Single-Shared Network with Prior-Inspired Loss for Parameter-Efficient Multi-Modal Imaging Skin Lesion Classification

Peng TANG[a,b], Tobias Lasser[a,b]

[a]organization=Department of Informatics, School of Computation, Information, and Technology, Technical University of Munich,city=Garching, country=Germany

[b]organization=Munich Institute of Biomedical Engineering, Technical University of Munich,city=Garching, country=Germany

Abstract

In this study, we introduce a multi-modal approach that efficiently integrates multi-scale clinical and dermoscopy features within a single network, thereby substantially reducing model parameters. The proposed method includes three novel fusion schemes. Firstly, unlike current methods that usually employ two individual models for for clinical and dermoscopy modalities, we verified that multimodal feature can be learned by sharing the parameters of encoder while leaving the individual modal-specific classifiers. Secondly, the shared cross-attention module can replace the individual one to efficiently interact between two modalities at multiple layers. Thirdly, different from current methods that equally optimize dermoscopy and clinical branches, inspired by prior knowledge that dermoscopy images play a more significant role than clinical images, we propose a novel biased loss. This loss guides the single-shared network to prioritize dermoscopy information over clinical information, implicitly learning a better joint feature representation for the modal-specific task. Extensive experiments on a well-recognized Seven-Point Checklist (SPC) dataset and a collected dataset demonstrate the effectiveness of our method on both CNN and Transformer structures. Furthermore, our method exhibits superiority in both accuracy and model parameters compared to currently advanced methods.

Keywords: skin lesion classification, multi-modal learning, single-shared network, biased loss

1. Introduction

The skin serves as the body's largest organ, safeguarding against external threats and invasion. Additionally, it plays crucial roles in thermoregulation, metabolism and sensory perception the body (Kolarsick et al., 2011). Over the past 30 years, skin cancer has emerged as one of the most lethal and rapidly spreading malignancies worldwide (Siegel et al., 2022). Among skin cancers, melanoma is the most fatal, as it can rapidly metastasize throughout the body and lead to a painful death. The five-year survival rate of melanoma can be improved to 95% if it is treated at an early stage (Balch et al., 2009). However, early prevention of melanoma is hindered by a large number of patients with skin diseases and a shortage of experienced dermatologists Therefore, there is an expectation that utilizing deep learning (DL) methods as an automatic aided system can enhance the diagnostic accuracy and efficiency of dermatologists.

With the development of deep learning, single-modality-based methods have experienced significant improvements compared to former hand-crafted methods. However, from a data-driven perspective, deep learning models tend to achieve more accurate predictions when they are provided with more information. Therefore, an increasing number of researchers have begun to explore the complementary information between clinical and dermoscopy images to achieve more robust results in complex clinical scenarios. (Kawahara et al., 2018; Yap et al., 2018) were among the first to propose fusing multimodal features using concatenation for skin lesion classification. Subsequent research of (Tang et al., 2022; Fu et al., 2022) improved performance by integrating prediction information in addition to feature fusion. To further enhance the diagnostic accuracy, (Bi et al., 2020; He et al., 2023; Zhang et al., 2023) introduced more advanced fusion modules for the feature interaction of clinical and dermoscopy images. They argued that simple concatenation cannot fully exploit the information from both modalities. However, the introduction of fusion modules requires significant computational costs, limiting their applications in real-world scenarios.

In this paper, we propose a novel parameter-efficient multi-modal (PEMM) framework for skin lesion classification, achieving state-of-the-art classification performance while using fewer parameters compared to current advanced methods. There are four difference between previous methods and our method. Firstly, unlike previous approaches that commonly employed ResNet as feature encoder, we conduct a comprehensive comparison between
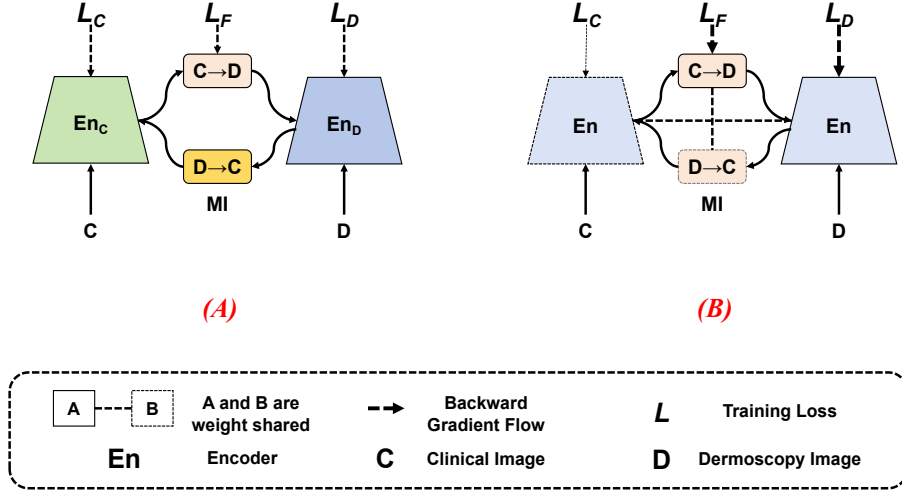
Figure 1: The overview structure of former methods and our PEMM framework.

ResNet and more advanced backbones, i.e., DenseNet (Huang et al., 2017), ConvNext (CXT) (Liu et al., 2022a), and SwinTransformer (ST) (Liu et al., 2021), which demonstrate that the latter three backbones can achieve higher accuracy with fewer parameters compared to ResNet. Secondly, given that the encoder accounts for the majority of the model's parameters, we naturally consider the idea of fusing multimodal features within a single network rather than using two individual encoders (See Fig. 1). Therefore, we explore and verify that multimodal features can be efficiently learned in a single-shared network with strong capacity by merely remaining the modal-specific classifiers, such as CXT and ST, resulting in significant parameter reduction while maintaining or subtly affecting accuracy. Thirdly, building on the concept of a 'shared network', we extend it to the fusion module and introduce a new shared cross-attention mechanism to efficiently conduct modality interaction on multi-scale multimodal features. Finally, inspired by the prior knowledge that dermoscopy images provide more useful information for diagnosis than clinical images, we introduce a new biased loss function. This function enables the model to focus more on the dermoscopy branch and less on the clinical branch, learning a better joint feature representation for the modal-specific classification task. Evaluations were conducted on two public datasets, and the results demonstrate the superiority of the proposed PEMM

framework in both accuracy and model parameter efficiency compared to current state-of-the-art methods. Extensive experiments validate the effectiveness of our method across both CNN and Transformer structures. The main contributions of our method can be summarized as follows:

1. We validated that both clinical and dermoscopy modalities can be input into a single-shared network with strong capacity, achieving similar performance while reducing a large number of parameters compared to commonly-used two individual networks.
2. We introduced a new shared cross-attention module to efficiently integrate multimodal features at different layers.
3. We propose a novel prior-biased loss that guides the single-shared network to learn more meaningful information for accurate diagnosis.
4. Our fusion method significantly outperforms state-of-the-art fusion methods, with only about 1 additional parameter increase on single-modal-based networks.

2. related works

2.1. single-modal imaging based methods for skin lesion classification

Most current deep learning-based skin lesion classification methods predict results based on single-modal images, such as clinical or dermoscopy images. Clinical images, captured by standard digital cameras or smartphone cameras (Ge et al., 2017; Pacheco et al., 2020), primarily display the geometry and color of the lesion (Yang et al., 2018).

To facilitate research in this area, (Sun et al., 2016) released a large clinical skin diseases dataset, known as SD-198, serving as a benchmark for comparison between convolutional neural networks (CNNs) and hand-crafted features. (Yang et al., 2018) presented effective feature representations by incorporating dermatologist's criteria, enhancing diagnostic performance, and capturing the manifestations of skin lesions. Additionally, they introduced a new metric called the 'complexity of image category' to guide self-paced balanced learning, addressing the class-imbalanced problem in classification tasks (Yang et al., 2019). Compared to hand-crafted methods, DL-based methods have achieved significant improvements in clinical-image (CI) based skin lesion classification. However, there still exists a considerable gap between CI-based and dermoscopy image-based methods (Togawa et al., 2023; Dascalu et al., 2022). For instance, in a comparative study by (Dascalu et al.,

2022), dermoscopy image (DI)-based CNN models significantly increased the accuracy of skin cancer diagnosis from 75% to 88% when compared to smartphone images

More researches (Esteva et al., 2017; Gu et al., 2019; Tang et al., 2020; Yao et al., 2021; Liu et al., 2022b; Gao et al., 2024) is directed towards dermoscopy images (DI) rather than clinical images (CI) due to two main factors. Firstly, as mentioned earlier, dermoscopy images offer higher diagnostic accuracy compared to clinical images (Dascalu et al., 2022). Secondly, the availability and high quality of numerous dermoscopy image datasets in challenges organized by the International Skin Imaging Collaboration also plays a significant role. (Esteva et al., 2017) demonstrated that a CNN trained using 129,450 images can achieve comparable performance with 21 board-certificated dermatologists on both classification tasks: benign seborrheic keratoses vs. keratinocyte carcinomas and benign nevi vs. malignant melanomas. (Gu et al., 2019) presented a progressive transfer learning method to address the generalization ability problem of fully-supervised methods and improve recognition performance, where adversarial learning was introduced to learn invariant attributes. (Yao et al., 2021) combined several techniques, including DropOut-related regularization, modified RandAugment, and a multi-weighted new loss, to address the class-imbalanced problem of skin lesion datasets. (Gao et al., 2024) explored and integrated information from different views, including RGB, HSL, and YCbCr, rather than only the RGB view, thereby enhancing skin lesion classification.

2.2. multi-modal imaging based methods for skin lesion classification

Despite their success, current methods are developed based on only a single modality, thus not exploiting the complementary information provided by both clinical and dermoscopy images (He et al., 2023). There is a growing body of research focused on utilizing multi-modal images for skin lesion classification (Ge et al., 2017; Kawahara et al., 2018; Yap et al., 2018; Bi et al., 2020; Tang et al., 2022; Wang et al., 2022; Fu et al., 2022; He et al., 2023; Zhang et al., 2023). (Kawahara et al., 2018) released the first multi-modal dataset for multi-label skin lesion classification, known as the Seven-point Checklist (SPC) dataset, laying the foundation stone in this field. (Tang et al., 2022; Fu et al., 2022) improved upon previous methods by incorporating a late fusion scheme to combine both feature and prediction information from different modalities. (Tang et al., 2022) employed a weighted averaging scheme, while (Fu et al., 2022) adopted a graph learning technique. (Wang

5

et al., 2022) proposed a modality discriminator that guides feature encoders to capture correlated and complementary information from two modalities. The most advanced methods (Bi et al., 2020; He et al., 2023; Zhang et al., 2023) focused on designing an additional fusion branch to facilitate modality interaction using both extracted features at multiple layers. The branch in HcCNN (Bi et al., 2020) was based on concatenation-based fusion and multi-scale attention modules, while CFANet (He et al., 2023) and TFormer (Zhang et al., 2023) mainly utilized cross-attention modules. However, the introduction of additional fusion branches incurs significant computational costs, which may hinder their application in various scenarios, such as deploying on mobile devices or implementing local AI-enabled family doctor systems for skincare.

2.3. Parameter-Sharing Network

Parameter-sharing network (PSN) or weight-sharing networks (WSN) are commonly employed in self-supervised learning as siamese networks. They are fed with multiple variants from the same source and then minimize the loss between their corresponding outputs to obtain task-related feature representations (Huang et al., 2022; Schürholt et al., 2021; Tao et al., 2022). Additionally, some works utilize PSN to improve performance while achieving lower memory consumption (Aich et al., 2020; Wang et al., 2023, 2020). For instance, (Wang et al., 2023) presented a parameter-sharing transformer block that captures scale-invariant information for 3D medical image segmentation. Similarly, (Wang et al., 2020) introduced a WSN that efficiently fuses RGB images and depth input for semantic segmentation tasks. However, there is a significant gap between the application scenarios due to the different types of data and tasks. Therefore, these methods cannot be directly applied to our task.

In the task of multi-modal skin lesion classification, TFormer (Zhang et al., 2023) employed a weight-sharing scheme to alleviate the overfitting problem. However, they did not thoroughly explore the impact of weight-sharing schemes on reducing parameters, leading to a confusing conclusion. For instance, in their configuration, the parameters of the introduced fusion branch are nearly identical to those of the feature encoder. It is highly probable that the weight-sharing scheme is achieved through the fusion branch rather than the encoder's capacity. In this paper, we verified that the single-shared network for parameter reduction is achieved based on encoder's capacity and maintaining individual classifiers, and further explored its gen-

6

eralization ability across different backbones by conducting extensive experiments. Moreover, in comparison to TFormer, we propose a new shared cross-attention module to efficiently reduce parameters on the fusion branch. Additionally, we introduce a novel biased loss mechanism that guides the single-shared network to be better optimized for the classification task.

3. Method: Parameter-Efficient Multi-Modal (PEMM) framework

The first step of our work is to explore utilizing different backbones as feature encoders instead of directly using ResNet for our classification task. Since many advanced backbones have been proposed and achieved better performance than ResNet for natural image recognition, such as DenseNet, ConvNext, and SwinTransformer. The results in Table 8 demonstrate the superiority of advanced backbones in improving classification accuracy and parameter reduction compared to the commonly-used ResNet. After that, we gradually introduce three main components, namely a single-shared network, shared cross-attention modules, and a biased loss function, as shown in Figure 2, into our Parameter-Efficient Multi-Modal (PEMM) framework.
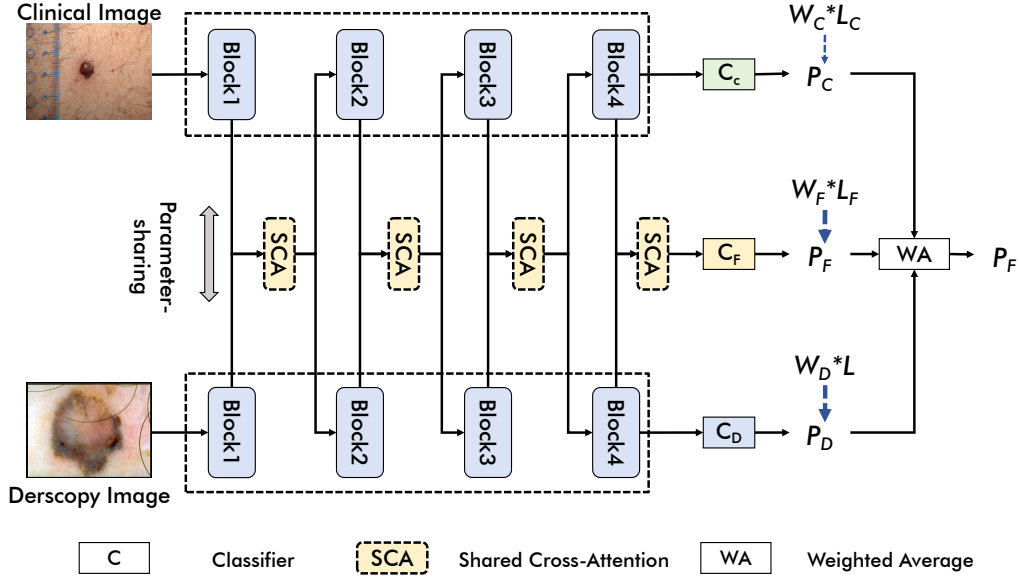

Figure 2: The detailed pipeline of our PEMM framework.

7

3.1. Single-Shared Network

Following (Tang et al., 2022; Fu et al., 2022; He et al., 2023), we also adopt two extra classifiers that can predict on clinical C_C and dermoscopy C_D branches and then conduct the late fusion on the prediction-level for more accurate results. Therefore, our baseline model contains two individual encoders, and three individual classifiers.

In deep learning-based multi-modal methods, feature encoders are indispensable as they are responsible for extracting individual features from different modalities, often occupying the majority of parameters in the entire model. Therefore, to build a parameter-efficient multi-modal method, we explore the extraction of modality-specific features from both clinical and dermoscopy images using a single-shared encoder (as illustrated in Fig.1(b)), rather than using two individual encoders as commonly done in previous methods (See Fig.1(a)). More specifically, as depicted in Fig. 2, the Single-Shared Network (SSN) adopts weight-sharing encoders to extract multi-modal features, while individual classifiers are built upon fully connected layers to predict on the extracted modality-specific features. We also attempted to share the parameters of the classifiers of dermoscopy and clinical branches, denoted as C_D and C_C respectively. However, the results were unsatisfactory, which is attributed to the robustness of convolution layers and the sensitivity of fully connected layers (More details can be found in Table 6). While this parameter-sharing scheme significantly compresses the parameters of our multi-modal fusion model, it is only effective with the ConvNext and SwinTransformer backbones. It fails to maintain accuracy compared to the corresponding non-parameter-sharing fusion model when ResNet and DenseNet are used as encoders (See Table)

3.2. Shared Cross-Attention Module

The effectiveness of the current advanced fusion module, i.e., cross-attention (CA) for multi-modal skin image fusion has been demonstrated in (He et al., 2023). As illustrated in Fig. 3, the CA module employs three individual convolutions to project the input clinical feature C into three feature vectors: C_k, C_v, and C_q. Subsequently, C_v and C_q are utilized to generate the attention map M_c through feature transformation and matrix multiplication. A dot product operation is then applied to C_v and M_c to obtain the attentive features A_c. Finally, the refined clinical feature $C_{refined}$ is obtained through matrix summation between the input feature C and A_c. Additional three convolutions are necessary to refine the dermoscopy feature $D_{refined}$.
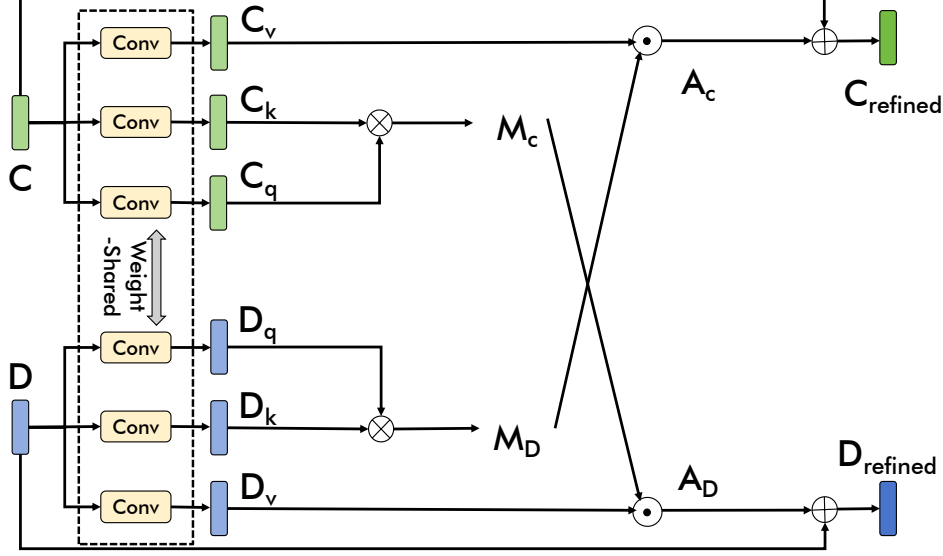
Figure 3: The detailed pipeline of shared cross-attention module.

Following the concept of "parameter-sharing", we further refined the CA modules by sharing the parameters of the three convolutions for the projections of input features from both modalities (refer to Fig. 3). Consequently, we can save half of the parameters of each CA module.

3.3. Biased Loss Function

In the training of previous methods, three branches are equally optimized, so their loss function can be formulated as Eq. 1:

$$L_{total} = (L_C + L_D + L_F)/3 \tag{1}$$

where L_{total} represents the total loss function and L_C, L_D and L_F are the loss function for clinical, dermoscopy and fusion branches, respectively (See Fig. 3).

However, equally optimization of these three branches seems like not reasonable based on the prior knowledge, which demonstrated that dermoscopy image-based model outperforms clinical image-based model (Dascalu et al., 2022). Inspired by the prior knowledge, we can have a hypothesis that dermoscopy information is more useful than clinical one in the multi-modal task

and thus proposed a new biased loss function, which are achieved by adjusting the corresponding weights of loss functions for different branches, as shwon in Eq. 2.

$$L_{total} = W_C \cdot L_C + W_D \cdot L_D + W_F \cdot L_F \tag{2}$$

where W_C, W_D and W_F are the corresponding weights of L_C, L_D and L_F, respectively. Specifically, in this function, W_D is set to bigger than W_C, and W_F are the sum of W_C and W_D as the fusion information is the combination of clinical and dermoscopy information. So, Eq. 2 can be simplified into Eq. 3

$$L_{total} = W \cdot L_C + (0.5 - W) \cdot L_D + 0.5 \cdot L_F \tag{3}$$

Where W is the weight factor and $W \in [0, 0.1, 0.2, 0.3, 0.4]$. With using this loss function, more backward gradient flows will pass the dermoscopy and branches and explicitly enforce the multi-modal model to concentrate more on the information from these two branches than clinical branch.

4. Experiments

4.1. Implementation Details

During training, the Adam optimizer (Kingma and Ba, 2014) is employed with a batch size of 24. The initial learning rate is set to 3e-5 and is adjusted every epoch following the CosineAnnealing learning schedule. Random transformations such as vertical and horizontal flipping, rotation, shifting, and enhancing brightness and contrast are applied during training. Stochastic weight averaging (Izmailov et al., 2018) is utilized to generate the final weight used for testing. We set varying training epochs of 250 and 150 to the SPC dataset and the collected ISIC dataset, respectively, owing to disparities in data number and the complexity of classification tasks. All images are resized to $224 \times 224 \times 3$ for both training and testing. During the testing, we followed (Tang et al., 2022) that searches the weights on the validation set and then form the final predictions by a weighted averaging scheme. All the experiments are based on the backbone of SwinTransformer are on the SPC dataset, unless specified. The weight factor in Eq. 3 is set to 0.1, as it yields the best performance (See Table 7).

4.2. Datasets and Metrics

Two datasets are used to evaluate the effectiveness of our method: the Seven-Point Checklist (SPC) (Kawahara et al., 2018) and the collected ISIC dataset.

SPC dataset The SPC dataset consists of 1011 cases, with each case containing both a dermoscopy image and a clinical image, along with diagnosis labels and seven-point checklist labels. For more detailed label information, refer to (Kawahara et al., 2018). The dataset was pre-split by the creator, so we followed the default setting in our experiments. In the task of multi-label skin lesion classification, we follow previous methods (Kawahara et al., 2018; Tang et al., 2022; He et al., 2023; Fu et al., 2022) by using the area under the curve (Avg AUC) and accuracy (Acc) as comparison metrics. Additionally, we provide precision (Prec), specificity, and sensitivity for supplementary analysis.

Collected ISIC dataset We collected 290 pairs of clinical and dermoscopy images from the ISIC Archive to build an additional multi-modal skin image dataset, named the collected ISIC dataset. It consists of 109 benign and 181 malignant cases. The collected ISIC dataset is divided into training, validation, and testing sets based on commonly used ratios of 0.7, 0.1, and 0.2, respectively. For evaluation, we used the metrics of averaged precision (AP) from the ISIC 2016 skin lesion classification challenge and Avg AUC. These metrics are chosen because the challenge involves classifying lesions into benign and malignant categories, aligning well with our task.

4.3. Comparison with state-of-the-art methods

We undertake a comparative analysis of our PEMM model with several existing methodologies, including TFormer (Zhang et al., 2023), GIIN (Fu et al., 2022), FusionM4Net-FS (Tang et al., 2022), AMFAM (Wang et al., 2022), HcCNN (Bi et al., 2020), and Inception-combination (Kawahara et al., 2018), on the SPC dataset. The comparative results concerning Averaged AUC and accuracy are presented in Tables 1 and 2, respectively. It is noteworthy that all reported results are extracted from the respective literature and are presumed to represent the optimal performance of each model, except for TFormer, which reported an averaged accuracy value. Therefore, for the comparison, we also opt for the model's weights demonstrating the best performance in terms of Avg AUC. For the ensuing experiments, our model underwent training five times, and the mean values alongside the standard

11

Table 1: The comparison between our PEMM and currently advanced methods on the SPC dataset in terms of AUC. The highest and second highest values in each column are bolden and italicized respectively. Incep-com: Inception-combined, FM-FS: FusionM4Net-FS, Avg: Averaged (%)

Methods	Diag					PN		STR		PIG		RS	DAG		BWV	VS		AVG
	BCC	NEV	MEL	MISC	SK	TYP	ATP	REG	IR	REG	IR	PRS	REG	IR	PRS	REG	IR	
Inception-com	92.9	89.7	86.3	88.3	*91*	84.2	79.9	87	78.9	74.9	79	82.9	76.5	79.9	89.2	85.5	76.1	83.7
HcCNN	94.4	87.7	85.6	88.3	80.4	*85.9*	78.3	87.8	77.6	*83.6*	81.3	81.9	77.7	82.6	89.8	87	82.7	84.3
AMFAM	94.1	89.7	89.1	90.6	81.7	84.5	82.0	*89.5*	80.7	**85.1**	83.4	**86.7**	77.7	81.9	91.1	**88.8**	80.9	85.7
FM-FS	*95.3*	92.6	89	*94*	89.2	*85.9*	*83.9*	87.9	81.4	80.9	83.5	81.7	*79.1*	80.1	90.6	87.8	78	86
GIIN	92.8	86.8	87.6	88.8	79.8	80.1	**87.5**	84.9	81.2	81.1	83.6	79	78.6	*83.1*	90.8	80.7	75.4	83.6
CAFNet	**97.1**	*92.7*	**92.2**	92.5	*91*	81.9	75.3	87.4	**85.4**	76.1	85	*85.4*	75.2	78.7	**94.7**	84.8	*83.5*	85.8
PEMM (Ours)	94.7	**93.0**	*90.8*	**94.9**	**91.7**	**86.7**	83.8	**90.1**	*84.4*	79.4	**86.1**	84.9	**80.7**	84.0	*93.9*	88.5	**85.4**	**87.6**

deviation from these five iterations were employed for a more robust analysis of our model.

As demonstrated in Table 1, CAFNet, FM-FS, and AMFAM notably outperform Incep-com, HcCNN, and GIIN, highlighting the effectiveness of cross-attention modules, weighted late fusion schemes, and adversarial learning schemes, respectively. Moreover, our PEMM model attains the highest performance in terms of Avg AUC value (87.6%), surpassing significantly the next three methods (FM-FS: 76.0%, CAFNet: 75.7%, and AMFAM: 75.7%), underscoring the superiority of our approach. Our PEMM model achieves the top-3 highest values across amolst all categories except PIG-REG, with the highest values in nine categories and the second-highest values in four categories, showing the robustness of our method across eight classification tasks. In Table 2, similar phenomena are observable. The proposed PEMM method attains the highest values in five label tasks (PN, BWV, PIG, DaG, and RS) out of eight label tasks, with CAFNet, AMFAM, and FM-SM ranking in the 2nd to 4th positions in terms of Avg ACC. Specifically, PEMM achieves the highest value of 77.4 % improve the Avg ACC values of CAFNet (76.8%), AMFAM (76.0%) and FM-FS (75.7%) by 0.7%, 1.3% and 1.6 %, respectively.

For further analysis, we adhere to the methodology outlined in (Bi et al., 2020; Fu et al., 2022) to present the results of melanoma-related features in Table 4. From this table, it is evident that our PEMM model attains the highest performance in terms of Avg AUC at 86.7% and Avg SEN at 64.1%, thereby affirming the efficacy of our method in detecting melanoma-related features. Regarding the Avg PRE value, GIIN attains the highest value of 70.3%, surpassing all other methods. We attribute this to the unbalanced

Table 2: The comparison between our PEMM and currently advanced methods on the SPC dataset in terms of accuracy. The highest and second highest values in each column are bolden and italicized respectively. Incep-com: Inception-combined, FM-FS: FusionM4Net-FS, Avg: Averaged (%)

Methods	PN	BWV	VS	PIG	STR	DaG	RS	Diag	AVG
Incep-com	*70.9*	87.1	79.7	66.1	74.2	60.0	77.2	74.2	73.7
HcCNN	70.6	87.1	**84.8**	68.6	71.6	**65.6**	80.8	69.9	74.9
AMFAM	70.6	*88.1*	83.3	70.9	74.7	63.8	*82.3*	75.4	76,0
FM-FS	*70.9*	86.8	81.8	*72.4*	74.4	61.0	**83.0**	74.9	75.7
CAFNet	70.1	87.8	*84.3*	**73.4**	**77.0**	61.5	*81.8*	**78.2**	*76.8*
TFormer	*70.9*	86.4	83.5	68.8	74.0	*64.9*	81.3	73	75.3
PEMM (ours)	**73.7**	**88.9**	82.5	71.9	*76.0*	**65.6**	**83.0**	*77.7*	**77.4**

distribution of VS-IR (irregular vascular structure), which comprises only 71 positive samples compared to 950 negative samples. This imbalance tends to lead GIIN to over-fit the negative samples of VS-IR, resulting in 100% values for SPE and PRE but only 3.6% for SEN. This indicates its effectiveness in detecting negative samples but its limited ability in identifying positive ones. Conversely, our PEMM achieves the second-highest value (33.3%) in SEN for VS-IR, showcasing its superior performance in detecting positive VS-IR samples even within an extremely unbalanced distribution.

The comparison of model parameters is illustrated in Table 3. Since there were no descriptions of the parameters for the compared methods in their respective papers, and only the source codes of TFormer and FM-FS are publicly available, we conducted a rough estimation of the parameters for other methods. Considering that Incep-com, AMFAM, and GIIN do not incorporate an additional third branch and solely utilize two InceptionV3 (57.4Mb) or two ResNet-50 (51.2Mb) as encoders along with fully connected layers as classifiers, we estimate the parameters of Incep-com to be slightly more than 57.4Mb, and the parameters of AMFAM and GIIN to be slightly more than 51.2Mb. Regarding HcCNN and CAFNet, which incorporate an additional branch, we estimate that their model parameters exceed those of their two encoders (ResNet-50: 51.2Mb). From the presented table, it is evident that our PEMM model achieves the highest Avg AUC and Avg ACC values while utilizing approximately 60% fewer parameters compared to the second-best methods, FM-FS (in terms of Avg AUC) and CAFNet (in terms

Table 3: The comprehensive comparison between our AMMFM and other methods in terms of model's parameters. $>$: slightly more, $>>$: much more

Method	Avg AUC (%)	Avg ACC (%)	Parameters
Incep-com	83.7	73.7	>57.4M
HcCNN	84.3	74.9	>65.0M
AMFAM	85.7	76	>51.2M
FM-FS	86	75.7	54.5M
GIIN	83.6	-	>51.2M
CAFNet	85.8	76.8	>>51.2M
TFormer	-	75.3	77.76M
PEMM(Ours)	**87.6**	**77.4**	31.12M

of Avg ACC). This result substantiates the effectiveness of our method.

4.4. Ablation studies

The ablation studies of our PEMM are shown in Table 5 to analyze the effect of three components, i.e., parameter-sharing (PS) encoder, shared cross-attention modules (SCA), biased loss (BL). The baseline model is a commonly-built multi-modal skin lesion classification model that adopts two individual encoders with a concatenation operation to fuse the features of the final layer of boto modalities, and trained by equally optimization (See Eq. 1). From the data presented in the table, it is apparent that by implementing parameter sharing among encoders, the total parameters of the baseline model experience a significant reduction from 58.49M to 30.14M. Despite this reduction, the decrease in diagnostic performance is negligible, with only a 0.2% decrease in AVG AUC value and a 0.23% decrease in AVG ACC (as observed in the 1st and 2nd columns). Furthermore, with the incorporation of shared cross-attention (SCA) modules into the PS encoder, there is an improvement in performance from 86.9% to 87.0% in AUC value and from 76.3% to 76.6% in ACC value, respectively, with only a subtle increase of 0.98M parameters (as shown in the 2nd and 4th columns). Moreover, the implementation of biased loss (BL) further enhances the performance of the PS-SCA model to 87.1% in AUC and 76.8% in ACC values without incurring any increase in computational cost (as shown in the 4th and 5th columns).. These results illustrate the effectiveness of parameter-sharing networks in parameter reduction, and the efficiency of SCA and BL in enhancing diag-

14

Table 4: Further comparison in melanoma-related features (%).

Metric	Method	DIAG MEL	PN ATP	STR IR	PIG IR	RS PRS	DaG IR	BWV PRS	VS IR	Avg
AUC	Incep-com	86.3	79.9	78.9	79	82.9	79.9	89.2	76.1	81.5
	HcCNN	85.6	78.3	77.6	81.3	81.9	82.6	89.8	82.7	82.5
	AMFAM	89.1	82.0	80.7	83.4	86.7	81.9	91.1	80.9	84.5
	FM-FS	89.0	83.9	81.4	83.5	81.7	80.1	90.6	78.9	83.7
	GIIN	87.6	87.5	81.2	83.6	79	83.1	90.8	75.4	83.5
	CAFNet	92.2	75.3	85.4	85.0	85.4	78.7	94.6	83.4	85.0
	PEMM	90.9	83.8	84.4	86.1	84.9	84.0	93.9	85.4	**86.7**
PRE	Incep-com	65.3	61.6	52.7	57.8	56.5	70.5	63.0	30.8	57.3
	HcCNN	62.8	62.3	52.4	65.1	81.6	69.6	91.9	50.0	67.0
	AMFAM	76.2	51.6	54.3	61.3	46.2	82.5	56.0	0.0	53.5
	FM-FS	65.7	82.2	56.2	67.6	82.0	67.2	64.9	42.9	68.5
	GIIN	65.6	48.4	50.4	82.3	73.5	74.9	67.4	100	**70.3**
	CAFNet	77.9	50.8	54.8	70.1	76.7	67.8	75.4	58.3	66.5
	PEMM	65.4	57.0	52.1	64.5	52.8	78.0	73.3	16.7	57.5
SEN	Incep-com	61.4	48.4	51.1	59.7	66	62.1	77.3	13.3	54.9
	HcCNN	58.4	40.9	35.1	55.7	95.2	80.2	92.2	20.0	59.7
	AMFAM	65.8	58.5	57.3	67.9	72.1	66.7	75.0	0.0	57.9
	FM-FS	62.4	49.5	47.9	58.9	47.1	68.4	66.7	20.0	52.6
	GIIN	59.0	77.5	67.0	39.2	21.9	70.1	69.9	3.6	51.0
	CAFNet	75.3	65.9	67.1	60.3	42.7	74.1	68.8	45.0	62.4
	PEMM	73.3	62.4	57.7	68.4	76.7	71.1	69.6	33.3	**64.1**
SPE	Incep-com	88.8	90.7	85.7	80.1	81.3	78.9	89.4	97.5	86.6
	HcCNN	88.1	92.4	90.0	86.3	41.5	71.6	65.3	98.4	79.2
	AMFAM	91.4	85.6	85.9	83.0	82.6	82.4	90.3	92.4	86.7
	FM-FS	88.8	90.1	88.4	88.1	96.2	72.9	91.6	97.8	89.2
	GIIN	89.5	79.0	80.3	95.8	96.8	78.8	91.0	100	88.9
	CAFNet	93.6	90.2	91.2	89.1	96.5	74.0	95.1	98.7	**91.1**
	PEMM	88.5	87.1	85.5	84.2	84.5	80.6	93.7	93.4	87.2

nostic accuracy with minimal or no increase in the model's parameters. In our comparison between CA (He et al., 2023) and our shared CA, we observed that CA does not outperform our SCA and even performs worse than simple concatenation operations (See 2nd-4th columns). This discrepancy may arise from the attention mechanism making the PS network more susceptible to overfitting, thereby resulting in poorer performance compared to concatenations, especially in smaller datasets. However, this issue can be mitigated by employing the PS scheme, akin to the phenomenon observed in (Zhang et al., 2023).

Table 5: Ablation studies of our PEMM in terms of AVG AUC, AVG ACC and model's parameters. FM: Fusion Module, PS: Parameter-Sharing, CA: Cross-Attention, SCA: Shared Cross-Attention, BL: Biased Loss (%).

| Encoder | FM | | BL | AVG AUC | AVG ACC | Parameters |
	CA	SCA				
Non-PS	Baseline			87.1±0.3	76.5±0.57	58.49M
PS				86.9±0.5	76.3±0.43	30.14M
	✓			86.7±0.4	75.9±0.43	32.10M
		✓		87.0±0.1	76.6±0.29	31.12M
		✓	✓	87.1±0.2	76.8±0.66	31.12M

4.5. Other experiments

4.5.1. The effect of individual classifiers

We also investigated the possibility of sharing parameters between the classifiers for the clinical and dermoscopy branches, denoted as C_C and C_D, respectively. However, as depicted in Table 6, compared to the Non-PS classifiers, the PS classifiers exhibit a significant reduction in AUC from 87.1% to 86.7% and ACC from 76.8% to 76.4%. This could be attributed to the sensitivity of fully connected layers to the input.

Table 6: Comparison between our model using parameter-sharing (PS) and non-PS classifiers. (%)

classifiers	AVG AUC	AVG ACC	Parameters(Mb)
PS	86.7±0.5	76.4±0.4	30.65
Non-PS	87.1±0.3	76.8±0.4	31.12

4.5.2. Comparison of different weight factor W

we conducted an experiment to explore the effect of different weight factors, denoted as W, in our biased loss function (Eq. 3). It is important to note that the weight factor W is assigned to L_C, while $0.5 - W$ is allocated to L_D. As illustrated in Table 7, we observed that the best and second-best overall performances are achieved by setting W to 0.1 and 0.2, respectively, surpassing the method trained using commonly-used equally optimized loss and other settings. This outcome supports our hypothesis that

improving classification performance is feasible by leveraging more information from the dermoscopy branch in multi-modal skin lesion classification. Furthermore, the best overall performance is attained when W is set to 0.1, indicating that specific clinical information can serve as supplementary data to enhance classification performance rather than disregarding it ($W=0$). Conversely, when W is set between 0.3 to 0.5, the corresponding diagnostic performances consistently deteriorate, with the worst performance observed at $W=0.5$. This underscores the significance of incorporating dermoscopy information in the classification process.

Table 7: The effect of different weight factor W in Eq. 3. EQ: Equally optimization that indicates the model is optimized by the loss function as shwon in Eq. 1. (%)

W	AVG AUC	AVG ACC
0	**87.23±0.36**	76.59±0.51
0.1	87.14±0.29	**76.84±0.42**
0.2	87.16±0.41	76.79±0.13
0.3	87.01±0.16	76.35±0.41
0.4	86.56±0.25	76.12±0.34
0.5	84.41±0.18	73.95±0.38
EQ	87.05±0.12	76.57±0.29

4.5.3. The effectiveness our method on different backbones

In addition to the SwinTransformer (Tiny), we further assessed the effectiveness of our method on different backbones, including ResNet50, DenseNet201, and Convnext (Tiny). As shown in Table 8, the proposed PEMM exhibits a decrease in diagnostic performance compared to the Baseline model when utilizing ResNet50 and DenseNet201 as the parameter-sharing encoder. Specifically, for ResNet50 and DenseNet201, both AUC and ACC metrics of PEMM decrease by over 0.5% compared to the corresponding baseline models. Notably, the AVG ACC value of both backbones even falls below the corresponding single-modality model trained using dermoscopy images (PEMM: 73.3% vs. Derm: 74.2% for ResNet50; PEMM: 74.4% vs. Derm: 75.0% for DenseNet201). Conversely, when our PEMM is applied to the Convnext and SwinTransformer backbones, the model's parameters can be significantly compressed (nearly 50%) while maintaining diagnostic accuracy compared to the baseline models, and in the case of SwinTransformer, even performing better. We believe that this discrepancy may be attributed to differences

Table 8: Comparisons between single-modal, baseline multi-modal and our PEMM methods based on different backbone. Params: Parameters (%).

Backbone	Model	AVG AUC	AVG ACC	Params(Mb)
ResNet50	Derm	84.4±0.3	74.2±0.4	26.68
	Clic	76.8±0.3	67.6±0.3	
	Baseline	85.2±0.1	74.3±0.2	55.52
	PEMM	84.6±0.2	73.3±0.6	36.93
DenseNet201	Derm	85.1±0.2	75.1±0.1	20.17
	Clic	76.8±0.2	68.5±0.4	
	Baseline	86.4±0.4	75.7±0.5	44.17
	PEMM	85.9±0.3	74.6±0.4	29.88
Convnext	Derm	86.5±0.5	75.9±0.2	29.05
	Clic	77.6±0.2	69.1±0.2	
	Baseline	87.0±0.4	76.6±0.5	58.96
	PEMM	87.1±0.2	76.4±0.4	31.35
ST	Derm	76.3±0.4	86.8±0.3	28.82
	Clic	69.0±0.7	77.3±0.4	
	Baseline	87.1±0.3	76.5±0.6	58.49
	PEMM	87.1±0.2	76.8±0.7	31.12

in the capacity of the backbones. More advanced backbones possess the capacity for parameter-sharing, while traditional backbones may lack this capability.

4.5.4. The effectiveness our method on the collected ISIC dataset

To further assess the effectiveness of our PEMM method, we collected a dataset from ISIC Archive for evaluation. Two backbones whose efficacy has been validated on the SPC dataset were continued to be evaluated on this dataset. As demonstrated in Table 9, compared to the baseline model, our PEMM achieves better performances while utilizing approximately 50% fewer model parameters, with both Convnext and SwinTransformer backbones. This underscores that our PEMM consistently achieves efficient multimodal skin lesion classification across different datasets. Additionally, it's noteworthy that the multi-modal baseline models perform even worse than

Table 9: The effectiveness of our method on the collected ISIC dataset. (%)

Backbone	Model	AVG AUC	AP	Params(Mb)
Convnext	Derm	87.3±0.9	87.0±0.9	29.05
	Clic	83.9±0.5	80.1±0.4	
	Baseline	87.2±0.3	84.7±0.5	58.95
	PEMM	88.3±0.4	86.8±1.0	31.35
ST	Derm	86.6±0.9	85.9±1.5	28.81
	Clic	85.3±1.0	80.9±1.6	
	Baseline	85.0±2.0	84.6±1.8	58.48
	PEMM	87.2±1.3	86.9±1.4	31.11

the corresponding single-modal models trained solely on dermoscopy images. Conversely, our PEMMs outperform the dermoscopy image-based models in terms of overall performance, indicating the vulnerability of the baseline model when applied to small datasets and the robustness of our PEMM across different datasets.

5. Conclusion

In this paper, we introduce a novel Parameter-Efficient Multi-Modal (PEMM) method for skin lesion classification. Our approach offers several key contributions: Firstly, by sharing the parameters of encoders with strong capacity while retaining individual classifiers, PEMM achieves approximately 50% compression in model parameters while preserving classification accuracy compared to models employing two separate encoders. Secondly, our proposed shared cross-attention module enhances modality interactions within the parameter-sharing network (PSN) with fewer parameters compared to commonly-used cross-attention mechanisms. Finally, we introduce a biased loss function, which leverages the prior knowledge that dermoscopy information is more critical than clinical images. This biased loss guides the PSN to prioritize learning from dermoscopy images, leading to improved optimization and classification. Extensive experiments validate the effectiveness of our PEMM method in compressing model parameters while maintaining accuracy. Furthermore, compared to current state-of-the-art methods, the results demonstrate that PEMM significantly outperforms them while utiliz-

ing fewer parameters on the SPC dataset.

References

Aich, S., Yamazaki, M., Taniguchi, Y., Stavness, I., 2020. Multi-scale weight sharing network for image recognition. Pattern Recognition Letters 131, 348–354.

Balch, C.M., Gershenwald, J.E., Soong, S.j., Thompson, J.F., Atkins, M.B., Byrd, D.R., Buzaid, A.C., Cochran, A.J., Coit, D.G., Ding, S., et al., 2009. Final version of 2009 ajcc melanoma staging and classification. Journal of clinical oncology 27, 6199.

Bi, L., Feng, D.D., Fulham, M., Kim, J., 2020. Multi-label classification of multi-modality skin lesion via hyper-connected convolutional neural network. Pattern Recognition 107, 107502.

Dascalu, A., Walker, B., Oron, Y., David, E., 2022. Non-melanoma skin cancer diagnosis: a comparison between dermoscopic and smartphone images by unified visual and sonification deep learning algorithms. Journal of cancer research and clinical oncology , 1–9.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. nature 542, 115–118.

Fu, X., Bi, L., Kumar, A., Fulham, M., Kim, J., 2022. Graph-based intercategory and intermodality network for multilabel classification and melanoma diagnosis of skin lesions in dermoscopy and clinical images. IEEE Transactions on Medical Imaging 41, 3266–3277.

Gao, G., He, Y., Meng, L., Huang, H., Zhang, D., Zhang, Y., Xiao, F., Yang, F., 2024. Multi-view compression and collaboration for skin disease diagnosis. Expert Systems with Applications , 123395.

Ge, Z., Demyanov, S., Chakravorty, R., Bowling, A., Garnavi, R., 2017. 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. pp. 250–258.

Gu, Y., Ge, Z., Bonnington, C.P., Zhou, J., 2019. Progressive transfer learning and adversarial domain adaptation for cross-domain skin disease classification. IEEE journal of biomedical and health informatics 24, 1379–1393.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

He, X., Wang, Y., Zhao, S., Chen, X., 2023. Co-attention fusion network for multimodal skin cancer diagnosis. Pattern Recognition 133, 108990.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708.

Huang, L., You, S., Zheng, M., Wang, F., Qian, C., Yamasaki, T., 2022. Learning where to learn in cross-view self-supervised learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14451–14460.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., Wilson, A.G., 2018. Averaging weights leads to wider optima and better generalization. arXiv preprint arXiv:1803.05407 .

Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G., 2018. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE journal of biomedical and health informatics 23, 538–546.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Kolarsick, P.A., Kolarsick, M.A., Goodwin, C., 2011. Anatomy and physiology of the skin. Journal of the Dermatology Nurses' Association 3, 203–213.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 10012–10022.

Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022a. A convnet for the 2020s, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11976–11986.

Liu, Z., Xiong, R., Jiang, T., 2022b. Ci-net: clinical-inspired network for automated skin lesion recognition. IEEE Transactions on Medical Imaging 42, 619–632.

Pacheco, A.G., Lima, G.R., Salomao, A.S., Krohling, B., Biral, I.P., de Angelo, G.G., Alves Jr, F.C., Esgario, J.G., Simora, A.C., Castro, P.B., et al., 2020. Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. Data in brief 32, 106221.

Ricci Lara, M.A., Rodríguez Kowalczuk, M.V., Lisa Eliceche, M., Ferraresso, M.G., Luna, D.R., Benitez, S.E., Mazzuoccolo, L.D., 2023. A dataset of skin lesion images collected in argentina for the evaluation of ai tools in this population. Scientific Data 10, 712.

Schürholt, K., Kostadinov, D., Borth, D., 2021. Self-supervised representation learning on neural network weights for model characteristic prediction. Advances in Neural Information Processing Systems 34, 16481–16493.

Siegel, R.L., Miller, K.D., Fuchs, H.E., Jemal, A., 2022. Cancer statistics, 2022. CA: a cancer journal for clinicians 72, 7–33.

Sun, X., Yang, J., Sun, M., Wang, K., 2016. A benchmark for automatic visual classification of clinical skin disease images, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14, Springer. pp. 206–222.

Tang, P., Liang, Q., Yan, X., Xiang, S., Zhang, D., 2020. Gp-cnn-dtel: Global-part cnn model with data-transformed ensemble learning for skin lesion classification. IEEE journal of biomedical and health informatics 24, 2870–2882.

Tang, P., Yan, X., Nan, Y., Xiang, S., Krammer, S., Lasser, T., 2022. Fusionm4net: A multi-stage multi-modal learning algorithm for multi-label skin lesion classification. Medical Image Analysis 76, 102307.

Tao, C., Wang, H., Zhu, X., Dong, J., Song, S., Huang, G., Dai, J., 2022. Exploring the equivalence of siamese self-supervised learning via a unified gradient framework, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14431–14440.

Togawa, Y., Yamamoto, Y., Matsue, H., 2023. Comparison of images obtained using four dermoscope imaging devices: An observational study. JEADV Clinical Practice 2, 888–892.

Wang, H., Xu, Y., Chen, Q., Tong, R., Chen, Y.W., Hu, H., Lin, L., 2023. Adaptive decomposition and shared weight volumetric transformer blocks for efficient patch-free 3d medical image segmentation. IEEE Journal of Biomedical and Health Informatics .

Wang, Y., Feng, Y., Zhang, L., Zhou, J.T., Liu, Y., Goh, R.S.M., Zhen, L., 2022. Adversarial multimodal fusion with attention mechanism for skin lesion classification using clinical and dermoscopic images. Medical Image Analysis 81, 102535.

Wang, Y., Sun, F., Lu, M., Yao, A., 2020. Learning deep multimodal feature representation with asymmetric multi-layer fusion, in: Proceedings of the 28th ACM International Conference on Multimedia, pp. 3902–3910.

Yang, J., Sun, X., Liang, J., Rosin, P.L., 2018. Clinical skin lesion diagnosis using representations inspired by dermatologist criteria, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1258–1266.

Yang, J., Wu, X., Liang, J., Sun, X., Cheng, M.M., Rosin, P.L., Wang, L., 2019. Self-paced balance learning for clinical skin disease recognition. IEEE transactions on neural networks and learning systems 31, 2832–2846.

Yao, P., Shen, S., Xu, M., Liu, P., Zhang, F., Xing, J., Shao, P., Kaffenberger, B., Xu, R.X., 2021. Single model deep learning on imbalanced small datasets for skin lesion classification. IEEE transactions on medical imaging 41, 1242–1254.

Yap, J., Yolland, W., Tschandl, P., 2018. Multimodal skin lesion classification using deep learning. Experimental dermatology 27, 1261–1267.

Zhang, Y., Xie, F., Chen, J., 2023. Tformer: A throughout fusion transformer for multi-modal skin lesion diagnosis. Computers in Biology and Medicine 157, 106712.