, the feature representation ability of our model becomes more and more powerful, and the high-level features are better clustered.

Our first finding is that fusing visual and text knowledge brings significant improvement, which demonstrates that the extra tabular information could help substantially in learning. Second, incorporating unsupervised contrastive learning in the supervised learning framework could also improve the feature representation ability of the model.

5 Conclusions

In conclusion, we propose a multi-modality contrastive learning model for sarcopenia screening using hip X-ray images and clinical information. The proposed model consists of a Non-local CAM Enhancement module, a Visual-text Feature Fusion module, and an Auxiliary contrastive representation for improving the feature representation ability of the network. Moreover, we collect a large dataset for screening sarcopenia from heterogeneous data. Comprehensive experiments and explanations demonstrate the superiority of the proposed method. Our future work includes the extension of our approach to other multi-modality diagnosis tasks in the medical imaging domain.

Acknowledgment. This work was supported by the Fundamental Research Funds for the Central Universities, the National Natural Science Foundation of China [Grant No. 62201460 and No. 62072329], and the National Key Technology R&D Program of China [Grant No. 2018YFB1701700].

References

1. Ackermans, L.L., et al.: Screening, diagnosis and monitoring of sarcopenia: when to use which tool? Clinical Nutrition ESPEN (2022)
2. Braman, N., Gordon, J.W.H., Goossens, E.T., Willis, C., Stumpe, M.C., Venkataraman, J.: Deep orthogonal fusion: multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 667–677. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_64
3. Chen, R.J., et al.: Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. IEEE Trans. Med. Imaging **41**(4), 757–770 (2020)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR (2020)
5. Cruz-Jentoft, A.J., et al.: Sarcopenia: revised European consensus on definition and diagnosis. Age Ageing **48**(1), 16–31 (2019)

6. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: AutoAugment: learning augmentation strategies from data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 113–123 (2019)
7. Dodds, R.M., Granic, A., Davies, K., Kirkwood, T.B., Jagger, C., Sayer, A.A.: Prevalence and incidence of sarcopenia in the very old: findings from the Newcastle 85+ study. J. Cachexia, Sarcopenia Muscle **8**(2), 229–237 (2017)
8. Giovannini, S., et al.: Sarcopenia: diagnosis and management, state of the art and contribution of ultrasound. J. Clin. Med. **10**(23), 5552 (2021)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
10. Omeiza, D., Speakman, S., Cintas, C., Weldermariam, K.: Smooth grad-CAM++: an enhanced inference level visualization technique for deep convolutional neural network models. arXiv preprint arXiv:1908.01224 (2019)
11. Pang, B.W.J., et al.: Prevalence and associated factors of Sarcopenia in Singaporean adults-the Yishun Study. J. Am. Med. Direct. Assoc. **22**(4), e1-885 (2021)
12. Ryu, J., Eom, S., Kim, H.C., Kim, C.O., Rhee, Y., You, S.C., Hong, N.: Chest X-ray-based opportunistic screening of sarcopenia using deep learning. J. Cachexia, Sarcopenia Muscle **14**(1), 418–428 (2022)
13. Shafiee, G., Keshtkar, A., Soltani, A., Ahadi, Z., Larijani, B., Heshmat, R.: Prevalence of sarcopenia in the world: a systematic review and meta-analysis of general population studies. J. Diab. Metab. Disord. **16**(1), 1–10 (2017)
14. Van Der Maaten, L.: Accelerating t-SNE using tree-based algorithms. J. Mach. Learn. Res. **15**(1), 3221–3245 (2014)
15. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
16. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)
17. Yan, K., Guo, Y., Liu, B.: Pretp-2l: identification of therapeutic peptides and their types using two-layer ensemble learning framework. Bioinformatics **39**(4), btad125 (2023)
18. Yan, K., Lv, H., Guo, Y., Peng, W., Liu, B.: samppred-gat: prediction of antimicrobial peptide by graph attention network and predicted peptide structure. Bioinformatics **39**(1), btac715 (2023)
19. Zhang, J., Xie, Y., Xia, Y., Shen, C.: Attention residual learning for skin lesion classification. IEEE Trans. Med. Imaging **38**(9), 2092–2103 (2019)
20. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)