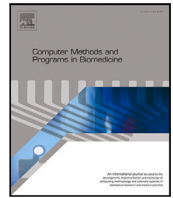




Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: <https://www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine>



Self-supervised multi-modality learning for multi-label skin lesion classification

Hao Wang^{a,c}, Euijoon Ahn^b, Lei Bi^{c,*}, Jinman Kim^a

^a School of Computer Science, Faculty of Engineering, The University of Sydney, Sydney, NSW 2006, Australia

^b College of Science and Engineering, James Cook University, Cairns, QLD 4870, Australia

^c Institute of Translational Medicine, National Center for Translational Medicine, Shanghai Jiao Tong University, Shanghai, China

ARTICLE INFO

Keywords:

Skin lesion classification
Self-supervised learning
Multi-label learning
Multi-modality learning

ABSTRACT

Background: The clinical diagnosis of skin lesions involves the analysis of dermoscopic and clinical modalities. Dermoscopic images provide detailed views of surface structures, while clinical images offer complementary macroscopic information. Clinicians frequently use the seven-point checklist as an auxiliary tool for melanoma diagnosis and identifying lesion attributes. Supervised deep learning approaches, such as convolutional neural networks, have performed well using dermoscopic and clinical modalities (multi-modality) and further enhanced classification by predicting seven skin lesion attributes (multi-label). However, the performance of these approaches is reliant on the availability of large-scale labeled data, which are costly and time-consuming to obtain, more so with annotating multi-attributes.

Methods: To reduce the dependency on large labeled datasets, we propose a self-supervised learning (SSL) algorithm for multi-modality multi-label skin lesion classification. Compared with single-modality SSL, our algorithm enables multi-modality SSL by maximizing the similarities between paired dermoscopic and clinical images from different views. We introduce a novel multi-modal and multi-label SSL strategy that generates surrogate pseudo-multi-labels for seven skin lesion attributes through clustering analysis. A label-relation-aware module is proposed to refine each pseudo-label embedding, capturing the interrelationships between pseudo-multi-labels. We further illustrate the interrelationships of skin lesion attributes and their relationships with clinical diagnoses using an attention visualization technique.

Results: The proposed algorithm was validated using the well-benchmarked seven-point skin lesion dataset. Our results demonstrate that our method outperforms the state-of-the-art SSL counterparts. Improvements in the area under receiver operating characteristic curve, precision, sensitivity, and specificity were observed across various lesion attributes and melanoma diagnoses.

Conclusions: Our self-supervised learning algorithm offers a robust and efficient solution for multi-modality multi-label skin lesion classification, reducing the reliance on large-scale labeled data. By effectively capturing and leveraging the complementary information between the dermoscopic and clinical images and interrelationships between lesion attributes, our approach holds the potential for improving clinical diagnosis accuracy in dermatology.

1. Introduction

Melanoma is one of the deadliest forms of skin cancer in the world, and the number of incidences has been increasing steadily in recent years [1]. Early diagnosis is particularly important as melanoma can be cured with simple excision [2]. In clinical practice, the suspected skin lesions are assessed by examining clinical images and dermoscopy images [3]. Clinical images are acquired by a digital camera, showing geometry and color of the skin lesion. On the other hand, dermoscopy

images are acquired with a dermatoscope, providing better view of the skin lesion subsurface structures. These combined imaging modalities provide complementary information to assist dermatologists in the diagnosis. Seven-point checklist [4] is the most commonly used algorithm for diagnosis, where each attribute in the checklist, as denoted in Fig. 1, is assigned with 1 point to rate the likelihood of having a melanoma. The examined lesion is diagnosed as melanoma when the sum of the score surpasses a given threshold, e.g., greater than three [5]. Fig. 1

* Corresponding author.

E-mail addresses: hwan7885@uni.sydney.edu.au (H. Wang), euijoon.ahn@jcu.edu.au (E. Ahn), lei.bi@sjtu.edu.cn (L. Bi), jinman.kim@sydney.edu.au (J. Kim).

<https://doi.org/10.1016/j.cmpb.2025.108729>

Received 12 August 2024; Received in revised form 10 March 2025; Accepted 16 March 2025

Available online 1 April 2025

0169-2607/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).







Modality	Dermoscopy Images			
	Clinical Images			
Diagnosis	Skin Lesion Type	Melanoma	Nevus	Seborrheic Keratosis
Seven-point Checklist	Blue Whitish Veil	Present	Absent	Absent
	Dots and Globules	Irregular	Regular	Absent
	Pigmentation	Absent	Absent	Regular
	Pigment Network	Absent	Typical	Absent
	Regression Structures	Combinations	Absent	Absent
	Streaks	Irregular	Absent	Absent
	Vascular Structures	Within Regression	Absent	Arborizing

Fig. 1. Examples of multi-modal skin lesion images comprising of clinical images and dermoscopy images. Each pair of images is labeled with seven-point checklist and its diagnosis (Melanoma, Nevus or Seborrheic Keratosis).

gives three example studies showing clinical and dermoscopic images and their corresponding seven attributes. However, such manual examination is time consuming and difficult, particularly for untrained or less experienced dermatologists [6]. Moreover, dermatologists are consistently in short supply in rural areas, and consultation costs are rising [7]. Motivated by these difficulties, computer-aided diagnosis (CAD) systems have been developed to automate such manual process and provide second opinions to clinicians. In recent years, many CAD systems based on Convolutional Neural Networks (CNNs) have been successful in skin lesion image segmentation and classification related tasks [8,9]. For example, Yu et al. [8] proposed to use a deep CNN with over 50 layers to acquire richer and more discriminative skin features. However, these methods are limited to using dermoscopy images (a single modality), and therefore discarding useful information contained in the clinical images. More recently, researchers [1,10] have attempted to develop multi-modality fusion networks to simultaneously learn image features from dermoscopic and clinical images. A common approach in these networks is to use CNNs for each imaging modality separately, with the subsequent step of concatenating the feature outputs from each CNN. Although these studies exploited complementary information from modalities, they directly predicted the diagnosis of melanoma from images without inferring the seven-point checklist, having the likelihood of making a misdiagnosis [1].

Multi-label classification (MLC), where an image can be assigned to multiple labels simultaneously, is a pertinent approach to solve the issue. In MLC setting, the seven attributes are considered as seven labels with the diagnosis as the eighth label for each image. It then learns the interrelationships among the seven attributes and the 8th diagnosis label. For instance, Fu et al. [11] proposed a graph-based model to leverage the interrelationships in the seven-point checklist to improve skin lesion classification. However, the performance of these methods is highly dependent on the availability of large-scale labeled training data. Unfortunately, there is a scarcity of large annotated multi-modal skin lesion datasets due to the expensive data acquisition and annotation process. Earlier studies [12,13] have mitigated this issue by adopting transfer learning such that models, pre-trained with ImageNet [14], can be fine-tuned on the target medical imaging dataset. Despite the

effectiveness of transfer learning, there still exists large domain shifts between sources, e.g., ImageNet and skin lesion images [15].

An alternative approach is to use a self-supervised learning (SSL) approach to learn meaningful features using only unlabeled data. Many recent SSL approaches [16–21] have been successfully introduced for various tasks in both natural and medical image analysis. For example, Chen et al. [16] designed an SSL approach that maximizes similarities between augmented views of the same image and minimizes the similarities with other images. Doing so provided a more robust transfer ability than approaches that used ImageNet-pre-trained weights. Existing SSL-based approaches, however, are not optimized for multi-modality and multi-label skin lesion images, as they do not consider (1) how image features from different modalities could be fused to complement each other; (2) and how to enable the model to learn interrelationships between skin lesion attributes.

In an attempt to address the issues discussed above, we propose a Self-supervised learning framework for Multi-Modality Multi-label skin lesion classification (SM3). Our contributions are summarized as follows:

- We introduce a new SSL pre-training algorithm in which the image features of different modalities are contrasted. Our innovation is to exploit the inherent complementary information within the dermoscopic and clinical images of the same patient which is expected to possess the highest mutual information compared to random pairing between different patients. This is in contrast to existing SSL pre-training methods which are exclusively designed to work with a single imaging modality. Our pre-training task facilitates the fusion of multi-modal image features, thereby fostering improved discrimination and differentiation among various skin lesion classes.
- We further innovate in SSL pre-training algorithm designed for multi-label learning. Our approach enables the learning of correlations among the seven attributes and the final diagnosis without using labeled data. To achieve this, we propose a pseudo-multi-labeling scheme (we refer to the seven attributes as seven additional labels from now on) that is constructed by multiple cluster analysis for each label, with the centroid of each resulting

cluster representing a class label. These pseudo-multi-labels are used to facilitate multi-label self-supervised learning during the pre-training.

- We improve the learning of correlations by introducing a label-relation-aware module. It uses the distribution of features within pseudo-multi-labels to better capture interrelationships between them. We also visualize the learned relationships.

2. Related works

2.1. Deep learning based skin lesion classification

Skin lesion classification has seen significant advancements with the introduction of CNNs [22] which has become the preferred technology for developing CAD systems [23–25]. In skin lesion classification, CNNs have shown superior performance compared to traditional methods based on handcrafted features [12,26]. Researchers have extended this approach by using more advanced techniques, such as deep CNNs with more layers [8], attention learning mechanism [9], and regularization strategies for small and unbalanced datasets [27]. Furthermore, researchers investigated multi-class skin lesion classification, which has fine-grained sub-classes for malignant and benign lesions. For example, Qian et al. [28] proposed to use a grouping of multi-scale attention blocks and class-specific loss weighting to solve the category imbalance issue. Hsu et al. [29] used hierarchy-aware contrastive learning to improve model performance by reducing penalties for outputs that correctly predict major class (e.g., malignant) but misclassify sub-classes (e.g., basal cell carcinoma, melanoma, and squamous cell carcinoma).

However, these methods often overlooked clinical images, which are crucial for precise decision making. To address this limitation, Ge et al. [10] proposed a multi-modality learning method that utilized both dermoscopy and clinical images by applying separate CNNs for each modality. Subsequent research has built on this approach by incorporating multi-scale feature fusion modules [3,30] and adversarial learning with attention mechanisms [31] to capture both correlated and complementary information from two image modalities. In addition, researchers have explored the detection of dermoscopic attributes (multi-label classification) from the seven-point checklist to improve the classification performance [1]. For example, Fu et al. [11] proposed a graph-based model to capture the interrelationships between different labels. Similarly, Tang et al. [32] developed a two-stage learning scheme, where dermoscopy and clinical image features were integrated in the first stage, which were then integrated with patient metadata in the second stage to capture correlations between labels. However, to ensure the performance, these supervised methods required large-scale labeled data.

2.2. SSL in medical imaging

SSL has emerged as a promising alternative to alleviate the problem of expensive and time-consuming annotation processes [33]. This is particularly relevant in the context of medical imaging, where annotated data are scarce due to the complicated data acquisition procedures [34]. One of common SSL approaches is to use Contrastive Learning [35] which uses a pretext task that maximizes similarities between similar data instances while minimizing them with dissimilar instances. For example, it maximizes the similarities between augmented views (e.g., rotated, masked views) of the same image and minimize similarities with augmented views of different images [16–19]. For example, Azizi et al. [34] trained a model on an unlabeled dataset and used a pretext task based on multiple images of the same clinical case to improve skin lesion classification. Öztürk et al. [36] proposed a deep clustering approach for melanoma detection that overcomes class imbalance issues, and this was achieved by maximizing cluster separation in the embedding space, where existing methods

focus on minimizing classification errors. Recently, there was a systematic review that evaluated the use of various SSL algorithms for skin lesion classification [37]. Other studies focus on using SSL to address common skin lesion classification challenges, e.g., long-tail out-of-distribution problem [38] and light weight models [39]. However, all these SSL-based methods focused on using a single medical imaging modality and cannot be directly applied to multi-modality learning. While multi-modality learning can be implemented through the simple concatenation of multi-modality image features, this is not optimal to derive complementary information in an SSL setting. As also pointed by Li et al. [40], naïve concatenation is not an efficient way and could heavily decrease the model performance due to the domain differences between different imaging modalities. To address this issue, Atli et al. [41] proposed channel-mixed Mamba blocks with a spiral-scan trajectory for modality synthesis, aiming to bridge the domain shifts between multi-contrast MRI and MRI-CT. Zhang et al. [42] proposed to align multi-modal feature maps by designing a new contrastive loss to enforce the network to focus on the similarities of segmentation masks from paired modalities as well as dissimilarities of unpaired multi-modal data. Huang et al. [43] proposed an SSL algorithm for four-modality ultrasound learning, where Mean Absolute Error across different modalities was minimized to ensure that high-level image features extracted from different modalities can be similar. However, these methods were designed for specific medical areas and cannot be applied to solve the multi-modality and multi-label problem in skin lesion classification.

2.3. MLC in medical imaging

MLC considers a scenario where a set of class labels is assigned to a single data instance. In this context, prior research has primarily concentrated on devising models for understanding relationships between labels [44], including approaches such as one-vs-all classifiers [45], tree structures [46], and graph structures [47]. Medical imaging field also has adopted MLC, as exemplified by Guan et al.'s use of a residual attention learning framework for chest X-ray image classification. It assigned different weights to different spatial regions based on multi-labels [48]. HydraViT [49] integrated a transformer backbone with a multi-branch output module to separately model disease-specific features and their co-occurrence for multi-label chest X-ray classification. Liu et al. [50] enhanced model robustness by maintaining the consistency of relationships among different samples under perturbations. Zhang et al. [51] employed a triplet attention network with a Transformer to make use of multi-labels together with spatial and category attention features. However, none of these works attempted to solve the MLC task with SSL which requires to design a pretext task to learn label interrelationships.

3. Materials and method

3.1. Materials

We used the Derm7pt [1] dataset for our experiments. It is currently the only publicly available dataset that provides aligned multi-modality and multi-label skin lesion images. It contains a total of 1,011 studies. The dataset is divided into 413 studies for training, 203 studies for validation, and 395 studies for testing, according to [1]. Each study contains a pair of dermoscopy and clinical images, a diagnosis (DIAG) label, and seven-point checklist labels. The DIAG label consists of 5 types of skin lesions including Basal Cell Carcinoma (BCC), Nevus (NEV), Melanoma (MEL), Miscellaneous (MISC), and Seborrheic Keratosis (SK). The seven-point checklist labels contain Pigment Network (PN), Blue Whitish Veil (BWV), Vascular Structures (VS), Pigmentation (PIG), Streaks (STR), Dots and Globules (DaG), and Regression Structures (RS). Each seven-point checklist label has different number of classes including Absent (ABS), Typical (TYP), Atypical (ATP), Present (PRS), Regular (REG), and Irregular (IR). The size of dermoscopy images varies from 474×512 to 532×768 pixels and the clinical images vary from 480×512 to 532×768 pixels.

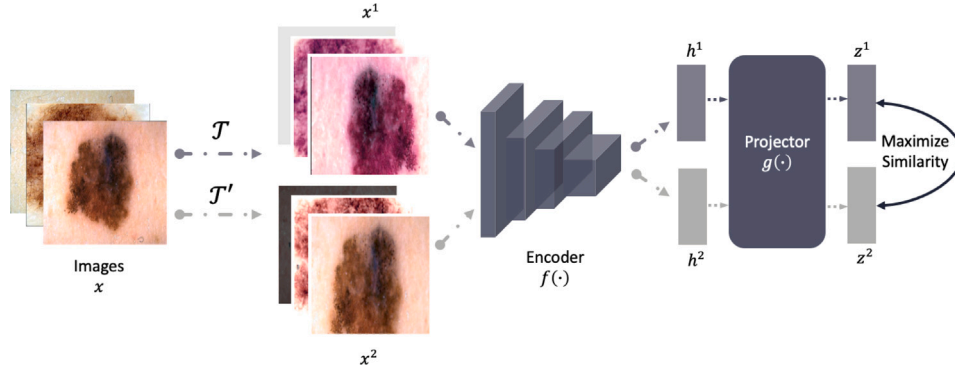


Fig. 2. Pipeline of SimCLR applied to skin lesion classification.

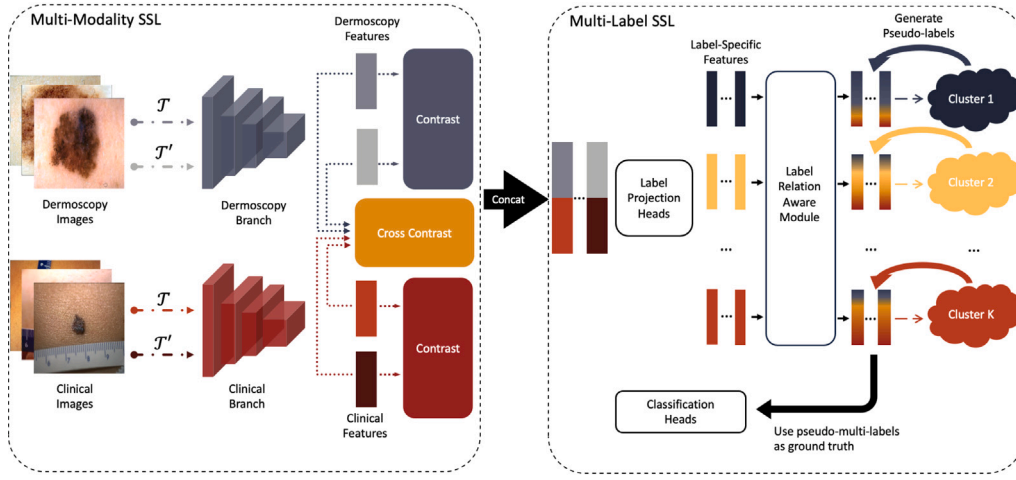


Fig. 3. A schematic of SM3. The multi-modality SSL component utilizes two separate branches to extract modality-specific features using SimCLR. A multi-modal fusion is then enabled by contrasting paired dermoscopic and clinical images. In the multi-label SSL component, the concatenated image features are projected into label-specific features, and a label-relation-aware module is applied to learn label correlations in a self-supervised manner. Each label-specific feature is then grouped into similar features and used to generate pseudo-multi-labels. These are used to update the classification heads.

3.2. Preliminaries: SSL pre-training strategy

In this work, we used a SSL method SimCLR [16] as the base pre-training strategy. The workflow of SimCLR is shown in Fig. 2. Firstly, two separate data augmentation sets $\{\mathcal{T}, \mathcal{T}'\}$ from the same family of augmentations (including random resized cropping, color jitter, random horizontal flip, and random Gaussian blur) randomly transform any given image sample x into two augmented views x^1 and x^2 , which are considered as a *positive pair*. Then, an encoder network $f(\cdot)$ extracts image features h^1, h^2 from augmented views, respectively. The choice of encoder network is flexible and can be any CNN architecture. Afterward, a projection head $g(\cdot)$ maps learned image representations into a latent space z . We used the Multi-Layer Perceptron (MLP) as our projection head. Lastly, the contrastive loss is applied to the latent space z , aiming to maximize the similarities between positive pairs. The loss function for image x_i is defined as:

$$\mathcal{L}_{NT-Xent,i} = -\log \frac{e(\sigma(z_i^1, z_i^2)/\tau)}{\sum_{j=1}^N e(\sigma(z_i^1, z_j^2)/\tau) + 1_{[i \neq j]} e(\sigma(z_i^1, z_j^1)/\tau)} \quad (1)$$

where N denotes the batch size, τ is a temperature hyperparameter. $1_{[i \neq j]} \in \{0, 1\}$ is an indicator function, which was set to 1 if and only if $i \neq j$. $e(\cdot)$ is the exponentiation operation, and $\sigma(u, v) = u^T v / \|u\| \|v\|$ denotes the cosine similarity function.

3.3. Overview

The overview of our method is shown in Fig. 3. Initially, modality-specific features from dermoscopic and clinical images were extracted using SimCLR. Subsequently, we pre-trained the multi-modality models by maximizing similarities between paired multi-modality images of the same patient. Following this, the extracted image features were projected into distinct label-specific embedding spaces. A label-relation-aware module was then used to learn correlation between labels. Lastly, we channeled the outputs into clusters, generating pseudo-multi-labels for SSL multi-label pre-training. Our pre-training process consists of two stages: first, we performed multi-modality SSL to obtain a trained multi-modality image feature extractor; after that, we conducted the multi-label SSL with the frozen feature extractor.

3.4. Self-supervised multi-modality learning

Given pairs of dermoscopy and clinical images, we utilized separate CNNs, named dermoscopy branch and clinical branch, to extract corresponding image modality features. These two branches have identical architecture but independent weight updates, which helps to optimize each branch for different image modalities. To enable efficient multi-modality representation learning, three pretext tasks are defined. The first and second pretext task are to apply SimCLR in each model branch to extract specific features from corresponding modalities. We defined the loss functions using Eq. (1) as follows:

$$\mathcal{L}_{derm} = \mathcal{L}_{NT-Xent} \quad (2)$$

$$\mathcal{L}_{clinic} = \mathcal{L}_{NT-Xent} \quad (3)$$

where z in \mathcal{L}_{derm} comes from dermoscopy images, whereas z in \mathcal{L}_{clinic} is derived from clinical images. We used these loss functions to solve the first and second pretext tasks. In addition, we propose a third task that jointly utilizes both dermoscopy and clinical images, allowing complementary representation learning of the two modalities for multi-modal fusion. The intuition behind our design is that pairs of multi-modality images have more similarities than others, i.e., dermoscopy and clinical images of the same case have maximum mutual information under different augmented views. We implemented this idea by (1) introducing two extra projection heads to map extracted dermoscopy and clinical features into a shared embedding space and (2) maximizing the agreement between randomly augmented views of the same case but different modality data sample by the modified contrastive loss:

$$\begin{aligned} \mathcal{L}_{mm} = & -\log \frac{e\left(\sigma\left(z_i^1, z_i^1\right) / \tau\right)}{\sum_{j=1}^N e\left(\sigma\left(z_i^1, z_j^1\right) / \tau\right) + 1_{[i \neq j]} e\left(\sigma\left(z_i^1, z_j^1\right) / \tau\right)} \\ & -\log \frac{e\left(\sigma\left(z_i^1, z_i^2\right) / \tau\right)}{\sum_{j=1}^N e\left(\sigma\left(z_i^1, z_j^2\right) / \tau\right) + 1_{[i \neq j]} e\left(\sigma\left(z_i^1, z_j^2\right) / \tau\right)} \end{aligned} \quad (4)$$

where z is computed from dermoscopy images while z' is calculated from clinical images. We applied these three tasks to train the model by adopting multi-task learning and defined the final loss function as:

$$\mathcal{L}_{ssl} = \mathcal{L}_{derm} + \mathcal{L}_{clinic} + \mathcal{L}_{mm} \quad (5)$$

3.5. Self-supervised multi-label learning

Naïve solution. Since multiple label predictions are derived from the single image representation and the number of classes is different for each label, we adopted separate classifiers for every label prediction. The classifier $h(\cdot)$ was built by a label projection head $p(\cdot)$ and a classification head $q(\cdot)$ such that $h(\cdot) = q(p(\cdot))$. Here, $p(\cdot)$ consists of an MLP aiming to filter label-specific features and $q(\cdot)$ is a single fully-connected (FC) layer to make final predictions. To enable self-supervised learning of the multi-label classifier, we used the clustering algorithm to generate pseudo-label. We independently iterated the clustering process for K times to generate K labels, with K equals to the number of labels in the dataset. We then utilized these generated pseudo-multi-labels to update the parameters of the classifier by optimizing the cross-entropy loss function which is defined as follows:

$$\mathcal{L}_{ce}\left(x_i, y_i\right)=\sum_{k=1}^K \text { CrossEntropy }\left(h_k\left(x_i\right), y_{i, k}\right) \quad (6)$$

where x_i is the i_{th} image and y_i is the corresponding pseudo-multi-labels containing $\{y_{i,1}, \dots, y_{i,K}\}$. $h_k(\cdot)$ denotes the k_{th} classifier.

Label-relation-aware solution. The above naïve solution, however, yielded degraded results in our preliminary experiments, which overlooked the relationships between labels. We therefore further refined each label embedding by understanding the relationships between other embeddings using an attention mechanism [52]. Self-attention was applied to access all label embeddings before the clustering analysis, ensuring each label embedding was refined by the weighted influence of other embeddings based on their importance. Formally, we inserted self-attention method $W(\cdot)$ before $q(\cdot)$ and fed outputs of all $p(\cdot)$ into it. Then, we rewrote the classifier function as:

$$h(\cdot)=q\left(W\left(p_1(\cdot), \dots, p_K(\cdot)\right)\right) \quad (7)$$

where $\{p_1, \dots, p_K\}$ are label projection heads from K classifiers. Here, the features specific to each label only contain information about that label. However, our label-relation-aware module come into play to connect all labels information and learn the relationships between them. As a result, the prediction of a single label involves contributions from other label information. In this work, we used $W(\cdot)$ as the encoder layer of Transformer [52].

3.6. Inference pipeline

We adopted a vanilla implementation of multi-modality and multi-label classification scheme (as shown in Fig. 4) to emphasize the effectiveness of our proposed SSL pre-training method. Following the design of multi-modality model fusion in preliminary work EmbeddingNet [53], we applied two separate branches to extract dermoscopic and clinical features. These two modality features were then concatenated and fed into a subsequent classifier to make the final predictions. This model was referred as the Baseline in Section 4. Experiments, and we initialized both the branches and the classifier using our SM3 pre-trained weights.

4. Experiments

4.1. Experiment configurations

We used a deep learning library PyTorch [54] to implement our algorithm. All the experiments were conducted on two NVIDIA RTX 3080Ti 12 GB GPUs. For fair comparisons, we used ResNet-50 [25] as the CNN backbone since ResNet-50 has been commonly applied for various skin lesion classification and segmentation tasks including [3,11,31,32]. The output dimension of the projectors in contrastive learning was set to 128. For multi-label SSL, we applied k-means [55] as the clustering algorithm, and ran it for $K = 8$ times, with the number of clusters in each iteration determined by the number of classes of the corresponding label. The label projection head was a single-layer MLP with a dimension of 512. We used a single Transformer encoder layer with a head of 1, a feed forward dimension of 128 and a dropout rate of 0.1 for our label-relation-aware module. All these hyperparameters were chosen based on our empirical studies. We resized all of images into 224×224 . We followed the existing literature [3,11,31,32] and evaluated the proposed method with the official data split [1]. Code is available at <https://github.com/Dylan-H-Wang/skin-sm3>.

4.1.1. Pre-training

For multi-modality SSL, we set the batch size to 96, learning rate to $1e-6$, the number of epochs to 400, and temperature τ to 0.1. For multi-label SSL, we used batch size of 256, learning rate of $1e-4$ and the number of epochs as 150. AdamW [56] was used as the optimizer with default parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay = 0.01). We followed the commonly used SimCLR data augmentation design during the pre-training.

4.1.2. Linear probing and fine-tuning

We adopted linear probing protocol where CNNs were frozen and only the classifier was fine-tuned [17]. We set learning rate as $1e-3$, batch size as 128 and the number of epochs as 50. The optimizer was AdamW with default parameters as in the pre-training. We used data augmentations including random resized crop and random horizontal flip. We also evaluated the non-linear quality of learned representations [57] such that CNNs were initialized with pre-trained weights and fine-tuned with all layers. We set learning rate as $1e-4$, batch size as 64 and the number of epochs as 50. The settings of optimizer and data augmentation were the same as linear probing experiments.

4.2. Evaluation setups

4.2.1. Performance metrics

The model performances were evaluated using metrics including area under receiver operating characteristic curve (AUC), sensitivity (Sens), specificity (Spec), and precision (Prec). We computed the evaluation metrics for multi-class labels using a one-vs-rest approach. For each class, we treated it as the positive class and combined all other classes as the negative class, then calculated the AUC, sensitivity, specificity, and precision accordingly. For clarity, we present only the melanoma class results of the DIAG label and the class results for the seven attributes that increase the chance of melanoma.

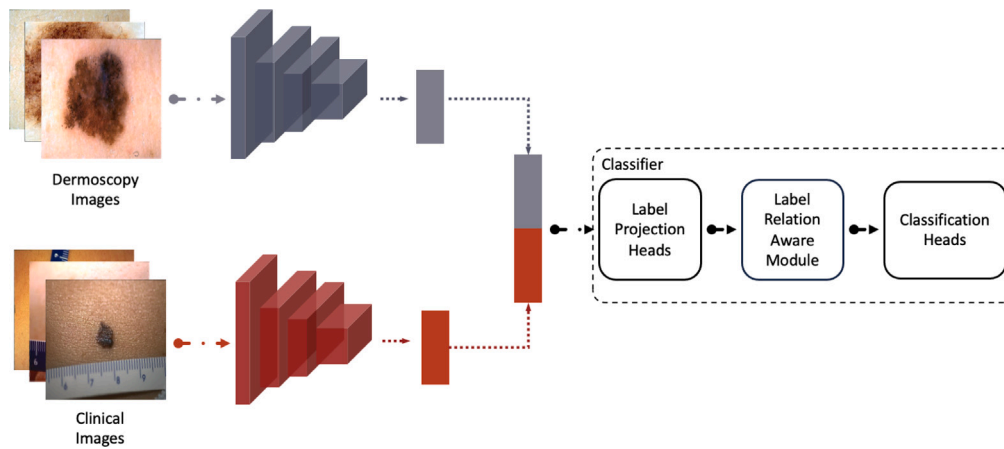


Fig. 4. Pipeline of the inference module. Initially, it employs two distinct branches to extract features from dermoscopic and clinical images separately. These extracted features are subsequently fused by concatenation. The combined feature set is then passed to a subsequent classifier for generating the final prediction.

4.2.2. Comparison to the state-of-the-arts

The Baseline model is a vanilla implementation of multi-modality and multi-label classification as shown in Fig. 4. We setup the upper bound of linear probing experiments by initializing the Baseline with ImageNet-pre-trained weights (Baseline-50-ImageNet). We also initialized the Baseline using our proposed SM3 pre-trained weights and conducted linear probing (SM3-linear) and fine-tuning (SM3-fine-tune) experiments. We benchmarked state-of-the-art (SOTA) SSL methods, including general SSL method SimCLR [16] and recent SSL method optimized for dermoscopic image analysis named SSD-KD [39], along with supervised SOTAs including commonly used baseline Inception-combined [1] and HcCNN [3], and recent graph-based GIIN [11], adversarial-based AMFAM [31] and patient-meta-based FM4-FS [32]. In addition, we initialized FM4-FS with our SM3 (FM4-FS + SM3) and fine-tuned with all layers.

4.2.3. Ablation studies

For each of the ablation studies, hyperparameters were grid-searched and the metrics were based on the AUC scores. We conducted ablation studies on the Multi-Modality SSL (MMSSL) and the Multi-Label SSL (MLSSL) components by incorporating them into the Baseline separately. For MMSSL, we evaluated three different fusion strategies including: (i) *concat*: concatenating dermoscopic and clinical features and maximizing the similarities between different views of the concatenation; (ii) *sep_shared*: maximizing the similarities between views of paired dermoscopy and clinical images using a shared projection head $g(\cdot)$; (iii) *sep_sep*: same as *sep_shared* but applying separate projection heads for each modality. We also used weights pre-trained on ImageNet (ImageNet) and weights pre-trained by SOTA SSL algorithms, including SwAV [58], BYOL [59], and SimCLR, for comparisons. The weights pre-trained on ImageNet were downloaded from Torchvision¹ and initialized to each branch. The SSL pre-trained weights were obtained by using the respective SSL algorithms to pre-train each branch separately on the corresponding image modalities. We also adopted the fusion module from F4M-FS to be part of the multi-modality SSL baseline. Additionally, to validate the effectiveness of multi-modality inputs, we pre-trained SimCLR using only dermoscopic images (SimCLR-derm).

For MLSSL, we evaluated five different strategies: (i) *no.proj*: there was no label projection head; (ii) *proj*: naively applying a label projection head for each label; (iii) *msa*: applying the multi-head self-attention; (iv) *tel*: applying a Transformer encoder layer; (v) *te*: applying a Transformer based encoder.

In addition, we conducted pair match experiments that query clinical images using dermoscopic images as the keys. It aims to find

matching corresponding clinical images. It helps to explore how different MMSSL fusion strategies utilize the mutual information between two modalities. Based on the assumption that paired dermoscopic and clinical image features contain the highest similarity, we generated a cosine similarity score matrix for ranking. We assessed the top-1 accuracy (Acc@1) that determines whether paired image features have the highest similarity score, and the top-5 accuracy (Acc@5) that considers the top five highest scores. Moreover, the average rank was computed by averaging the ranks of each paired image's similarity score relative to others. By calculating these metrics, we quantify the extent to which complementary information between two modalities were effectively leveraged by different strategies, and thus guiding the selection of an optimized method for efficient multi-modality fusion. We conducted pair matching experiments using the training set of Derm7pt, which contains 413 samples.

5. Results

5.1. Comparisons to the state-of-the-arts

The primary results of the experiment are presented in Tables 1–4. We first evaluated the effectiveness of the proposed SM3 representations via linear probing (SSL + Linear Probing). When compared to SimCLR, our SM3 showed consistent improvements including a 2.2% increase in mean AUC, 3.8% increase in mean Sens, 2.1% increase in mean Spec, and 3.6% in mean Prec, although SimCLR performed better in some categories, e.g., RS category. SSD-KD, based on SSL and knowledge distillation, achieved a second-best performance with an AUC of 79.6, mean Sens of 9.8, mean Spec of 96.7, and mean Prec of 31.7. In comparison, our SM3 had much higher mean Sens (+25.1%), mean Prec (+29.8%) and relatively higher mean AUC (+0.8%) but lower mean Sens (−3.1%). To understand the upper bound of our SSL linear probing, we compared the SM3-initialized Baseline model with the supervised ImageNet-initialized Baseline model, with SM3 resulting in a higher mean Specificity (+2.7%), a small gap in terms of mean AUC (−0.9%), mean Sensitivity (−7.6%), and mean Precision (−5.2%).

We conducted experiments to assess the effectiveness of SM3 in improving various backbones (SSL + Fine-tuning). After fine-tuning the SM3-initialized Baseline model, the mean AUC improved from 81.3 to 82.9, surpassing both Inception-Combined (mean AUC of 81.5) and HcCNN (mean AUC of 82.5). This improvement was consistent regarding mean Spec and mean Prec. We also experimented with another existing method, F4M-FS by replacing the ImageNet-pre-train weights with our SM3-pre-trained weights while keeping other components unchanged. We found that the SM3-initialized FM4-FS achieved 1% increase in mean AUC, 4.2% in mean Spec and 12.6% in mean Prec. Compared to

¹ <https://github.com/pytorch/vision>.

Table 1

Classification performance evaluated by AUC on the Derm7pt dataset. Three settings are presented: Supervised, SSL + Linear Probing, and SSL + Fine-tuning. The best result are in bold.

Strategy	Method	BWV PRS	DaG IR	PIG IR	PN ATP	RS PRS	STR IR	VS IR	DIAG MEL	AVG \pm STD
Supervised	Inception-Combined	89.2	79.9	79.0	79.9	82.9	78.9	76.1	86.3	81.5 \pm 4.3
	HcCNN	89.8	82.6	81.3	78.3	81.9	77.6	82.7	85.6	82.5 \pm 3.9
	GIIN	90.8	83.1	83.6	87.5	79.0	81.2	75.4	87.6	83.5 \pm 5.0
	AMFAM	91.1	81.9	83.4	82.0	86.7	80.7	80.9	89.1	84.5 \pm 4.0
	F4M-FS	90.6	80.1	83.5	83.9	81.7	81.4	78.9	89.0	83.6 \pm 4.2
	Baseline-50-ImageNet	87.4	78.0	82.0	79.4	80.6	76.1	80.5	86.7	81.3 \pm 4.0
SSL + Linear Probing	SimCLR	85.6	76.1	77.0	75.4	78.7	72.8	76.2	83.6	78.2 \pm 4.3
	SSD-KD	86.7	76.8	81.5	79.1	77.1	79.2	69.7	86.6	79.6 \pm 5.5
	SM3-50-linear	90.4	76.5	80.4	78.2	77.4	75.4	79.4	85.0	80.4 \pm 5.0
SSL + Fine-tuning	SM3-50-finetune	91.1	81.9	82.8	78.0	82.5	79.8	81.2	86.1	82.9 \pm 4.1
	FM4-FS + SM3	92.9	80.3	84.5	82.3	84.2	84.5	77.8	90.1	84.6 \pm 4.9

Table 2

Classification performance evaluated by Sensitivity on the Derm7pt dataset. Three settings are presented: Supervised, SSL + Linear Probing, and SSL + Fine-tuning. The best result are in bold.

Strategy	Method	BWV PRS	DaG IR	PIG IR	PN ATP	RS PRS	STR IR	VS IR	DIAG MEL	AVG \pm STD
Supervised	Inception-Combined	77.3	62.1	59.7	48.4	66.0	51.1	13.3	61.4	54.9 \pm 19.0
	HcCNN	92.2	80.2	55.7	40.9	95.2	35.1	20.0	68.8	61.0 \pm 27.7
	GIIN	69.9	70.1	39.2	77.5	21.9	67.0	3.6	59.0	51.0 \pm 26.7
	AMFAM	75.0	66.7	67.9	58.5	72.1	57.3	0.0	65.8	57.9 \pm 24.2
	F4M-FS	66.7	68.4	58.9	49.5	47.1	47.9	20.0	62.4	52.6 \pm 15.6
	Baseline-50-ImageNet	49.3	66.1	46.8	40.9	34.0	47.9	3.3	51.5	42.5 \pm 18.3
SSL + Linear Probing	SimCLR	20.0	70.1	46.0	26.9	12.3	24.5	0.0	49.5	31.1 \pm 22.6
	SSD-KD	0.0	66.7	9.7	0.0	0.0	0.0	0.0	2.0	9.8 \pm 23.2
	SM3-50-linear	50.7	40.1	31.5	39.8	18.9	41.5	0.0	56.4	34.9 \pm 18.1
SSL + Fine-tuning	SM3-50-finetune	34.7	51.4	68.5	32.3	25.5	48.9	20.0	50.5	41.5 \pm 16.1
	FM4-FS + SM3	70.7	52.5	48.4	35.5	45.3	37.2	3.3	31.7	40.6 \pm 19.4

Table 3

Classification performance evaluated by Specificity on the Derm7pt dataset. Three settings are presented: Supervised, SSL + Linear Probing, and SSL + Fine-tuning. The best result are in bold.

Strategy	Method	BWV PRS	DaG IR	PIG IR	PN ATP	RS PRS	STR IR	VS IR	DIAG MEL	AVG \pm STD
Supervised	Inception-Combined	89.4	78.9	80.1	90.7	81.3	85.7	97.5	88.8	86.6 \pm 6.3
	HcCNN	65.3	71.6	86.3	92.4	41.5	90.0	98.4	85.4	78.9 \pm 18.6
	GIIN	91.0	78.8	95.8	79.0	96.8	80.3	100.0	89.5	88.9 \pm 8.6
	AMFAM	90.3	82.4	83.0	85.6	82.6	85.9	92.4	91.4	86.7 \pm 4.1
	F4M-FS	91.6	72.9	87.1	90.1	96.2	88.4	97.8	88.8	89.1 \pm 7.6
	Baseline-50-ImageNet	98.4	72.0	87.8	86.1	94.5	93.4	99.7	95.6	90.9 \pm 9.0
SSL + Linear Probing	SimCLR	99.7	69.7	85.2	94.4	97.2	94.0	100.0	91.8	91.5 \pm 10.0
	SSD-KD	100.0	74.3	99.3	99.7	100.0	100.0	100.0	100.0	96.7 \pm 9.0
	SM3-50-linear	95.9	89.9	94.5	89.4	99.3	86.0	100.0	93.9	93.6 \pm 4.9
SSL + Fine-tuning	SM3-50-finetune	99.1	90.4	80.4	94.7	96.9	89.7	96.4	93.5	92.6 \pm 5.9
	FM4-FS + SM3	92.5	85.3	91.9	92.4	94.5	91.9	100.0	98.0	93.3 \pm 4.4

Table 4

Classification performance evaluated by Precision on the Derm7pt dataset. Three settings are presented: Supervised, SSL + Linear Probing, and SSL + Fine-tuning. The best result are in bold.

Strategy	Method	BWV PRS	DaG IR	PIG IR	PN ATP	RS PRS	STR IR	VS IR	DIAG MEL	AVG \pm STD
Supervised	Inception-Combined	63.0	70.5	57.8	61.6	56.5	52.7	30.8	65.3	57.3 \pm 12.0
	HcCNN	91.9	69.6	65.1	62.3	81.6	52.4	50.0	54.5	65.9 \pm 14.7
	GIIN	67.4	74.9	82.3	48.4	73.5	50.4	100.0	65.6	70.3 \pm 16.7
	AMFAM	56.0	82.5	61.3	51.6	46.2	54.3	0.0	76.2	53.5 \pm 24.9
	F4M-FS	64.9	67.2	67.6	60.5	82.0	56.2	42.9	65.6	63.4 \pm 11.1
	Baseline-50-ImageNet	88.1	65.7	63.7	47.5	69.2	69.2	50.0	80.0	66.7 \pm 13.7
SSL + Linear Probing	SimCLR	93.8	65.3	58.8	59.5	61.9	56.1	0.0	67.6	57.9 \pm 26.2
	SSD-KD	0.0	67.8	85.7	0.0	0.0	0.0	0.0	100.0	31.7 \pm 44.6
	SM3-50-linear	74.5	76.3	72.2	53.6	90.9	48.1	0.0	76.0	61.5 \pm 28.3
SSL + Fine-tuning	SM3-50-finetune	89.7	81.3	61.6	65.2	75.0	59.7	31.6	72.9	67.1 \pm 17.5
	FM4-FS + SM3	68.8	74.4	73.2	58.9	75.0	73.2	100.0	84.2	76.0 \pm 12.0

Table 5

Ablation on multi-modality and multi-label SSL. The metrics are based on mean AUC scores and the best results are in bold.

Strategy	AUC								
	BWV	DaG	PIG	PN	RS	STR	VS	DIAG	AVG
	PRS	IR	IR	ATP	PRS	IR	IR	MEL	
SimCLR-derm	82.6	68.4	75.8	74.0	78.2	73.7	71.3	80.1	74.7
ImageNet	83.4	72.0	77.6	78.0	73.0	73.0	79.6	82.7	77.4
SwAV	86.1	73.6	78.5	77.4	75.0	73.6	76.5	83.3	76.9
BYOL	88.0	75.8	74.2	74.8	74.6	75.9	81.8	81.6	77.3
SimCLR	85.6	76.1	77.0	75.4	78.7	72.8	76.2	83.6	78.2
F4M-FS	88.4	76.3	79.5	76.6	79.0	76.4	78.7	85.5	79.0
MMSSL _{concat}	86.2	76.0	74.2	74.5	76.5	70.9	74.0	83.6	77.0
MMSSL _{sep_shared}	88.2	75.3	75.6	75.1	74.7	71.7	75.5	83.9	77.5
MMSSL _{sep_sep}	87.7	75.2	79.4	77.8	77.8	73.4	79.8	84.6	79.5
MLSSL _{no_proj}	89.5	75.3	79.6	77.2	77.8	73.1	75.2	83.5	78.9
MLSSL _{proj}	86.9	75.7	76.3	76.1	77.6	72.1	77.1	84.1	78.2
MLSSL _{msa}	89.3	75.3	79.7	74.2	78.0	73.8	78.8	85.7	79.3
MLSSL _{tel}	90.4	76.5	80.4	78.2	77.4	75.4	79.4	85.0	80.4
MLSSL _{te}	90.0	75.3	80.2	78.6	78.3	72.5	74.9	84.5	79.3

the current supervised state-of-the-art AMFAM which obtained a mean AUC of 84.5, mean Sens of 57.9, mean Spec of 86.7, and mean Prec of 53.5, SM3-initialized FM4-FS outperformed it by 0.1% in mean AUC, 6.6% in mean Spec and 22.5% in mean Prec but with 17.3% lower in mean Sens.

5.2. Ablation studies

5.2.1. Efficacy of multi-modality SSL

The ablation results of MMSSL are presented in Table 5. *SimCLR* strategy obtained the best mean AUC of 78.2, outperforming *SwAV* (mean AUC of 76.9) and *BYOL* (mean AUC of 77.3). Pre-training on multi-modality inputs, *SimCLR* strategy improved the mean AUC by 3.5%, increasing it from 74.7 (*SimCLR-derm*). Compared to the ImageNet-pre-trained weights (mean AUC of 77.4), *SimCLR* strategy achieved a higher score with a mean AUC of 78.2. Compared to the *SimCLR*, *concat* strategy resulted in a decreased mean AUC of 77. The strategy *sep_shared* helped to improve the performance by 0.5% compared to naïve concatenation. In contrast, only *sep_sep* strategy, which maximized the mutual information between paired dermoscopic and clinical images with separate projection heads, resulted in an improved model performance with a mean AUC of 79.5 (increased by 1.3% compared to *SimCLR*). The SOTA baseline *F4M-FS* achieved a mean AUC of 79.0, outperforming the *SimCLR* strategy but falling short of our *sep_sep* strategy.

5.2.2. Efficacy of multi-label SSL

The choice of MLSSL strategies affected performance differently as shown in Table 5. Firstly, we assessed when there was no label projection head (*no_proj*) applied, i.e., using concatenated multi-modality features to generate pseudo-multi-labels. We found that without label projection, the pre-trained multi-label classifier was not able to learn meaningful representations, which decreased the mean AUC from 79.5 to 78.9. Moreover, naively applying a label projection head for each label (*proj*), without the proposed label-relation-aware module, resulted in worse representations due to the ignorance of structure and relationships among labels, and resulting in decreased mean AUC by an additional 0.7%. The multi-head self-attention (*msa*) was capable of learning and building correlations among labels during SSL pre-training such that it can achieve a similar mean AUC of 79.3. We also evaluated the inclusion of a Transformer encoder layer (*tel*) to measure the benefit of label associations. This strategy improved the result by 0.9%. The use of different Transformer based encoder (*te*) did not improve the performances.

Table 6

Ablation on pair matching between different image modalities for different SSL multi-modality strategies. The metrics are based on mean AUC scores and the best results are in bold.

Strategy	Avg rank (total 413)	Acc@1	Acc@5
ImageNet	97.67	0.21	0.37
SimCLR	83.16	0.22	0.38
MMSSL _{concat}	90.55	0.20	0.34
MMSSL _{sep_shared}	13.69	0.32	0.57
MMSSL _{sep_sep}	7.23	0.42	0.73

5.2.3. Pair matching between different modalities

The pair matching results of different fusion strategies are shown in Table 6. The *sep_sep* strategy outperformed others by a large margin, with the average rank surpassing that of the *ImageNet* strategy (97.67) by approximately 90 points, achieving top 1% (7.23/413) rank. The accuracy metrics followed the same trend with *sep_sep* strategy obtaining the highest Acc@1 of 0.42 and Acc@5 of 0.73. Strategies without multi-modality pre-training generally had worse performance in pair matching, for example, *ImageNet* had average rank of 97.67, Acc@1 of 0.21, and Acc@5 of 0.37, and *SimCLR* obtained average rank of 83.16, Acc@1 of 0.22 and Acc@5 of 0.38. Additionally, naïve concatenation did not aid in mutual information learning with it achieving average rank of 90.55, Acc@1 of 0.2 and Acc@ 5 of 0.34. Analogously, directly contrasting the two modality images boosted the performance of average rank by 76.86, Acc@ of 0.22, and Acc@5 of 0.23.

5.3. Empirical evaluation of hyperparameters

We conducted an empirical study to select the hyperparameters for multi-modality SSL and multi-label SSL. The results are shown in Fig. 5. We experimented with multi-modality SSL learning rate (5e-3, 1e-3, 5e-4, 1e-4, 5e-5, 1e-5, 5e-6, 1e-6) and temperature (10, 0.1, 0.01, 0.001), and multi-label SSL batch size (64, 128, 256), learning rate (1e-3, 1e-4, 1e-5, 1e-6), label projection head dimension (512, 1024, 2048, 4096), transformer head number (1, 2, 4, 8, 16), and transformer feed-forward dimension (128, 256, 512, 1024, 2048). We identified the best performing parameters and used them for the pre-training, i.e., multi-modality SSL learning rate is 1e-6, temperature is 0.1, multi-label SSL batch size is 256, learning rate is 1e-4, label projection head dimension is 512, transformer head number is 1, and transformer feed-forward dimension is 128.

6. Discussion

The main findings are that: (1) Our SM3 had superior performances compared with other SOTA SSL methods in a multi-modality and multi-label setting; (2) SM3 was shown to be effective in improving other existing methods, outperforming ImageNet-pre-trained counterparts; and (3) the ablation studies showed that both the MMSSL and MLSSL components contributed to the overall performance improvements.

6.1. Comparisons to the state-of-the-arts

As shown in Tables 1–4, most methods performed relatively well due to the use of complex multi-modality fusion techniques, such as class-balanced sampling and multi-task loss in Inception-Combined [1], concatenation of intermediary image features in HcCNN [3], adversarial fusion with attention mechanism in AMFAM [31], and hierarchy fusion at the feature and decision levels in F4M-FS [32]. However, most of these methods ignored the MLC setting and did not exploit the interrelationships among labels. GIIN [11], on the other hand, additionally leveraged a graph module to model the label relationships which resulted in improved performance. Compared to Baseline-50-ImageNet, our SM3-fine-tuned method achieved a 1.6% increase in mean AUC, surpassing established multi-modal fusion strategies like

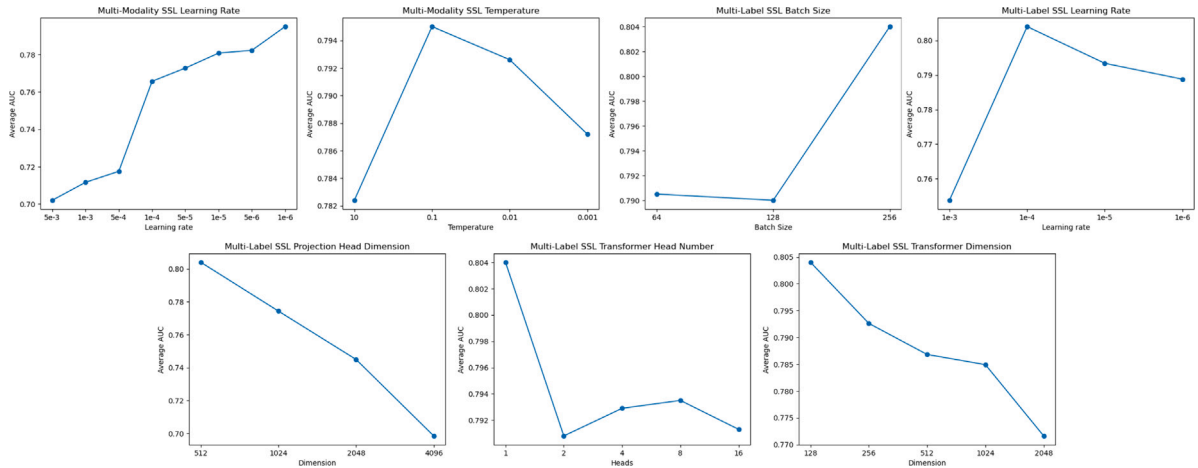


Fig. 5. Empirical evaluation for hyperparameter selection.

Inception-Combined and HcCNN. Similarly, our SM3 increased the recently published F4M-FS performance with an improved mean AUC of 1%. This indicates that SM3-pre-trained weights could be an alternative to ImageNet-pre-trained weights for improving the generalizability of different methods in skin lesion analysis.

In linear probing experiments, our method demonstrated a mean AUC enhancement of 2.2% when compared to the established SimCLR [16]. Furthermore, in comparison to the recent SSL skin lesion classification approach of SSD-KD [39], SM3 exhibited a mean AUC increase of 0.8%. It is worth noting that the degradation in performance with mean Sens metric is expected. This is mainly attributed to the fact that a large proportion of contributions are coming from the VS category, which tends to have low Sens with a relatively high Spec. As indicated by Kawahara et al. [1], the low sensitivity and high specificity is likely attributed to the substantial class imbalance issue in the dataset, where there only exists 71 IR studies in the VS category out of 1011 studies. By pre-training the model on a skewed dataset without labels, this discrepancy was accumulated resulting in a lower Sens score with a higher Spec score. It is worth noting that SSD-KD was pre-trained on a different skin lesion dataset, as such the out-of-distribution nature further exacerbates class imbalance issues, which results in imbalanced performance such as 0 Sens, 100 Spec, and either 0 or 100 Prec in certain sub-classes. Nevertheless, we identified that SM3 was effective in dealing with imbalanced data in the seven-point dataset when compared to the other two SSL methods. For instance, IR in the VS category has extremely limited number of cases when compared to the more prevalent REG, and both compared methods struggled to correctly classify IR. In contrast, our method leveraged multi-modality and multi-label techniques managing to correctly classify this minority class. These results underscore the efficacy of our multi-modality SSL design, which harnesses the supplementary potential of dermoscopic and clinical image features to extract enhanced discriminative skin attributes. Moreover, the inclusion of self-supervised multi-label pre-training augments the model's capacity to glean intricate interrelationships among labels, thereby contributing to a more consistent and reliable classification performance.

6.2. Ablation studies

Among the SSL algorithms we evaluated, SimCLR proved to be the most effective for skin lesion classification. We attribute this to the fact that skin lesions often have overlapping visual features across different classes, which can result in less discriminative representations for clustering-based methods like SwAV and for negative-sample-free approaches such as BYOL. Furthermore, training with multi-modality inputs using the SimCLR strategy demonstrated superior performance

than training with only dermoscopic images (*SimCLR-derm*), highlighting the advantages of incorporating multiple modalities for skin lesion classification. We found that naïve SSL pre-training (i.e., *SimCLR* strategy) with multi-modal data contributed to improving the baseline performance when compared to the commonly used ImageNet-pre-trained weights. This finding is consistent with previous work by Menegola et al. [15], where the domain gap between natural images and skin lesion images degraded performance. The *SimCLR* strategy, however, is not optimal for multi-modality learning since the mutual information between two modalities are not leveraged during the pre-training. In contrast, the SOTA baseline *F4M-FS* achieved better results, and these results illustrate the effectiveness of incorporating multi-modal inputs during pre-training. Nevertheless, for the multi-modal SSL, inappropriate multi-modal fusion (i.e., *concat*) could hinder the learning of meaningful representations. This observation is consistent with findings from a previous work [40] that naïve concatenation may result in a worse performance due to domain shift among different modalities. For example, the strategy *sep_shared* was better than *concat* but still inferior to the *SimCLR* strategy in terms of the mean AUC. In contrast, the *sep_sep* strategy demonstrated higher performance overall and we attribute this to the application of separate projection heads which can focus on independent image modality features and decide how to map them to increase their similarities. Compared to summing multi-modal features in *F4M-FS*, our contrast-based *sep_sep* strategy proved more effective in capturing complementary information between modalities. Furthermore, we also identified that contrasting naïve concatenation (*concat*) did not contribute to the learning of mutual information between the two modalities, and directly contrasting two modality features (*sep_shared* and *sep_sep*) was more effective giving better pair matching performance. It is noteworthy that although *sep_shared* was capable of learning more mutual information than *concat* and *SimCLR* strategies, its classification accuracy was lower than that of *SimCLR*. This suggests that an inefficient multi-modality pre-training, i.e., sharing the same projection head, learns trivial complementary multi-modality information and hinders the extraction of individual modality features.

In addition, we observed that directly pre-training a multi-label classifier on image features without label projection heads (*no_proj*) was not helpful, and simply projecting image features into label embeddings (*proj*) can disrupt self-supervised multi-label learning. We hypothesize that such failure was caused by the independent learning of label projection heads. A naïve label correlation learner (*msa*) had trivial contributions for multi-label SSL, whereas complex model (*te*) cannot achieve reasonable results neither attributing to the fact that the performance of the Transformer based architecture is heavily reliant on the use of large training dataset, which cannot be satisfied with

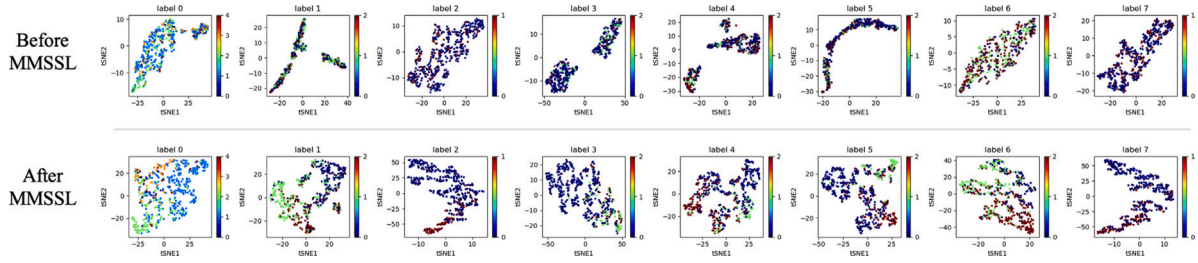


Fig. 6. t-SNE visualization of feature spaces of classification head for each label. The top row consists of feature spaces before MMSSL pre-training. The bottom row is features spaces after MMSSL pre-training.

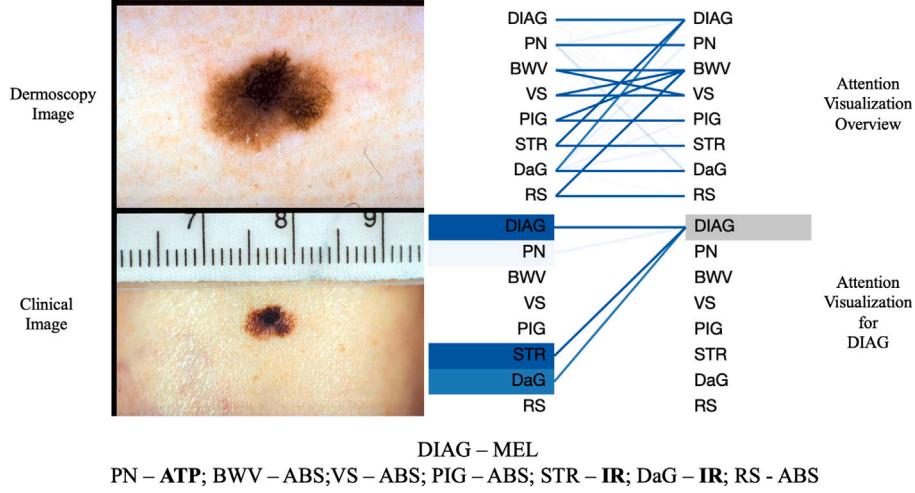


Fig. 7. Attention visualization of the label-relation-aware module. The left column displays an example image pair, while the right column shows the corresponding attention visualizations from the proposed label-relation-aware module. In the visualization, each line represents the attention from one label (left) to another (right), with line weight indicating the attention value (ranging from 0 to 1). The upper visualization provides an overview of all labels, and the lower visualization focuses on the DIAG label. The bottom text indicates the corresponding label, with attributes that increase the chance of melanoma highlighted in bold.

the current experimental dataset. Therefore, it is essential to define an optimal learning strategy, i.e., the Transformer encoder layer, for learning the correlations among labels.

6.3. Visualization

To demonstrate that our MMSSL enables preferable model initialization, we investigated the distribution of feature spaces in the last layer of the classifier and visualized them in Fig. 6. We selected outputs from the classification head of each label and applied t-SNE [60] for the visualization. We chose t-SNE for its ability to reduce high-dimensional data to two dimensions while preserving local structure, making it ideal for revealing patterns and clusters of the features. Before MMSSL pre-training, the feature spaces of classes (indicated by different colors) in each label exhibited a scattered distribution and largely overlapped with each other, indicating that the model struggled to differentiate the data effectively. The color gradients appear less distinct, suggesting lower intra-class compactness and higher inter-class confusion. After being pre-trained by MMSSL, we observe that the feature spaces of classes in each label demonstrate a more organized clustering (i.e., points with the same class are grouped together and separate with other class data points), indicating an improved ability of the model to differentiate between the labels. Also, the color gradients within the clusters are more pronounced, reflecting higher intra-class compactness and reduced inter-class confusion. The t-SNE visualizations provide clear evidence that MMSSL pre-training enhances the model's capability to differentiate between different labels in an MLC task.

In addition, we applied BertViz [61] to visualize the attention in our label-relation-aware module to better understand how it establishes the

relationship between seven attributes and the skin lesion type, as shown in Fig. 7. In this visualization, each line represents the attention from one label (left) to another (right), with line transparency indicating the attention value (ranging from 0 to 1). Higher attention values indicate a stronger relationship between the two labels, represented by more prominent lines. From the attention visualization overview, we can find that each label feature contains information from itself and others after the label-relation-aware module. Specifically, when DIAG is MEL, we observe that the main contribution comes from DIAG, STR, DaG, and PN. This aligns with the corresponding ground truth that STR, DaG and PN are IR/ATP, which increases the chance of melanoma. Based on BertViz visualization results, our MLSSL demonstrates promising results for establishing the relationship between seven attributes and skin lesion type.

6.4. Limitations and future works

In this study, we focused on developing a multi-modality and multi-label SSL framework for skin lesion classification. Therefore, addressing the challenge of an imbalanced dataset was not our primary focus. It is a known issue that pre-training on an imbalanced dataset, without labels can raise the issue of discrepancies. One potential solution to mitigate this challenge is to incorporate ensemble learning, where predictions from multiple models trained on different balanced subsets of the dataset are combined to improve overall performance. Additionally, we aim to introduce adaptive sampling techniques or loss reweighting mechanisms that can emphasize minority class samples during both pre-training and fine-tuning stages. Another limitation of our work is that the proposed SSL framework currently relies on pre-defined clustering techniques for generating pseudo-labels. While clustering

provides a useful surrogate for label supervision, it is inherently sensitive to hyperparameter selection and may introduce noise into the learning process. Future work will explore adaptive clustering strategies or hybrid approaches that incorporate weak supervision to improve the reliability of pseudo-labels. Leveraging large-scale foundation models trained on diverse medical imaging datasets could serve as a complementary approach to enhance pseudo-label quality. Although our method was demonstrated for skin lesion dataset, we suggest that it is applicable to other multi-modality and multi-label imaging datasets where it retains informative mutual features between modalities and relationships between labels. For example, our method can be adapted for multi-modality PET-CT and PET-MR datasets, where PET scans provide functional metabolic activity while CT and MR scans offer anatomical details. By preserving informative mutual features between modalities and capturing interrelationships between multiple diagnostic labels, our approach has the potential to enhance classification performance in these complex imaging scenarios.

7. Conclusion

In this paper, we introduced a new SSL algorithm for multi-modality and multi-label skin lesion classification. Specifically, maximum complementary information between dermoscopic and clinical images were leveraged during the pre-training when we directly contrasted these two modalities with separate projection heads. Experiments showed that this multi-modality SSL scheme can improve the accuracy of skin lesion classification. Furthermore, we found that generating pseudo-multi-labels using clustering analysis was a surrogate solution for self-supervised multi-label training. With our label-relation-aware module, SM3 was able to capture the interrelationships between labels. Our SM3 outperformed other SOTA SSL methods and helped to improve existing methods by using SM3-pre-trained weights.

CRediT authorship contribution statement

Hao Wang: Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Methodology, Formal analysis, Conceptualization. **Euijoon Ahn:** Writing – review & editing, Supervision. **Lei Bi:** Writing – review & editing, Supervision, Resources. **Jinman Kim:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by Australian Research Council (ARC) grants (DP200103748).

References

- [1] J. Kawahara, S. Daneshvar, G. Argenziano, G. Hamarneh, Seven-point checklist and skin lesion classification using multitask multimodal neural nets, *IEEE J. Biomed. Heal. Inform.* 23 (2) (2019) 538–546, <http://dx.doi.org/10.1109/JBHI.2018.2824327>.
- [2] D. Gutman, N.C. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, A. Halpern, Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC), 2016, arXiv preprint [arXiv:1605.01397](https://arxiv.org/abs/1605.01397).
- [3] L. Bi, D.D. Feng, M. Fulham, J. Kim, Multi-label classification of multi-modality skin lesion via hyper-connected convolutional neural network, *Pattern Recognit.* 107 (2020) 107502, <http://dx.doi.org/10.1016/j.patcog.2020.107502>.
- [4] G. Argenziano, G. Fabbrocini, P. Carli, V. De Giorgi, E. Sammarco, M. Delfino, Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: Comparison of the ABCD rule of dermoscopy and a new 7-point checklist based on pattern analysis, *Arch. Dermatol.* 134 (12) (1998) <http://dx.doi.org/10.1001/archderm.134.12.1563>.
- [5] G. Argenziano, C. Citalà, M. Ardicò, P. Bucci, P. De Simone, L. Eibenschutz, A. Ferrari, G. Mariani, V. Silipo, I. Sperduti, I. Zalaudek, Seven-point checklist of dermoscopy revisited, *Br. J. Dermatol.* 164 (4) (2011) 785–790, <http://dx.doi.org/10.1111/j.1365-2133.2010.10194.x>.
- [6] H. Kittler, H. Pehamberger, K. Wolff, M. Binder, Diagnostic accuracy of dermoscopy, *Lancet Oncol.* 3 (3) (2002) 159–165, [http://dx.doi.org/10.1016/S1470-2045\(02\)00679-4](http://dx.doi.org/10.1016/S1470-2045(02)00679-4).
- [7] Y. Liu, A. Jain, C. Eng, D.H. Way, K. Lee, P. Bui, K. Kanada, G. de Oliveira Marinho, J. Gallegos, S. Gabriele, V. Gupta, N. Singh, V. Natarajan, R. Hofmann-Wellenhof, G.S. Corrado, L.H. Peng, D.R. Webster, D. Ai, S.J. Huang, Y. Liu, R.C. Dunn, D. Coz, A deep learning system for differential diagnosis of skin diseases, *Nature Med.* 26 (6) (2020) 900–908, <http://dx.doi.org/10.1038/s41591-020-0842-3>.
- [8] L. Yu, H. Chen, Q. Dou, J. Qin, P.-A. Heng, Automated melanoma recognition in dermoscopy images via very deep residual networks, *IEEE Trans. Med. Imaging* 36 (4) (2017) 994–1004, <http://dx.doi.org/10.1109/TMI.2016.2642839>.
- [9] J. Zhang, Y. Xie, Y. Xia, C. Shen, Attention residual learning for skin lesion classification, *IEEE Trans. Med. Imaging* 38 (9) (2019) 2092–2103, <http://dx.doi.org/10.1109/TMI.2019.2893944>.
- [10] Z. Ge, S. Demnyanov, R. Chakravorty, A. Bowling, R. Garnavi, Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images, in: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11–13, 2017, Proceedings, Part III* 20, Springer, 2017, pp. 250–258.
- [11] X. Fu, L. Bi, A. Kumar, M. Fulham, J. Kim, Graph-based intercategory and intermodality network for multilabel classification and melanoma diagnosis of skin lesions in dermoscopy and clinical images, *IEEE Trans. Med. Imaging* 41 (11) (2022) 3266–3277, <http://dx.doi.org/10.1109/TMI.2022.3181694>.
- [12] J. Kawahara, A. BenTaieb, G. Hamarneh, Deep Features to Classify Skin Lesions, *IEEE, Prague, Czech Republic*, 2016, pp. 1397–1400, <http://dx.doi.org/10.1109/ISBI.2016.7493528>.
- [13] X. Sun, J. Yang, M. Sun, K. Wang, A benchmark for automatic visual classification of clinical skin disease images, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), *Computer Vision – ECCV 2016*, Vol. 9910, Springer International Publishing, Cham, 2016, pp. 206–222.
- [14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet large scale visual recognition challenge, *Int. J. Comput. Vis.* 115 (3) (2015) 211–252, <http://dx.doi.org/10.1007/s11263-015-0816-y>.
- [15] A. Menegola, M. Fornaciari, R. Pires, F.V. Bittencourt, S. Avila, E. Valle, Knowledge Transfer for Melanoma Screening with Deep Learning, *IEEE, Melbourne, Australia*, 2017, pp. 297–300, <http://dx.doi.org/10.1109/ISBI.2017.7950523>.
- [16] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A Simple Framework for Contrastive Learning of Visual Representations, *PMLR*, 2020, pp. 1597–1607.
- [17] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum Contrast for Unsupervised Visual Representation Learning, *IEEE, Seattle, WA, USA*, 2020, pp. 9726–9735, <http://dx.doi.org/10.1109/CVPR42600.2020.00975>.
- [18] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, A. Joulin, Unsupervised Learning of Visual Features by Contrasting Cluster Assignments, *NIPS '20, Curran Associates Inc., Red Hook, NY, USA*, 2020, pp. 9912–9924.
- [19] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P.H. Richemond, E. Buchatskaya, C. Doersch, B.A. Pires, Z.D. Guo, M.G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, M. Valko, Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning, *NIPS '20, Curran Associates Inc., Red Hook, NY, USA*, 2020, pp. 21271–21284.
- [20] H. Wang, E. Ahn, J. Kim, Self-supervised representation learning framework for remote physiological measurement using spatiotemporal augmentation loss, *Proc. AAAI Conf. Artif. Intell.* 36 (2) (2022) 2431–2439, <http://dx.doi.org/10.1609/aaai.v36i2.20143>.
- [21] H. Wang, E. Ahn, J. Kim, A dual-branch self-supervised representation learning framework for tumour segmentation in whole slide images, 2023, arXiv preprint [arXiv:2303.11019](https://arxiv.org/abs/2303.11019).
- [22] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, *Proc. IEEE* 86 (11) (1998) 2278–2324, <http://dx.doi.org/10.1109/5.726791>.
- [23] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, *Vol. 25*, Curran Associates, Inc., 2012.
- [24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going Deeper with Convolutions, *IEEE, Boston, MA, USA*, 2015, pp. 1–9, <http://dx.doi.org/10.1109/CVPR.2015.7298594>.
- [25] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, *IEEE, Las Vegas, NV, USA*, 2016, pp. 770–778, <http://dx.doi.org/10.1109/CVPR.2016.90>.
- [26] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, *Nature* 542 (7639) (2017) 115–118, <http://dx.doi.org/10.1038/nature21056>.

- [27] P. Yao, S. Shen, M. Xu, P. Liu, F. Zhang, J. Xing, P. Shao, B. Kaffenberger, R.X. Xu, Single model deep learning on imbalanced small datasets for skin lesion classification, *IEEE Trans. Med. Imaging* 41 (5) (2022) 1242–1254, <http://dx.doi.org/10.1109/TMI.2021.3136682>.
- [28] S. Qian, K. Ren, W. Zhang, H. Ning, Skin lesion classification using CNNs with grouping of multi-scale attention and class-specific loss weighting, *Comput. Methods Programs Biomed.* 226 (2022) 107166.
- [29] B.W.-Y. Hsu, V.S. Tseng, Hierarchy-aware contrastive learning with late fusion for skin lesion classification, *Comput. Methods Programs Biomed.* 216 (2022) 106666.
- [30] Y. Yang, F. Xie, H. Zhang, J. Wang, J. Liu, Y. Zhang, H. Ding, Skin lesion classification based on two-modal images using a multi-scale fully-shared fusion network, *Comput. Methods Programs Biomed.* 229 (2023) 107315.
- [31] Y. Wang, Y. Feng, L. Zhang, J.T. Zhou, Y. Liu, R.S.M. Goh, L. Zhen, Adversarial multimodal fusion with attention mechanism for skin lesion classification using clinical and dermoscopic images, *Med. Image Anal.* 81 (2022) 102535, <http://dx.doi.org/10.1016/j.media.2022.102535>.
- [32] P. Tang, X. Yan, Y. Nan, S. Xiang, S. Krammer, T. Lasser, FusionM4Net: A multi-stage multi-modal learning algorithm for multi-label skin lesion classification, *Med. Image Anal.* 76 (2022) 102307, <http://dx.doi.org/10.1016/j.media.2021.102307>.
- [33] L. Jing, Y. Tian, Self-supervised visual feature learning with deep neural networks: A survey, *IEEE Trans. Pattern Anal. Mach. Intell.* 43 (11) (2021) 4037–4058, <http://dx.doi.org/10.1109/TPAMI.2020.2992393>.
- [34] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, V. Natarajan, M. Norouzi, Big Self-Supervised Models Advance Medical Image Classification, *IEEE, Montreal, QC, Canada*, 2021, pp. 3458–3468, <http://dx.doi.org/10.1109/ICCV48922.2021.00346>.
- [35] A.v.d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, 2018, arXiv preprint [arXiv:1807.03748](https://arxiv.org/abs/1807.03748).
- [36] Ş. Öztürk, T. Çukur, Deep clustering via center-oriented margin free-triplet loss for skin lesion detection in highly imbalanced datasets, *IEEE J. Biomed. Heal. Inform.* 26 (9) (2022) 4679–4690.
- [37] L. Chaves, A. Bissoto, E. Valle, S. Avila, An evaluation of self-supervised pre-training for skin-lesion analysis, in: L. Karlinsky, T. Michaeli, K. Nishino (Eds.), *Computer Vision – ECCV 2022 Workshops*, Vol. 13804, Springer Nature Switzerland, Cham, 2023, pp. 150–166.
- [38] A. Guha Roy, J. Ren, S. Azizi, A. Loh, V. Natarajan, B. Mustafa, N. Pawlowski, J. Freyberg, Y. Liu, Z. Beaver, N. Vo, P. Bui, S. Winter, P. MacWilliams, G.S. Corrado, U. Telang, Y. Liu, T. Cemgil, A. Karthikesalingam, B. Lakshminarayanan, J. Winkens, Does your dermatology classifier know what it doesn't know? Detecting the long-tail of unseen conditions, *Med. Image Anal.* 75 (2022) 102274, <http://dx.doi.org/10.1016/j.media.2021.102274>.
- [39] Y. Wang, Y. Wang, J. Cai, T.K. Lee, C. Miao, Z.J. Wang, SSD-KD: A self-supervised diverse knowledge distillation method for lightweight skin lesion classification using dermoscopic images, *Med. Image Anal.* 84 (2023) 102693, <http://dx.doi.org/10.1016/j.media.2022.102693>.
- [40] X. Li, M. Jia, M.T. Islam, L. Yu, L. Xing, Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis, *IEEE Trans. Med. Imaging* 39 (12) (2020) 4023–4033, <http://dx.doi.org/10.1109/TMI.2020.3008871>.
- [41] O.F. Atli, B. Kabas, F. Arslan, M. Yurt, O. Dalmaz, T. Çukur, I2I-mamba: Multi-modal medical image synthesis via selective state space modeling, 2024, arXiv preprint [arXiv:2405.14022](https://arxiv.org/abs/2405.14022).
- [42] S. Zhang, J. Zhang, B. Tian, T. Lukasiwicz, Z. Xu, Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation, *Med. Image Anal.* 83 (2023) 102656, <http://dx.doi.org/10.1016/j.media.2022.102656>.
- [43] R. Huang, Z. Lin, H. Dou, J. Wang, J. Miao, G. Zhou, X. Jia, W. Xu, Z. Mei, Y. Dong, X. Yang, J. Zhou, D. Ni, AW3M: An auto-weighting and recovery framework for breast cancer diagnosis using multi-modal ultrasound, *Med. Image Anal.* 72 (2021) 102137, <http://dx.doi.org/10.1016/j.media.2021.102137>.
- [44] W. Liu, H. Wang, X. Shen, I.W. Tsang, The emerging trends of multi-label learning, *IEEE Trans. Pattern Anal. Mach. Intell.* 44 (11) (2021) 7955–7974.
- [45] Y. Prabhu, A. Kag, S. Harsola, R. Agrawal, M. Varma, Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising, in: *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 993–1002.
- [46] Y. Prabhu, A. Kag, S. Gopinath, K. Dahiya, S. Harsola, R. Agrawal, M. Varma, Extreme multi-label learning with label features for warm-start tagging, ranking & recommendation, in: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 2018, pp. 441–449.
- [47] D. Huynh, E. Elhamifar, Interactive multi-label cnn learning with partial labels, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9423–9432.
- [48] Q. Guan, Y. Huang, Multi-label chest X-ray image classification via category-wise residual attention learning, *Pattern Recognit. Lett.* 130 (2020) 259–266.
- [49] Ş. Öztürk, M.Y. Turali, T. Çukur, Hydravit: Adaptive multi-branch transformer for multi-label disease classification from chest X-ray images, *Biomed. Signal Process. Control.* 100 (2025) 106959.
- [50] Q. Liu, L. Yu, L. Luo, Q. Dou, P.A. Heng, Semi-supervised medical image classification with relation-driven self-ensembling model, *IEEE Trans. Med. Imaging* 39 (11) (2020) 3429–3440.
- [51] Y. Zhang, L. Luo, Q. Dou, P.-A. Heng, Triplet attention and dual-pool contrastive learning for clinic-driven multi-label medical image classification, *Med. Image Anal.* 86 (2023) 102772, <http://dx.doi.org/10.1016/j.media.2023.102772>.
- [52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, *Adv. Neural Inf. Process. Syst.* 30 (2017).
- [53] J. Yap, W. Yolland, P. Tschandl, Multimodal skin lesion classification using deep learning, *Exp. Dermatol.* 27 (11) (2018) 1261–1267, <http://dx.doi.org/10.1111/exd.13777>.
- [54] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, R. Garnett (Eds.), *PyTorch: An Imperative Style, High-Performance Deep Learning Library*, Curran Associates, Inc., 2019, pp. 8024–8035, <http://dx.doi.org/10.48550/arXiv.1912.01703>.
- [55] J. Macqueen, Some methods for classification and analysis of multivariate observations, 1967.
- [56] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2019, <http://dx.doi.org/10.48550/arXiv.1711.05101>.
- [57] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, 2022, pp. 16000–16009, <http://dx.doi.org/10.1109/CVPR52688.2022.01553>.
- [58] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, A. Joulin, Unsupervised learning of visual features by contrasting cluster assignments, *Adv. Neural Inf. Process. Syst.* 33 (2020) 9912–9924.
- [59] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al., Bootstrap your own latent: a new approach to self-supervised learning, *Adv. Neural Inf. Process. Syst.* 33 (2020) 21271–21284.
- [60] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, *J. Mach. Learn. Res.* 9 (11) (2008).
- [61] J. Vig, A multiscale visualization of attention in the transformer model, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, Florence, Italy, 2019, pp. 37–42, <http://dx.doi.org/10.18653/v1/P19-3007>, URL: <https://www.aclweb.org/anthology/P19-3007>.