Medical Image Analysis 81 (2022) 102535

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Adversarial multimodal fusion with attention mechanism for skin lesion classification using clinical and dermoscopic images

Yan Wang[a], Yangqin Feng[a], Lei Zhang[b], Joey Tianyi Zhou[a], Yong Liu[a], Rick Siow Mong Goh[a], Liangli Zhen[a,*]

[a] Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore 138632, Singapore
[b] Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu, 610065, P.R.China

ARTICLE INFO

Article history:
Received 2 November 2021
Revised 7 July 2022
Accepted 11 July 2022
Available online 13 July 2022

Keywords:
Multimodal learning
Multimodal fusion
Skin lesion classification
Correlated and complementary information

ABSTRACT

Accurate skin lesion diagnosis requires a great effort from experts to identify the characteristics from clinical and dermoscopic images. Deep multimodal learning-based methods can reduce intra- and inter-reader variability and improve diagnostic accuracy compared to the single modality-based methods. This study develops a novel method, named adversarial multimodal fusion with attention mechanism (AMFAM), to perform multimodal skin lesion classification. Specifically, we adopt a discriminator that uses adversarial learning to enforce the feature extractor to learn the correlated information explicitly. Moreover, we design an attention-based reconstruction strategy to encourage the feature extractor to concentrate on learning the features of the lesion area, thus, enhancing the feature vector from each modality with more discriminative information. Unlike existing multimodal-based approaches, which only focus on learning complementary features from dermoscopic and clinical images, our method considers both correlated and complementary information of the two modalities for multimodal fusion. To verify the effectiveness of our method, we conduct comprehensive experiments on a publicly available multimodal and multi-task skin lesion classification dataset: 7-point criteria evaluation database. The experimental results demonstrate that our proposed method outperforms the current state-of-the-art methods and improves the average AUC score by above 2% on the test set.

1. Introduction

According to the Global Cancer Statistics 2020, skin cancer is ranked as the fourth leading cause of new cancer cases and deaths worldwide for 36 cancers and all cancers combined in 2020 (Sung et al., 2021). Skin cancer is one of the most dangerous cancers, especially melanoma, one of the most dangerous with the highest mortality skin cancer (Rigel et al., 1996). In a recent study (Barata et al., 2017), researchers have shown that early detection and timely adjuvant treatment could significantly reduce skin cancer mortality. Fortunately, with advanced developments in medical technology, there are many approaches to detect different kinds of skin cancers. Among these approaches, dermoscopic combined with clinical imaging is one of the most commonly used lesion diagnosis approaches in clinical practice (Massone et al., 2007). Dermoscopic images are obtained using optical magnification with liquid immersion and low angle-of-incidence lighting or cross-polarized lighting to make the contact area translucent, making subsurface structures visible. These images usually provide away to pay more attention to the local features of the lesions. Clinical images obtained using a digital camera usually provide more information about the global features of the lesions, such as the geometry and color of the lesions (Bi et al., 2020). The images captured by dermoscopic and clinical digital cameras make a comprehensive multimodal assessment of skin lesions possible.

In clinical practice, multimodal assessment is conducted by human experts. However, it lacks well-trained experts to perform large-scale skin cancer screening promptly. In addition, human experts' diagnosis is quite subjective, which is prone to intra- and inter-reader variability, causing inaccurate and inconsistent results across experts. Many factors can affect the diagnosis results, such as empirical knowledge, visual fatigue, and the resolution of images. Developing automated computer-aided diagnosis (CAD) systems to assist the diagnosis procedure may help to mitigate the impact of these factors (Chen et al., 2020; Xu et al., 2020; He et al., 2020; Zhang et al., 2019; Polat and Koc, 2020). The classification module is the core part of an automated skin lesion diagnosis system. When it comes the development of CAD methods, convolutional neural networks (CNNs) (LeCun et al., 2015) have

* Corresponding author.
E-mail address: llzhen@outlook.com (L. Zhen).

https://doi.org/10.1016/j.media.2022.102535
1361-8415/

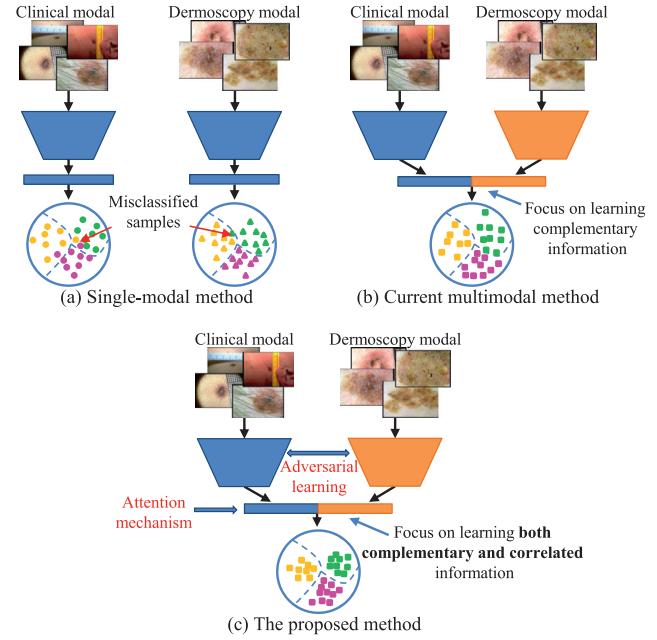


Fig. 1. The flow charts of single-modal classification and our proposed method. Different colors denote different data classes. Circular, triangle, and square stand for clinical, dermoscopic, and multimodal samples, respectively.

replaced the traditional methods (Claridge et al., 2003; Mendoza et al., 2009; Zhou et al., 2009; Ma and Staunton, 2013) and become the most effective approaches to learn the features of skin lesion images (Pereira et al., 2021; Thomas et al., 2021; Pérez et al., 2021).

In automated skin lesion classification, researchers have made an excellent effort to classify skin lesion images using CNNs. However, most of them only consider a single modality, i.e., clinical images or dermoscopic images (Yu et al., 2018; Zhang et al., 2019; Yu et al., 2016; Harangi, 2018). As mentioned above, clinical and dermoscopic imaging modalities capture different characteristics of skin lesions. Clinical images provide the global features of the lesions, while dermoscopic images provide the detailed features of lesions. One modality may not catch the critical information about the lesion and result in a wrong decision, as shown in Fig. 1(a). To overcome this problem, researchers have attempted to combine clinical and dermoscopic images to classify skin lesions (Ge et al., 2017; Yap et al., 2018; Kawahara et al., 2018). The key idea of these methods is to learn the complementary information from each modality to improve the classification performance, as shown in Fig. 1(b). The complementary information is the knowledge that is not visible in individual modalities on their own but is suitable for understanding the underline semantics of the target event/topic (Baltrušaitis et al., 2018). A typical way of learning complementary information is concatenating the feature vectors from different modalities. Each modality's feature vector provides information about different aspects of an object, event, or activity of interest (Liu et al., 2018). However, these methods mainly focus on learning the complementary information while ignoring the correlated information between the two input modalities. Correlated information is the correlation over the representations of different modalities. The correlated information can be leveraged to increase the confidence of the learned features for both modalities by encouraging the consistency of the feature vectors from the two modalities. The correlated information is represented in multiple aspects, such as the color, geometry information, and other potential shared characteristics between the two modalities; they are all critical for skin lesion classification.

In this paper, we propose a novel classification method, named adversarial multimodal fusion with attention mechanism (AM-FAM), to learn the discriminative feature representations from clinical and dermoscopic images. The flow chart of our method is shown in Fig. 1(c), from which we can see that, on the one hand, our method aims to learn high discriminative features from each modality by adopting attention-based reconstruction; on the other hand, it tries to restrain the CNN backbone to explicitly learn the correlated features from both modalities to maintain the essential shared characteristics. Then, we concatenate the feature vectors from the two modalities to gain high discriminative representations. Specifically, adversarial learning is adopted to guide the feature extractor to learn the correlated information. Moreover, we employ the gradient reversal layer (GRL) that forces the feature extractor to produce multimodal-invariant representations on multiple source images (Ganin et al., 2016). The multimodal-invariant representations are the correlated information we aim to learn, as well the shared characteristics we aim to maintain. At the same time, we design an attention-based image reconstruction procedure to encourage the feature extractor to learn more discriminative features for each modality by concentrating on the lesion area of its input image. Lastly, we combine the high-level feature vectors of the two modalities to obtain more discriminative representations and feed them to a classifier for the final classification. A multimodal skin lesion database, 7-point criteria evaluation database (Kawahara et al., 2018), is used to evaluate our proposed method. The experimental results show that our method outperforms the state-of-the-art methods and verify our method's effectiveness.

The main novelty and contributions of this work can be summarized as follows:

- A novel multimodal fusion method is proposed to perform automated skin lesion classification using clinical and dermoscopic images. Its effectiveness is verified on a widely-used skin lesion classification dataset, i.e., 7-point criteria evaluation database.
- By adopting the adversarial learning strategy, our method can learn the correlated information between the two modalities. More specifically, a modality discriminator is designed to guide the feature extractor to learn the correlated information **explicitly**.
- To extract more discriminative features for each modality, we propose a self-attention-based image reconstruction approach to enforce the feature extractor concentrating on lesion areas automatically.
- Unlike most existing methods that only consider the complementary information, our method simultaneously considers both the correlated and complementary information of the two modalities.

The rest of this paper is organized as follows. First, a review of related work is provided in Section 2. In Section 3, we present the details of the material and our proposed method. In Section 4, we describe the experimental setups and performance metrics and report the experimental results. The discussion and future work are presented in Section 5. At last, we conclude this work in Section 6.

2. Related work

This section reviews some related skin lesion classification approaches, including single-modality skin lesion classification, multimodal fusion methods, and multi-modality skin lesion classification. Also, we will highlight how the proposed method differs from the existing methods.

2.1. CNN-based skin lesion classification

In early studies, researchers usually used a CNN model to extract features from dermoscopic images. A commonly used strategy is using the model pre-trained on ImageNet (Russakovsky et al., 2015) and finetuning the model on the target dataset. For instance, Kawahara et al. proposed a CNN model which adopts ImageNet pre-trained parameters to initialize the model and then finetunes this pre-trained CNN on the skin lesion data (Kawahara and Hamarneh, 2016). Pomponiu et al. proposed a similar approach that uses pre-trained AlexNet to replace the hand-crafted features and achieves a performance improvement (Pomponiu et al., 2016). In recent years, researchers have started to design more sophisticated CNN architectures and used new strategies to improve classification performance. For example, Zhang et al. introduced an attention mechanism and proposed an attention residual approach for skin lesion classification (Zhang et al., 2019). The purpose of the attention mechanism is to use the high-level layer attention feature maps to guide the generation of the low-level layer feature maps. Similarly, Gessert et al. used an attention mechanism combined with diagnosis-guided loss weighting to improve the performance of skin lesion classification (Gessert et al., 2020). They divided the high-resolution image into several patches and used an attention mechanism to learn the global context between patches. Instead of designing a new architecture, Harangi et al. proposed an ensemble strategy to aggregate the feature vectors of different CNN models by computing ensemble weight for each model (Harangi, 2018). Finally, combining several related tasks, including classification, detection, and segmentation, Song et al. proposed an end-to-end multi-task framework to classify, detect, and segment the dermoscopic images simultaneously and improved each task's performance (Song et al., 2020).

However, these methods only use a single dermoscopy modality to classify the skin lesion. They have not used the clinical images' complementary information, which contains the overall characteristics of the lesions.

2.2. Multimodal skin lesion classification

Classifying the skin lesion with dermoscopic and clinical images can be regarded as a multimodal feature fusion procedure. For multimodal fusion, Wang et al. proposed a parameter-free multimodal fusion framework, namely, Channel-Exchanging-Network (CEN), that dynamically exchanges channels between sub-networks of different modalities (Wang et al., 2020). Based on the transformer, Nagrani et al. proposed an attention bottlenecks architecture that uses the proposed bottleneck module for multimodal feature fusion at multiple layers (Nagrani et al., 2021). Unlikely the architecture design methods, Liu et al. proposed a contrastive learning objective, TupleInfoNCE, to learn the complementary information by considering both strong and weak modalities (Liu et al., 2021). The previously mentioned methods are designed for specific natural scenario tasks. For multimodal skin lesion classification, we can sort current works into two main strategies to fusion different modalities' information. The first strategy is fusing the representations of each modality at the end of the CNN model. Based on this strategy, Yap et al. proposed a ResNet-50-based architecture to extract representations from both modalities and then concatenate the representations of these two modalities with meta representations to make the final classification (Bhardwaj and Rege, 2021). Similarly, Kawahara et al. also used two image modalities and one metadata to classify each patient (Kawahara et al., 2018). The difference between Yap et al. and Kawahara et al. is that Kawahara et al. used Inception-V3 to extract representations and tried different combinations of the three modalities to find the optimal performance. The second strategy is fusing the features of different modalities progressively on all levels of the CNN architecture. Based on this strategy, Bi et al. proposed a hyper-connected convolutional neural network (HcCNN) and designed a new fusion block to fuse the feature maps of two modalities to obtain the fused representations (Bi et al., 2020). HcCNN concatenates the representations from two single modalities to make the final classification for each sample. Ge et al. proposed a three-branch CNN architecture named Triplet: one for clinical images, one for dermoscopic images, and another sub-networks for extracting the representations from both modalities (Ge et al., 2017). Besides, Triplet applies a saliency map to make the network learn the area's features most likely to belong to the foreground, which contains crucial information about the input image.

However, the first strategy does not impose constraints on the feature representations, leading the final fusion representation suboptimal for the classification. In contrast, the second strategy introduces a new branch to learn the complementary information that increases the model complexity, thus the model requires more computing resources. Besides, most of the current multimodal feature fusion methods only focus on learning and using complementary information. The correlated information between different modalities is ignored. In our method, we consider both of them to learn more discriminative representations for the skin lesion classification.

3. Material and method

3.1. Material

We employ a publicly available multimodal skin lesion dataset, named 7-point criteria evaluation database (Kawahara et al., 2018), as our material. It contains three modalities (two image modalities and one text modality) for evaluating automated image-based prediction of the 7-point skin lesion malignancy checklist. There are 1011 cases, and each case contains one dermoscopic image, one clinical image, and metadata (such as patient gender and lesion location). Two image modalities are used for evaluation in this study, i.e., dermoscopic and clinical images. These 1011 cases are officially divided into three subsets for evaluating algorithms' performance. The train, validate, and the test sets contain 413, 203, and 395 cases, respectively. The sizes of dermoscopic images range from 474×512 to 532×768 pixels, and the size of clinical images range from 480×512 to 532×768 pixels. There are two classification tasks: the DIAGNOSIS (DIAG) classification and the 7-point criteria classification. The DIAG classification task is to classify each case into one of five categories: basal cell carcinoma (BCC), nevus (NEV), melanoma (MEL), miscellaneous (MISC), and seborrheic keratosis (SK). The 7-point criteria classification task includes seven classification sub-tasks: 1) Pigment Network (PN), 2) Blue Whitish Veil (BWV), 3) Vascular Structures (VS), 4) Pigmentation (PIG), 5) Streaks (STR), 6) Dots and Globules (DaG), and 7) Regression Structures (RS). Among them, PN is a three-category classification task: absent (ABS), typical (TYP), and atypical (ATP); BWV and RS aim to distinguish absent (ABS) and present (PRS) categories; VS, PIG, STR, and DaG aim to distinguish: absent (ABS), regular (REG), and irregular (IR) categories. The detailed statistics of the material are summarized in Table 1.

3.2. Framework of adversarial multimodal fusion with attention mechanism (AMFAM)

The framework of our proposed AMFAM is shown in Fig. 2, from which we can see that AMFAM contains two branches: the clinical image branch and the dermoscopic image branch. When the dermoscopic and the clinical images are input to AMFAM, they will

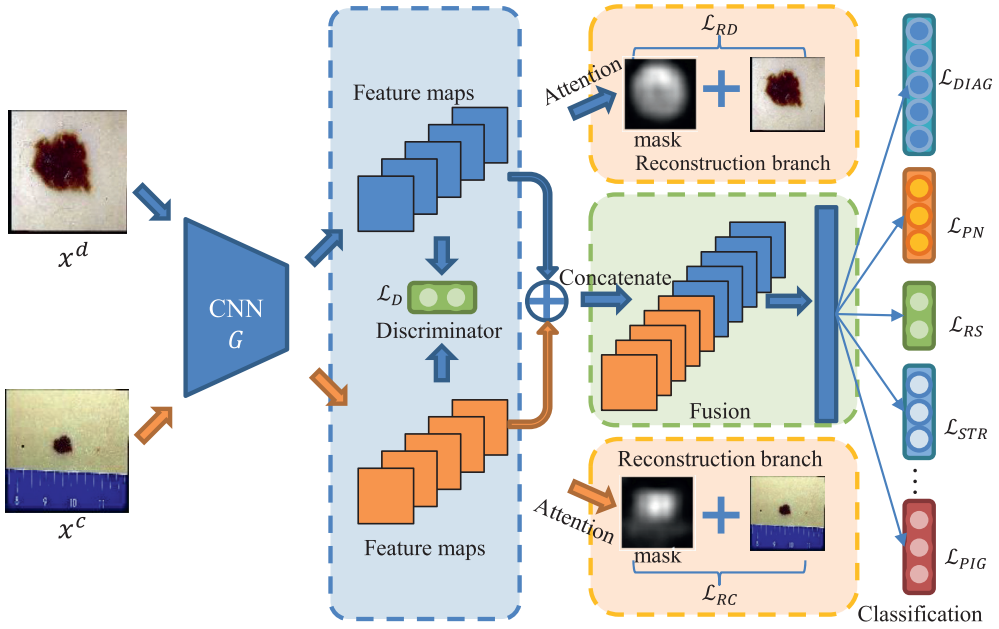


Fig. 2. The framework of our proposed method. The blue dashed box denotes an adversarial discriminator, the orange dashed boxes denote the attention mechanism-based reconstruction modules, and the green dashed box denotes the classification module. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
The detailed statistics for the 7-point criteria evaluation database. #Train, #Validate, and #Test denot the number of images in the training, validation, and test sets, respectively. #Total: the total number of images.

Task	Category	#Train	#Validate	#Test	#Total
DIAG	BCC	19	7	16	42
	NEV	256	100	219	575
	MEL	90	61	101	252
	MISC	32	25	40	97
	SK	16	10	19	45
PN	ABS	160	84	156	400
	TYP	160	75	146	381
	ATP	93	44	93	230
BWV	ABS	339	157	320	816
	PRS	74	46	75	195
VS	ABS	347	163	313	823
	REG	43	22	52	117
	IR	23	18	39	71
PIG	ABS	253	112	223	588
	REG	44	26	48	118
	IR	226	65	124	305
STR	ABS	273	123	257	653
	REG	39	24	44	107
	IR	101	56	94	251
DaG	ABS	84	45	100	229
	REG	156	60	118	334
	IR	173	98	177	448
RS	ABS	317	152	289	758
	PRS	96	51	106	253

go through a CNN for feature extraction. We obtain the CNN feature maps from its last layer as the extracted representations. Then our method will input the extracted representations into three modules: adversarial discriminator module, attention reconstruction module, and classification module.

These three modules and the feature extractor are trained jointly in an end-to-end manner to guide the feature extractor in learning both the correlated and complementary features. We will illustrate the details below.

Let $\mathcal{D}_s = \{(x^c, x^d, Y)_i\}_{i=1}^N$ be the set of 7-point criteria evaluation dataset, where $x_i^c \in \mathbb{R}^{w\times h\times 3}$ and $x_i^d \in \mathbb{R}^{w\times h\times 3}$ denote the i-th clinical image and dermoscopic image, respectively. w and h are the width and height of the input image. Each image is a 3-channel RGB image. $Y_i = \{y_i^1, y_i^2, \ldots, y_i^T\}$ denotes the label for DIAG, PN, BWV, VS, PIG, STR, DaG, and RS classification tasks, respectively. $T = 8$ denotes the number of classification tasks. Lastly, N is the total number of cases. The goal of our proposed method is to train the neural network as a function $F(\boldsymbol{\theta})$ to map the input clinical and dermoscopic images from input space to its label space, where $\boldsymbol{\theta} = \{\theta_G, \theta_{RC}, \theta_{RD}, \theta_D, \theta_C\}$ represents the trainable parameters of the neural network model. Moreover, $\theta_G, \theta_{RC}, \theta_{RD}, \theta_D$, and θ_C are the trainable parameters of the CNN backbone G, the clinical image reconstruction component RC, the dermoscopic image reconstruction component RD, the discriminator D, and the classification component C, respectively.

3.2.1. Adversarial multimodal fusion

The adversarial multimodal fusion uses adversarial learning to learn the correlated features. In our skin lesion classification task, the correlated features capture the shared characteristics of the two modalities, namely, the color, geometry information, and other potential shared characteristics between the two modalities. We use a discriminator to classify the input feature maps from the feature extractor into one of the two modalities. If the discriminator can accurately classify the feature maps, it can be inferred that the feature maps from different modalities have negligibly correlated information. Otherwise, if the discriminator has been confused and it is hard to distinguish the feature maps from different modalities, it indicates that the feature maps from different modalities contain significantly correlated information. Adversarial learning is a mini-max game that trains the feature extractor to extract the features that can minimize the discriminator's capability of assigning the correct label to both modalities. Meanwhile, the discriminator attempts to classify the feature maps into their corresponding modality as correctly as possible. Thus, the discriminator should maximize the probability of assigning the correct label to both modalities. When the mini-max optimization converges, ideally, the feature extractor should be able to extract the features which contain the correlated information from different modalities,

yet the discriminator cannot distinguish the feature maps from which modality.

Before the discriminator, we add one global average pooling layer that transforms the input feature maps into feature vectors. Besides, a gradient reverse layer (GRL) (Ganin et al., 2016) is connected to update the generator's gradient. In the forward pass, GRL performs as an identity layer. In the backward pass, GRL multiplies a gradient calculated from the discriminator error by a negative scaler and propagates the negative gradient to the feature extractor. The adversarial discriminator of AMFAM is a fully-connected neural network, which contains three layers: the input layer, the hidden layer of 512 neurons with the ReLU activation function, and the output layer (a softmax layer for classification). It outputs the probabilities of input feature maps for different modalities, namely, clinical modality and dermoscopy modality. To achieve the goal of adversarial learning, we optimize the following two-player minimax game:

$$\begin{aligned}\mathcal{L}_D(x^c, x^d; \theta_G, \theta_D) &= \min_G \max_D V(D, G) \\ &= \mathbb{E}[\log D(G(x^c, \theta_G), \theta_D)] \\ &\quad + \mathbb{E}[\log(1 - D(G(x^d, \theta_G), \theta_D))].\end{aligned} \tag{1}$$

3.2.2. Attention mechanism-based reconstruction

Different from the methods that incorporate the attention mechanism into the backbone. We use the attention mechanism-based reconstruction method to constrain the backbone to learn the lesion area features. The benefit of our strategy is that we can focus on learning the features of the lesion area while maintaining the full-image information for each modality. The attention mechanism-based reconstruction module aims to restrain the feature extractor from extracting the feature maps of each modality containing the discriminative information and with full-image information. To achieve this goal, we apply an attention mechanism to encourage the reconstruction by concentrating on reconstructing the lesion area instead of the background. The better reconstruction results are achieved by our approach, the more information will be extracted by the feature extractor. The detailed architecture of the attention reconstruction module is shown in Fig. 3. The input of this module is the feature maps extracted by the CNN backbone. This module includes two branches: the attention map computation branch and the reconstruction branch. The input of the attention reconstruction module is the feature maps extracted from dermoscopic or clinical images. In the attention map computation branch, we first adopt the spatial squeeze and channel excitation (cSE) (Roy et al., 2018) to compute the importance of the feature maps. Next, we use the weights to represent the importance. The cSE branch uses one convolutional layer to transform the input feature maps into a feature vector. Then, three fully-connected layers are connected to compute the weights for input feature maps. The neurons of the three layers are the number of feature map channels c, $\frac{c}{2}$, and c. By applying the weights, we emphasize the important feature maps from the input feature maps when we compute the weighted sum of all the feature maps to get the attention map m. Finally, we resize m into the input image size by using bicubic interpolation and normalize the values of m into the range of [0,1] by using $\hat{m} = \frac{m - \min(m)}{\max(m) - \min(m)}$, where $\hat{m}$ denotes the normalized attention map, $\min(\cdot)$ and $\max(\cdot)$ denote the functions that compute the minimum value and maximum value of m, respectively.

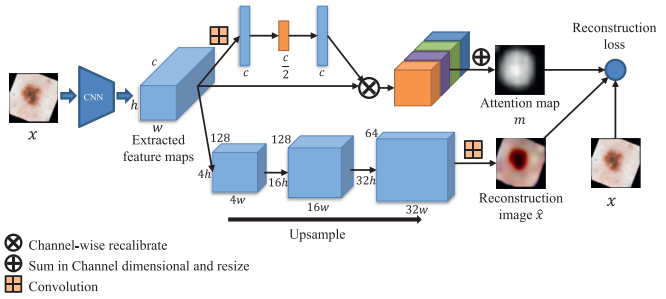


Fig. 3. The architecture of the attention-based reconstruction module. The upper branch is the cSE branch, which is used to compute the attention map. The lower branch is the reconstruction branch, which is used to reconstruct the input image. Lastly, we combine two branches to compute and update the parameters.

We use three upsampling blocks and one convolutional layer to reconstruct the input image in the reconstruction branch. Each upsampling block contains one upsampling layer, one convolutional layer, one batch normalization layer, and a leaky Rectified Linear Unit (ReLU) activation function. The scale factor of the first two upsampling blocks is four, and the followed convolutional layer's kernel size is 3×3 with 128 output channels. The scale factor of the third upsampling block is two. Then, the followed convolutional layer (3×3 kernel size) has 64 output channels. The last convolutional layer's kernel size is 3×3. The output of the last convolutional layer is the reconstructed RGB image $\hat{x} \in \mathbb{R}^{w \times h \times 3}$. Specifically, $\hat{x}^c = RC(G(x^c, \theta_G), \theta_{RC})$ and $\hat{x}^d = RD(G(x^d, \theta_G), \theta_{RD})$ are the reconstruction images for clinic and dermoscopic images by using reconstruct branch $RC(\cdot)$ and $RD(\cdot)$, respectively.

After obtaining the attention map and the reconstruction image, we construct the loss function to optimize the entire network. We define a new reconstruction loss function for each clinical image x^c and dermoscopic image x^d using the attention map $\hat{m}$ to constrain the feature extractor to focus more on the lesion. The loss function is the L2-norm reconstruction loss. After employing the attention mechanism, we introduce an exponential function applied to our attention map. The goal of adding an exponential function is to scale the loss of each pixel. Thus, making the reconstruction focus more on the attention area. The larger values in $\hat{m}$, the losses of the pixels will be more significant. For example, if one pixel is salient in the attention map, the value of this pixel is close to 1. Then, the loss weight of this pixel should be $e^1 = e$ times, and the entire network will pay more attention to learning the feature of this pixel. On the contrary, if one pixel is not salient in the attention map, the value of this pixel in $\hat{m}$ should be close to 0. Then, the loss weight of this pixel should be 1. At this time, the loss of this pixel degrades into the normal L2-norm loss. The detailed loss functions are as follows:

$$\begin{cases} \mathcal{L}_{RC}(x^c; \theta_G, \theta_{RC}) = \frac{1}{w \times h} \sum_{i=1}^{w} \sum_{j=1}^{h} e^{\hat{m}^c_{ij}} (\hat{x}^c_{ij} - x^c_{ij})^2, \\ \mathcal{L}_{RD}(x^d; \theta_G, \theta_{RD}) = \frac{1}{w \times h} \sum_{i=1}^{w} \sum_{j=1}^{h} e^{\hat{m}^d_{ij}} (\hat{x}^d_{ij} - x^d_{ij})^2, \end{cases} \tag{2}$$

where $\hat{m}^c_{ij}$ and $\hat{m}^d_{ij}$ denote the value of the i-th column and the j-th row of attention map $\hat{m}^c$ for clinical image and $\hat{m}^d$ for dermoscopic image, respectively. $\mathcal{L}_{RC}$ and $\mathcal{L}_{RD}$ denote the loss functions for clinical modality and dermoscopy modality, and w and h denote the width and height of the input image.

3.2.3. Classification module

The classification module is utilized to classify the multimodal input feature maps into their corresponding categories. Firstly, this module concatenates the feature maps from both modalities to obtain complementary information. Then the concatenated feature maps will go through a network of three layers. The first layer is a global average pooling layer which turns the input feature maps into feature vectors. Then, it is followed by a fully-connected layer with 512 neurons. Finally, the last layer is the softmax layer, which is used to classify the inputs into different lesion categories.

In our method, there are eight classifiers to achieve the classification of the DIAG and seven sub-tasks of 7-point criteria. The loss

Table 2
Comparison of different ablation study models in terms of accuracy (%).

Models	BWV	DaG	PIG	PN	RS	STR	VS	DIAG	Avg.
Clin. only	83.0	53.7	64.1	55.2	72.4	66.3	80.0	64.8	67.4
Derm. only	87.1	60.0	66.6	66.1	78.5	71.1	80.5	69.4	72.4
Concat.	88.6	62.0	67.1	70.1	79.7	74.7	82.0	70.1	74.3
Concat.+Recon.	89.4	59.5	69.1	68.6	81.0	74.4	82.8	73.2	74.8
Concat.+Recon.+Att.(without E)	89.4	60.8	67.3	**71.1**	80.3	74.2	81.5	70.9	74.4
Concat.+Recon.+Att.	**89.9**	**63.8**	69.6	68.9	77.5	**77.2**	82.8	72.7	75.3
Concat.+AD.	89.4	63.0	69.1	69.4	**82.3**	75.4	82.3	69.6	75.1
AMFAM	88.1	**63.8**	**70.9**	70.6	80.8	74.7	**83.3**	**75.4**	**76.0**

function of the classification module is defined as follows:

$$\mathcal{L}_P(x^c, x^d, Y, \hat{Y}; \theta_G, \theta_C) = \sum_{i}^{T} \mathcal{L}_{P_i}(x^c, x^d, y^i, \hat{y}^i; \theta_G, \theta_C), \tag{3}$$

where $\mathcal{L}_{P_i}(x^c, x^d, y^i, \hat{y}^i; \theta_G, \theta_C)$ is the standard cross-entropy loss for each classification task. $T = 8$ is the total number of classifiers. $Y = [y^1, \ldots, y^8]$ are the ground truth labels for the eight tasks and $\hat{Y} = [\hat{y}^1, \ldots, \hat{y}^8]$ are the predict labels for the eight tasks.

At last, we combine all the loss functions illustrated above to obtain the final loss function used to optimize the proposed AMFAM. Mathematically, the final loss function is shown in Eq. 4.

$$\begin{aligned}\mathcal{L} = \; & \mathcal{L}_P(x^c, x^d, Y, \hat{Y}; \theta_G, \theta_C) + \lambda \mathcal{L}_D(x^c, x^d; \theta_G, \theta_D) \\ & + \gamma\left(\mathcal{L}_{RC}(x^c, x^d; \theta_G, \theta_{RC}) + \mathcal{L}_{RD}(x^c, x^d; \theta_G, \theta_{RD})\right),\end{aligned} \tag{4}$$

where λ and γ are two trade-off parameters to control the contribution of reconstruction loss and discriminator loss, respectively.

4. Experimental study

4.1. Experimental settings

4.1.1. Training and testing details

We implement our proposed method using the PyTorch[1] library. We run all the training and testing processes on an NVIDIA QUADRO RTX 8000 GPU with 48 GB memory. For a fair comparison, we use ResNet-50 (He et al., 2016) as our CNN backbone, which keeps the same with HcCNN. The backbone is initialized with the ImageNet pre-trained parameters. During the training process, we used Adam (Kingma and Ba, 2014) optimizer with learning rate $l = 0.00001$ and weight decay $wd = 0.0001$ to optimize the entire neural network. We set batch size $b = 8$ and max training epoch $E = 150$ with the early stop. After grid search, we set the trade-off parameters $\lambda = 0.8$ and $\gamma = 0.4$. We use online data argumentation methods to augment training images from different modalities. Specifically, we firstly resize the input images into 299×299 pixels. Then, we augment each image from two modalities by using the following operations: padding 20 pixels with zero values, random cropping 299×299, random rotation (rotation angle $\phi \in [-45^\circ, 45^\circ]$), and random vertical flipping with the probability of 0.5. We only use the well-trained CNN backbone and classifiers to perform the classification task on the test set during the testing process.

4.1.2. Evaluation metrics

We evaluate the proposed method by using the following five widely-used evaluation metrics: accuracy (Acc.), sensitivity (Sen.), specificity (Spec.), precision (Prec.), and the area under the receiver operator characteristic curve (AUC). The definitions of accuracy, sensitivity, specificity and precision are as follows:

$$\begin{cases} Acc = \frac{TP+TN}{TP+FP+TN+FN} \\ Sen = \frac{TP}{TP+FN} \\ Spec = \frac{TN}{FP+TN} \\ Prec = \frac{TP}{TP+FP} \end{cases}, \tag{5}$$

where TP, FP, TN and FN are the numbers of the true positive, false positive, true negative and false negative samples, respectively. We keep the threshold of accuracy, sensitivity, specificity, and precision as the same with Reference (Bi et al., 2020) for a fair comparison, that is 0.5. AUC is computed by following the approach reported by Kawahara et al. (Kawahara et al., 2018), where the authors used the one-against-all approach to evaluate the labels positively related to the melanoma diagnosis. For all these metrics, the larger values indicate better performance. For a fair comparison with other methods, we use the average value on all these metrics to evaluate the performance of the models by following Reference (Kawahara et al., 2018; Bi et al., 2020) and considering the 7-point criteria dataset is a multi-task dataset.

4.2. Experimental results and analysis

4.2.1. Ablation study

To evaluate the effectiveness of each module of AMFAM, we compare seven different settings based on the ResNet-50 backbone. 1) Only using clinical modality to classify lesions (Clin. only); 2) Only using dermoscopy modality to classify lesions (Dermo. only); 3) Simply concatenating clinical modality feature representations and dermoscopy modality feature representations in the high-level feature space to classify lesions (Concat.); 4) Based on 3), adding reconstruction part to restrain the feature maps (Concat. + Recon.); 5) Based on 4), adding attention map to the reconstruction part and training by L2-norm reconstruction loss without the exponential weighting function (Concat. + Recon. + Att. (without E)); 6) Based on 4), adding attention map to the reconstruction part with our proposed exponential weight function (Concat. + Recon. + Att.); 7) Based on 3), adding adversarial discriminator (Concat. + AD.); and 8) Our proposed AMFAM, based on 5) and adding adversarial discriminator to restrain the feature maps of two modalities further.

The results of the ablation study models are reported in Table 2, in which the bold number and underline number mean the best and the second-best accuracy in each task. "Avg." denotes the average score over the entire row. Table 2 shows that all the multimodal learning models outperform the models using a single modality, which indicates the significance of multimodal fusion. Secondly, we can see that adding the reconstruction and the adversarial discriminator into the concatenation model can significantly improve the performance. It indicates that learning both correlated and complementary information is crucial. In addition, the attention mechanism improves the performance of classification. Adding reconstruction, attention map, and adversarial discriminator into

[1] https://pytorch.org/.

Table 3
The detailed comparison of ablation study models in terms of sensitivity (Sen.), specificity (Spec.), precision (Prec.), and AUC (%).

7pt criteria		BWV		DaG			PIG			PN			RS		STR			VS			Avg.
Method	Met.	ABS	PRS	ABS	REG	IR	ABS	REG	IR	ABS	TYP	ATP	ABS	PRS	ABS	REG	IR	ABS	REG	IR	
Clin. only	Sen.	85.2	61.1	58.3	48.3	55.7	66.0	11.3	59.3	63.1	60.4	40.0	76.9	47.5	74.3	69.2	44.1	80.2	66.7	0.0	56.2
	Spec.	61.1	85.2	76.8	78.0	74.6	67.3	88.5	79.1	73.6	75.1	83.6	47.5	76.9	57.4	90.8	83.3	83.3	87.7	92.4	77.0
	Prec.	95.6	29.3	14.0	48.3	79.7	83.4	6.3	51.6	57.1	55.5	51.6	88.9	27.4	80.9	20.5	47.9	99.7	7.7	0.0	49.8
	AUC	81.6	81.6	68.1	74.2	73.1	72.3	68.6	76.4	73.3	75.9	69.6	66.6	66.6	73.4	83.5	73.4	80.9	80.9	74.9	74.5
Derm. only	Sen.	87.2	85.3	61.7	52.8	64.8	65.1	72.2	73.6	73.3	74.2	46.4	83.1	62.1	75.5	1.0	52.4	83.1	48.3	0.0	61.2
	Spec.	85.3	87.3	81.2	83.0	74.3	83.1	90.7	75.1	83.8	80.1	85.3	62.1	83.1	68.8	90.9	83.9	69.0	89.6	92.4	81.5
	Prec.	98.4	38.7	37.0	63.6	70.6	94.6	27.1	31.5	75.6	63.0	54.8	88.6	50.9	88.7	20.5	46.8	97.1	26.9	0.0	56.5
	AUC	87.4	87.4	76.7	75.8	77.2	79.2	82.9	79.6	86.9	81.4	77.1	80.0	80.0	79.7	86.9	77.1	84.2	85.3	71.1	80.8
Concat.	Sen.	90.1	78.8	68.9	52.7	67.5	69.1	64.7	62.3	74.0	71.8	58.8	81.4	71.0	82.7	67.7	55.1	85.1	53.8	0.0	66.1
	Spec.	78.8	90.1	80.3	84.0	78.5	71.5	90.2	79.9	87.4	82.6	85.4	71.0	81.4	71.3	93.7	86.5	74.4	91.3	92.4	82.7
	Prec.	96.6	54.7	31.0	66.9	76.3	84.3	22.9	53.2	82.1	70.0	50.5	93.8	41.5	85.6	47.7	57.4	96.8	40.4	0.0	60.6
	AUC	86.7	86.7	75.5	77.8	81.5	78.9	81.4	82.2	89.2	84.1	84.9	85.1	85.1	84.8	88.6	82.4	89.0	89.0	77.9	83.7
Concat. +Recon.	Sen.	91.1	78.9	66.7	52.3	62.4	70.6	58.3	66.3	80.6	68.7	53.4	83.7	69.6	82.5	63.9	54.9	85.6	58.5	0.0	65.7
	Spec.	78.9	91.1	78.8	81.1	79.5	78.8	89.3	80.6	80.5	85.3	87.0	69.6	83.9	71.7	94.2	85.5	75.6	92.0	92.4	82.9
	Prec.	96.3	0.6	24.0	57.6	80.8	89.2	14.6	54.0	66.7	76.7	59.1	91.7	51.9	86.0	52.3	53.2	96.8	46.2	0.0	57.8
	AUC	88.7	88.7	72.8	74.7	80.4	80.9	81.4	83.7	87.7	85.5	82.1	82.4	82.4	84.7	90.0	81.1	87.8	88.2	80.1	83.3
Concat.+Recon.	Sen.	92.1	75.4	65.7	53.0	65.6	68.4	53.8	66.0	76.5	73.9	57.1	83.5	67.5	83.3	72.0	54.0	82.1	68.8	0.0	66.2
+Att.(without	Spec.	75.4	92.1	78.6	84.4	78.5	75.7	89.3	79.4	86.3	83.8	86.5	67.5	83.5	68.8	93.0	88.3	87.5	89.2	92.4	83.2
E)	Prec.	95.0	65.3	23.0	67.8	77.4	88.3	14.6	50.0	79.5	71.9	55.9	91.0	50.9	83.3	40.9	64.9	99.4	21.2	0.0	60.0
	AUC	90.3	90.3	72.9	75.7	80.0	79.1	80.4	81.9	89.6	85.3	82.8	84.4	84.4	83.9	88.6	82.5	88.6	88.0	79.4	83.6
Concat.	Sen.	91.4	80.7	76.7	51.1	73.0	69.5	55.0	74.0	70.9	70.9	61.0	79.2	66.0	84.4	69.7	61.0	83.6	70.8	0.0	67.8
+Recon.	Spec.	80.7	91.4	81.0	86.9	78.3	83.5	90.1	78.9	84.8	81.9	86.3	66.0	79.2	72.9	94.2	88.8	87.5	90.6	92.4	84.0
+Att.	Prec.	96.6	61.3	33.0	75.4	73.4	92.8	22.9	46.0	78.2	68.5	53.8	93.8	33.0	86.0	52.3	64.9	99.0	32.7	0.0	61.2
	AUC	92.4	92.4	74.2	78.0	82.0	79.8	82.3	81.7	88.0	84.3	85.4	82.1	82.1	86.3	90.8	84.5	87.5	87.3	79.4	84.2
Concat.	Sen.	90.6	81.1	68.0	54.4	68.4	67.1	85.7	76.4	74.1	69.9	60.0	85.9	70.0	81.2	59.4	60.5	84.8	59.0	0.0	**68.2**
+AD.	Spec.	81.1	90.6	80.9	85.0	78.4	86.1	89.2	78.6	84.5	81.7	87.2	70.0	85.9	77.8	93.1	85.0	71.8	91.9	92.4	83.7
	Prec.	96.9	57.3	34.0	68.6	75.7	95.1	12.5	44.4	76.9	68.5	58.1	90.7	59.4	90.7	43.2	48.9	96.5	44.2	0.0	61.1
	AUC	90.3	90.3	75.9	78.6	81.4	81.9	82.7	81.9	87.6	84.7	83.9	85.5	85.5	85.8	89.7	81.7	89.1	88.6	81.3	84.5
AMFAM	Sen.	90.3	75.0	68.4	56.3	66.7	72.7	63.2	67.9	77.2	70.3	58.5	82.6	72.1	80.6	70.4	57.3	83.8	75.0	0.0	67.8
	Spec.	75.0	90.3	82.0	81.5	82.4	76.3	90.4	83.0	85.7	84.6	85.6	72.1	82.6	72.4	93.2	85.9	91.7	90.8	92.4	**84.1**
	Prec.	95.6	56.0	39.0	56.8	82.5	86.1	25.0	61.3	78.2	74.7	51.6	93.4	46.2	87.5	43.2	54.3	99.4	34.6	0.0	**61.3**
	AUC	91.1	91.1	77.1	77.7	81.9	80.7	85.1	83.4	89.2	84.5	82.0	86.7	86.7	83.0	89.5	80.7	89.6	88.8	80.9	**84.7**

Table 4
The detailed comparison of ablation study models for the DIAG classification in terms of sensitivity (Sen.), specificity (Spec.), precision (Prec.), and AUC (%).

7pt criteria		DIAG					Avg.
Method	Met.	BCC	NEV	MEL	MISC	SK	
Clin. only	Sen.	0.0	75.2	49.2	40.0	0.0	32.9
	Spec.	95.9	75.8	88.7	90.6	95.2	89.2
	Prec.	0.0	83.1	69.3	10.0	0.0	32.5
	AUC	83.1	81.6	79.6	79.1	67.4	78.2
Derm. only	Sen.	14.3	73.0	71.1	34.8	0.0	38.6
	Spec.	96.1	88.5	86.5	91.4	95.2	91.6
	Prec.	6.3	94.1	58.4	20.0	0.0	35.7
	AUC	90.5	88.4	85.2	89.1	73.6	85.3
Concat.	Sen.	31.6	73.5	66.7	87.5	0.0	51.8
	Spec.	97.3	85.8	87.1	91.5	95.2	91.4
	Prec.	37.5	92.2	61.4	17.5	0.0	41.7
	AUC	91.0	87.6	87.6	85.5	79.8	86.3
Concat. +Recon.	Sen.	36.4	76.6	69.7	25.0	0.0	41.5
	Spec.	96.9	89.7	89.2	92.1	95.2	92.6
	Prec.	25.0	94.1	68.3	62.5	0.0	50.0
	AUC	90.7	90.5	89.3	92.5	78.0	88.2
Concat. +Recon. +Att.(without E)	Sen.	37.5	75.5	64.9	45.5	0.0	44.7
	Spec.	96.6	85.4	89.8	90.9	95.2	91.6
	Prec.	18.8	91.3	71.3	12.5	0.0	38.8
	AUC	89.3	88.7	88.7	87.9	72.2	85.4
Concat. +Recon. +Att.	Sen.	37.5	75.5	74.4	48.1	5.3	48.2
	Spec.	96.6	89.3	88.0	92.7	95.4	92.4
	Prec.	18.8	94.1	63.4	32.5	100.0	**61.7**
	AUC	91.9	90.9	90.3	89.5	82.6	**89.0**
Concat.+AD.	Sen.	30.0	70.7	75.7	75.0	0.0	50.3
	Spec.	97.3	90.8	85.2	91.2	95.2	92.0
	Prec.	37.5	95.9	52.5	15.0	0.0	40.2
	AUC	91.0	89.7	89.3	91.1	76.1	87.4
AMFAM	Sen.	40.0	84.1	65.8	68.0	40.0	**59.6**
	Spec.	97.4	85.8	91.4	93.8	95.6	**92.8**
	Prec.	37.5	89.5	76.2	42.5	10.0	51.1
	AUC	94.1	89.7	89.1	90.6	81.7	**89.0**

the concatenation model one by one to restrain the CNN backbone can gradually improve the classification performance. Similarly, adding our proposed exponential function to add weights to the reconstruction loss of pixels can improve the model performance. Finally, we observe that our proposed AMFAM model achieves the best performance compared to all other models with different settings. It means that learning both the correlated information and complementary information can further improve the multimodal classification performance. The classification results in Table 2 demonstrate the effectiveness of each part of our proposed method.

To further verify the effectiveness of each component in our proposed method, we report the results of other metrics achieved by the ablation study models in Tables 3 and 4. Table 3 shows the 7-point criteria results and Table 4 shows the results of DIAG. The results in Table 3 show that the average AUC of our proposed method is 3.9% higher than using a single modality. The results (in terms of all other metrics) increase gradually when reconstruction, attention map, and adversarial discriminator components are added to the basic concatenate model. Both Tables 3 and 4 demonstrate that the performance of adding reconstruction, attention and adversarial learning parts can improve the concatenate model's performance. These results also demonstrate the consistent conclusion with Table 2 that learning correlated and complementary information together can improve multimodal classification performance.

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

In this section, we compare the performance of our proposed method with the current state-of-the-art (SOTA) methods on the 7-point criteria evaluation database. All the methods are evaluated with the same database. Firstly, we compare three baseline methods proposed by Kawahara et al. (Kawahara et al., 2018). These methods are based on Inception-V3. They are the Inception-unbalanced method that directly sampled data into mini-batch to train the model; The Inception-balanced method sampled a balanced mini-batch according to the samples' label to train the model. The Inception-combined method combines three modalities to classify the lesion. All the results of these three methods come from Reference (Kawahara et al., 2018). Then, we compare several multimodal classification methods, i.e., TripleNet (Ge et al., 2017), EmbeddingNet (Yap et al., 2018), and HcCNN (Bi et al., 2020). Among all the methods, Inception unbalanced, Inception-balanced, Inception-combined, TripleNet, and EmbeddingNet fuse the representations of different modalities by concatenating the representations from different modalities at the end of CNN backbones. HcCNN is a gradually fused multimodal method that uses a separate CNN branch to fuse the features of different modalities. The details of these multimodal methods have been introduced in the related work section. The classification results of these comparison methods come from Reference (Bi et al., 2020). The average results of all the methods are computed using all the categories' results as Kawahara et al. (Kawahara et al., 2018) do, except for HcCNN. The average results for HcCNN are computed by using the numbers provided in the original paper.

Firstly, the results of the eight classification tasks in terms of accuracy are reported in Table 5. From Table 5, the average accuracies of different models indicate that most of the concatenation-based multimodal fusion methods' results are worse than the gradually fused HcCNN method, except for our proposed method. The main reason why the concatenation-based fusion method performs worse than HcCNN is that even the concatenation-based fusion method contains complementary information of different modalities. The features from each modality are extracted separately,

Table 5
Comparison between our proposed method with the SOTA multimodal learning methods in terms of accuracy (%).

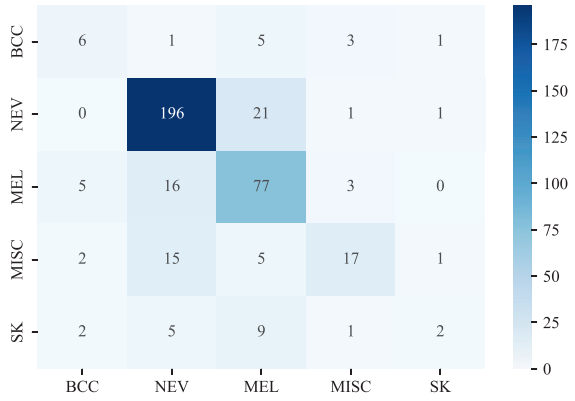
Methods	BWV	DaG	PIG	PN	RS	STR	VS	DIAG	Avg.
Inception-unbalanced (Kawahara et al., 2018)	87.6	56.7	65.6	68.1	78.2	**75.9**	81.3	68.4	72.7
Inception-balanced (Kawahara et al., 2018)	87.3	60.3	64.8	68.9	78.2	75.7	81.5	70.9	73.5
Inception-combined (Kawahara et al., 2018)	87.1	60.0	66.1	**70.9**	77.2	74.2	79.7	74.2	73.7
TripleNet (Ge et al., 2017)	87.9	61.3	67.3	63.3	76.0	74.4	83.0	68.6	72.7
EmbeddingNet (Yap et al., 2018)	84.3	57.5	64.3	65.1	78.0	73.4	82.5	68.6	71.7
HcCNN (Bi et al., 2020)	87.1	**65.6**	68.6	70.6	**80.8**	71.6	**84.8**	69.9	74.9
AMFAM	**88.1**	63.8	**70.9**	70.6	**80.8**	74.7	83.3	**75.4**	**76.0**

which may cause the sub-optimal solution (Bi et al., 2020). Though our method also concatenates the feature maps, it employs adversarial learning and attention mechanism-based reconstruction to learn highly discriminative features which contain correlated and complementary information. Thus, our method achieves the best average results among all the comparison methods. Even though HcCNN introduces a new branch to fusion the features, it only learns the complementary information. Besides, the additional neural network branch makes the model of HcCNN very complicated and requires a lot of additional computing resources. Our proposed method outperforms other methods in the eight classification tasks on most of the tasks (4 out of 8). One of the potential reasons is that by employing an attention mechanism-based mechanism to restrain the feature extractor, the feature representation of each modality is more discriminative. Besides, adversarial learning guarantees representations from different modalities to gain correlated information.

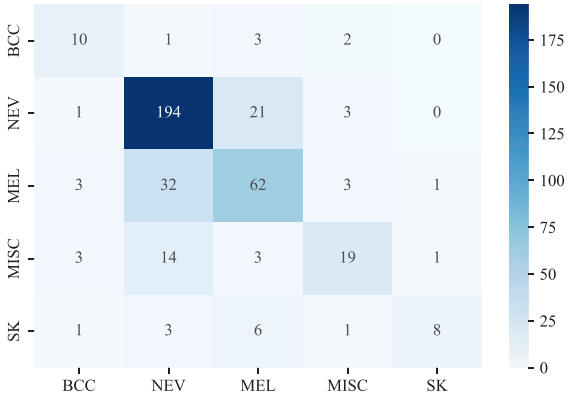
Considering the poor performance of TripleNet and EmbeddingNet, when comparing the performance of other classification metrics, we only compare Inception-unbalanced, Inception-balanced, Inception-combined, and HcCNN with our AMFAM method. The detailed results are reported in Table 6. Our proposed method achieves the best performance on most metrics (3 out of 4). Specifically, our proposed method improves more than 6% of sensitivity and specificity on this dataset and 2.7% AUC compared to the previous SOTA method (HcCNN).

4.2.3. Comparison on the DIAG classification task

Following the setting in Reference (Kawahara et al., 2018), we compare our method with the Inception-unbalanced, Inception-balanced, Inception-combined, and HcCNN on the DIAG classification task. For Inception-unbalanced, Inception-balanced, and Inception-combined methods, we use the results provided in the original paper (Kawahara et al., 2018). Since the authors of HcCNN did not provide the DIAG classification results, we re-implement their method according to the original paper (Bi et al., 2020). The results of our re-implementation can achieve similar results in terms of average accuracy and AUC on eight sub-tasks. Specifically, the reported average accuracy and AUC are 74.9% and 82.0% (Bi et al., 2020), respectively, and the results of our re-implemented code are 74.5% and 82.5% for average accuracy and AUC, respectively. Table 7 reports the detailed comparison results, which indicates that our proposed method achieves the best results in precision and second-best sensitivity and AUC. However, the average performance on its precision is worse than other methods. It is caused by the low precision of the SK category. To find out the reason for this phenomenon, we plot the confusion matrices of our method and the Inception-combined method using the test set predictions. The confusion matrices are shown in Fig. 4a. We can see that our proposed method performs better on the NEV and MEL categories than the Inception-combined method do but slightly worse on other categories. Table 1 shows that this database is unbalanced. The SK category only has 16 images in the training set. The NEV and MEL categories are the two majority



(a) The confusion matrix of our proposed method.



(b) The confusion matrix of the Inception-combined method.

Fig. 4. Visualization of the confusion matrices the test set predictions. The *x* axis is the output label of the model, and the *y*-axis is the ground truth label.

categories. Our method does not consider and aim at solving the unbalanced problem, resulting in its performance slightly worse on the minority categories than the Inception-combined methods, which use a data selection approach to handle the data unbalance issue.

We also plot the ROC curves in Fig. 5, from which one can see that the curves for all categories except the SK category have similar area sizes (around 90.0%) under the ROC curve. For the SK category, its ROC curve is obviously lower than other categories, its area under the ROC curve is less than 80.0%. Overall, these results show that our method can achieve a promising performance for skin lesion classification.

4.2.4. Visualization

We further use gradient-weighted class activation maps (Selvaraju et al., 2017) (Grad-CAM) to demonstrate AMFAM's effectiveness visually. Grad-CAM is a technique for producing visual explanations for decisions from a large class of CNN-based

Table 6
The detailed classification results (%) of our method and the SOTA methods for the 7-point criteria classification task.

7pt criteria		BWV		DaG			PIG			PN			RS		STR			VS			Avg.
Method	Met.	ABS	PRS	ABS	REG	IR	ABS	REG	IR	ABS	TYP	ATP	ABS	PRS	ABS	REG	IR	ABS	REG	IR	
Inception-unbalanced (Kawahara et al., 2018)	Sen.	96.6	49.3	34.0	59.3	67.8	83.0	6.2	57.3	78.8	77.4	35.5	95.5	31.1	98.1	36.4	34.0	98.7	23.1	0.0	55.9
	Spec.	49.3	96.6	92.2	72.2	67.4	53.5	99.4	80.1	80.8	75.5	93.7	31.1	95.5	47.8	98.6	94.0	22.0	97.1	100.0	76.1
	Prec.	89.0	77.1	59.6	47.6	62.8	69.8	60.0	56.8	72.8	64.9	63.5	79.1	71.7	77.8	76.2	64.0	82.8	54.5	0.0	64.7
	AUC	87.0	87.0	72.3	72.6	76.4	77.4	67.2	78.1	87.8	83.6	78.6	79.9	79.9	84.2	87.8	78.3	82.1	81.8	73.4	79.8
Inception-balanced (Kawahara et al., 2018)	Sen.	92.5	65.3	43.0	66.1	66.1	73.5	16.7	67.7	78.2	76.0	41.9	84.1	62.3	90.7	43.2	50.0	96.8	30.8	10.0	60.8
	Spec.	65.3	92.5	89.8	75.1	73.4	64.5	98.3	73.4	81.6	77.9	92.1	62.3	84.1	63.8	97.4	87.7	31.7	95.6	99.5	79.3
	Prec.	91.9	67.1	58.9	53.1	66.9	72.9	57.1	53.8	73.5	66.9	61.9	85.9	58.9	82.3	67.9	56.0	84.4	51.6	60.0	66.9
	AUC	87.5	87.5	73.0	76.5	78.0	78.8	75.2	79.4	88.6	83.6	78.9	83.5	83.5	84.9	87.1	78.7	85.0	84.0	76.1	81.6
Inception-combined (Kawahara et al., 2018)	Sen.	89.4	77.3	47.0	67.8	62.1	77.6	29.2	59.7	77.6	78.1	48.4	81.3	66.0	86.0	54.5	51.1	92.3	42.3	13.3	63.2
	Spec.	77.3	89.4	87.8	72.6	78.9	65.1	94.2	80.1	85.8	78.7	90.7	66.0	81.3	67.4	96.0	85.7	45.1	92.4	97.5	80.6
	Prec.	94.4	63.0	56.6	51.3	70.5	74.2	41.2	57.8	78.1	68.3	61.6	86.7	56.5	83.1	63.2	52.7	86.5	45.8	30.8	64.3
	AUC	89.2	89.2	74.1	76.5	79.9	79.0	74.9	79.0	89.9	84.2	79.9	82.9	82.9	86.1	87.0	78.9	86.2	85.5	76.1	82.2
HcCNN (Bi et al., 2020)	Sen.	-	92.2	-	-	80.2	-	-	55.7	-	-	40.9	-	95.2	-	-	35.1	-	-	20.0	59.9
	Spec.	-	65.3	-	-	71.6	-	-	86.3	-	-	92.4	-	41.5	-	-	90.0	-	-	98.4	77.9
	Prec.	-	91.9	-	-	69.6	-	-	65.1	-	-	62.3	-	81.6	-	-	52.4	-	-	50.0	**67.6**
	AUC	-	89.8	-	-	82.6	-	-	81.3	-	-	78.3	-	81.9	-	-	77.6	-	-	82.7	82.0
AMFAM	Sen.	90.3	75.0	68.4	56.3	66.7	72.7	63.2	67.9	77.2	70.3	58.5	82.6	72.1	80.6	70.4	57.3	83.8	75.0	0.0	**67.8**
	Spec.	75.0	90.3	82.0	81.5	82.4	76.3	90.4	83.0	85.7	84.6	85.6	72.1	82.6	72.4	93.2	85.9	91.7	90.8	92.4	**84.1**
	Prec.	95.6	56.0	39.0	56.8	82.5	86.1	25.0	61.3	78.2	74.7	51.6	93.4	46.2	87.5	43.2	54.3	99.4	34.6	0.0	61.3
	AUC	91.1	91.1	77.1	77.7	81.9	80.7	85.1	83.4	89.2	84.5	82.0	86.7	86.7	83.0	89.5	80.7	89.6	88.8	80.9	**84.7**

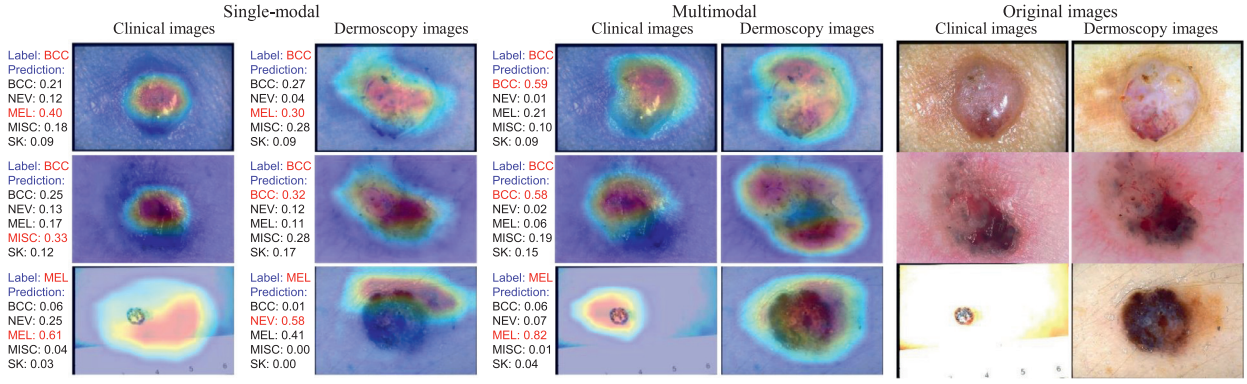
Table 7
The detailed classification results (%) of our method on the SOTA methods for the DIAG classification task.

Method	Met.	BCC	NEV	MEL	MISC	SK	Avg.
Inception-unbalanced (Kawahara et al., 2018)	Sen.	25.0	94.1	44.6	35.0	5.3	40.8
	Spec.	98.4	50.6	92.2	98.0	99.5	87.7
	Prec.	40.0	70.3	66.2	66.7	33.3	55.3
	AUC	92.2	87.7	83.2	86.3	88.7	87.6
Inception-balanced (Kawahara et al., 2018)	Sen.	25.0	91.3	55.4	42.5	15.8	46.0
	Spec.	98.9	62.5	88.4	97.2	99.7	89.3
	Prec.	50.0	75.2	62.2	63.0	75.0	65.1
	AUC	89.2	88.1	84.2	86.8	90.4	87.7
Inception-combined (Kawahara et al., 2018)	Sen.	62.5	88.6	61.4	47.5	42.1	60.4
	Spec.	97.9	71.6	88.8	97.5	99.5	91.1
	Prec.	55.6	79.5	65.3	67.9	80.0	**69.7**
	AUC	92.9	89.7	86.3	88.3	91.0	**89.6**
HcCNN (Bi et al., 2020)	Sen.	35.3	79.4	68.8	58.1	66.7	**61.6**
	Spec.	97.4	86.7	85.4	95.7	95.7	92.2
	Prec.	37.5	91.3	54.5	62.5	10.5	51.3
	AUC	90.8	90.4	87.9	91.5	83.0	88.7
AMFAM	Sen.	40.0	84.1	65.8	68.0	40.0	59.6
	Spec.	97.4	85.8	91.4	93.8	95.6	**92.8**
	Prec.	37.5	89.5	76.2	42.5	10.0	51.1
	AUC	94.1	89.7	89.1	90.6	81.7	89.0

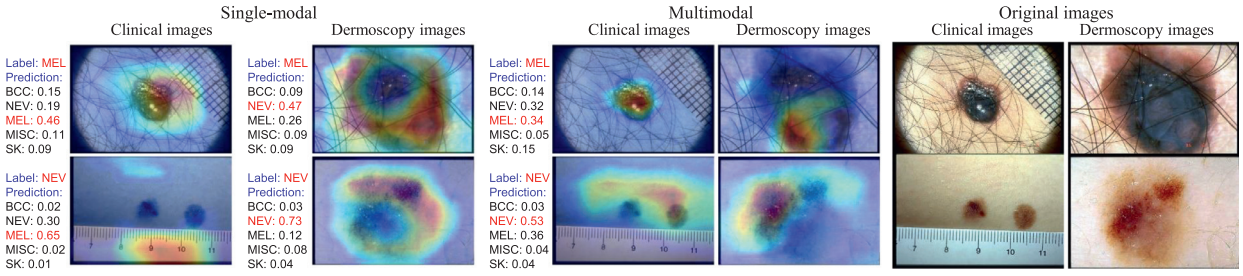
Fig. 5. ROC curves for each label in the DIAG classification task via the one-vs-all approach.

models, making them more transparent. It will highlight the important regions in the image for predicting the category.

Firstly, to evaluate the effectiveness of complementary and correlated information of our AMFAM method, we visualize five cases for the DIAG classification task, as shown in Fig. 6. In the figure, the first column represents the visualization of the clinical modality-based model; the second column represents the visualization of the dermoscopy modality-based model; the third and fourth columns represent the visualization of our proposed method for clinical images and dermoscopic images; the last two columns represent the original images for clinical and dermoscopy modalities. From the first two rows, we can see that even though the single-modal-based model makes decisions based on the right lesion area, it may misclassify the samples. For our proposed multimodal-based method, we can see that it not only can make decisions based on the right lesion area but also classify them accurately (with high confidence). Also, our proposed method can locate the lesion area in both modalities while predicting accurately when the single-modal-based model fails to focus precisely on the right lesion area, as shown in the third row of Fig. 6(a). It demonstrates that the correlated information learned from both modalities is helpful to improve the confidence of the final prediction in the lesion classification task. For the first and second rows of 6(b), we can see that the single-modal-based model can only predict accurately in either clinical modal or dermoscopy modal. Our proposed multimodal method can classify it accurately by consider-



(a) Samples that are classified correctly with high confidence by considering correlated information.



(b) Samples that are classified correctly by considering the complementary information from both modalities.

Fig. 6. Visualization results on evaluating the effectiveness of complementary and correlated information by using Grad-CAM. The first column is clinical modality-based visualization; the second column is dermoscopy modality-based visualization; the third and fourth columns are the clinical and dermoscopic images of multimodal-based visualization; and the last two columns are the original clinical and dermoscopic images. "Label" denotes the ground truth of the image. "Prediction" stands for the softmax output probabilities for each category of the DIAG classification task.

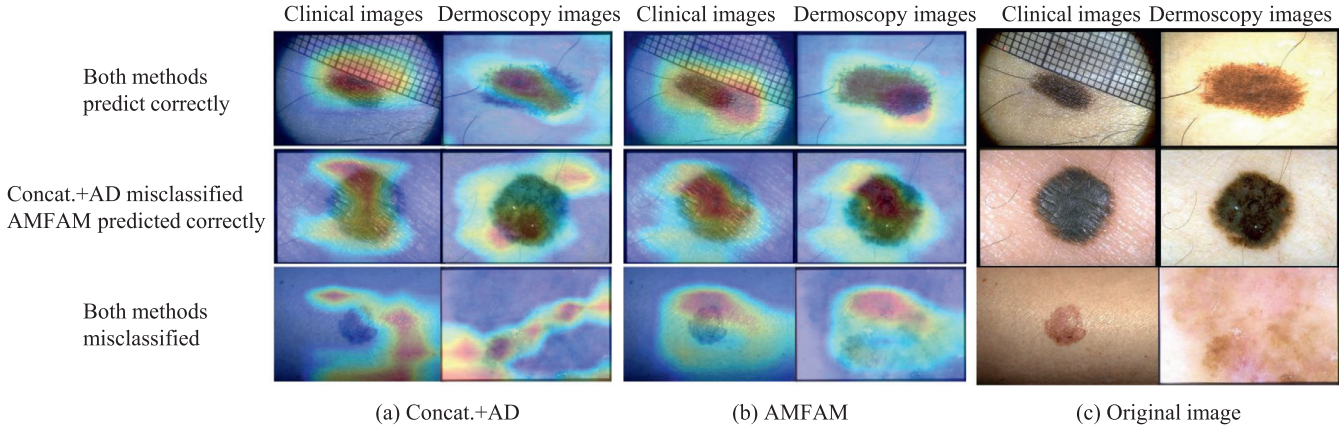


Fig. 7. Visualization results on evaluating the effectiveness of the attention mechanism-based module under three different classification situations.

ing the complementary information from both modalities. It indicates that the leaned complementary information of our proposed method is helpful to improve the model's performance.

Secondly, we demonstrate several Grad-CAM-based images from both clinical and dermoscopy modalities to show the effectiveness of attention mechanism-based reconstruction. We compare the visualization of our proposed method AMFAM with Concat.+AD (as mentioned in the ablation study). The difference between AMFAM and Concat.+AD. is that AMFAM used an attention mechanism-based reconstruction module to restrain the CNN backbone. We visualize three different classification situations of these two methods: 1) both methods classified correctly, 2) Concat.+AD misclassified but AMFAM classified correctly, and 3) both methods misclassified. Fig. 7 shows the specific visualization results. For all three different situations, we can see that for both clinical and dermoscopy modalities, the important areas focused by the AMFAM model are more compact and centered on the lesion. Even when both methods misclassified the sample, our AMFAM can still focus on the right lesions. It verifies the effectiveness of the attention mechanism-based reconstruction module. At the same time, from Figs. 6 and 7 we can find that our proposed multimodal learning method focuses on the lesion areas of both modalities for most of the samples. It verifies that the discriminator has been well trained, and the correlated information has been learned.

5. Discussion and future work

Although ablation studies and comparisons have proven the advantages of our proposed method, some phenomena need to be noted. First, the attention mechanism-based and adversarial train-

ing models in the ablation study do not improve the performance of every sub-tasks. For instance, the Concat. model achieves better accuracy on the DIAG category than Concat.+Recon.+Att and Concat.+AD. models, as shown in Table 2. By considering the data distribution, as shown in Table 1, most of the classification tasks are unbalanced classification tasks. The ratio of minority class and majority class vary from 1:1.72 (PN) to 1:16 (DIAG). We find that the performance improvement of attention-based mechanism and adversarial training models on the highly unbalanced sub-tasks is smaller than the less unbalanced sub-tasks. Moreover, the attention mechanism-based model improves different sub-tasks with the adversarial learning model. The potential reason is that the attention mechanism-based model and adversarial learning model enhance the features of the less unbalanced tasks but weaken the features of the highly unbalanced tasks. And the results also have shown that our proposed AMFAM method is sensitive to training samples' data distribution and number. If the dataset is highly unbalanced and the number of train samples is small, the performance of our method may decrease. Secondly, balanced and unbalanced models perform differently on the minority category for highly unbalanced sub-task. Table 6 and Table 7 show the results of two kinds of models which are balanced models and unbalanced models. The balanced models contain Inception-balanced and Inception-combined models, and the unbalanced models are the rest. We can see that even though our proposed method achieves the best performance via most metrics, the results in the highly unbalanced category are inferior to the balanced models, such as the IR category of VS classification task and SK category of DIAG classification task. Especially, our proposed model obtains 0% sensitivity and precision of IR category in VS classification task, but two balanced models obtain 10% and 13.3% via sensitivity and 60% and 30.8% via precision, respectively. However, for the less unbalanced category, such as DaG and BWV, our method can also achieve comparable results on the minority category compared to the balanced models and better performance on the majority category.

For the multimodal multi-task classification, a simple data sample strategy can not make each sub-classification task balanced. If we want the data samples of each sub-classification task to be balanced, we need a large batch size to contain the balanced samples as the Inception-balanced model do (Kawahara et al., 2018). However, when the dataset is small, increasing the batch size may cause poor generalization (Keskar et al., 2016). Thus, a potential research topic in future work is solving the unbalanced problem for the multi-task classification but under the limited data samples. Besides, combining efficient image-based modality and text modality information to improve the diagnosis performance further is worth studying.

6. Conclusion

In this study, to leverage multiple modalities of medical data, we proposed a multimodal deep neural network, AMFAM, for multimodal and multi-task skin lesion classification. Our proposed method can learn both correlated and complementary information from different modalities. Specifically, to learn the correlated information, we adopted adversarial learning to train the model. Furthermore, to make the CNN backbone pay more attention to the lesion for better extracting complementary information, we designed an attention-based reconstruction sub-network with a new loss function to force the network to learn the discriminate features of the lesion instead of the background. Then, we concatenated the extracted different modalities' features to obtain the complementary information. The comprehensive experiments on a publicly available dataset, 7-point criteria evaluation, demonstrated that our method can achieve the SOTA performance in most classification evaluation metrics and significantly improve the sensitivity and specificity by above 6% and AUC 2.7% compared to the previous SOTA methods.

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.

Acknowledgements

This work was supported by the Agency for Science, Technology and Research (A*STAR) through its AME Programmatic Funding Scheme under Project A20H4g2141.

References

Baltrušaitis, T., Ahuja, C., Morency, L.-P., 2018. Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell 41 (2), 423–443.
Barata, C., Celebi, M.E., Marques, J.S., 2017. Development of a clinically oriented system for melanoma diagnosis. Pattern Recognit 69, 270–285.
Bhardwaj, A., Rege, P.P., 2021. Skin lesion classification using deep learning. In: Advances in Signal and Data Processing. Springer, pp. 575–589.
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 Recognit 107, 107502.
Chen, M., Zhou, P., Wu, D., Hu, L., Hassan, M.M., Alamri, A., 2020. Ai-skin: skin disease recognition based on self-learning and wide data collection through a closed-loop framework. Information Fusion 54, 1–9.
Claridge, E., Cotton, S., Hall, P., Moncrieff, M., 2003. From colour to tissue histology: physics-based interpretation of images of pigmented skin lesions. Med Image Anal 7 (4), 489–502.
Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. The Journal of Machine Learning Research 17 (1), 2030–2096.
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: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 250–258.
Gessert, N., Sentker, T., Madesta, F., Schmitz, R., Kniep, H., Baltruschat, I., Werner, R., Schlaefer, A., 2020. Skin lesion classification using cnns with patch-based attention and diagnosis-guided loss weighting. IEEE Trans. Biomed. Eng. 67 (2), 495–503. doi:10.1109/TBME.2019.2915839.
Harangi, B., 2018. Skin lesion classification with ensembles of deep convolutional neural networks. J Biomed Inform 86, 25–32.
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, Q., Li, X., Kim, D.N., Jia, X., Gu, X., Zhen, X., Zhou, L., 2020. Feasibility study of a multi-criteria decision-making based hierarchical model for multi-modality feature and multi-classifier fusion: applications in medical prognosis prediction. Information Fusion 55, 207–219.
Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G., 2018. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE J Biomed Health Inform 23 (2), 538–546.
Kawahara, J., Hamarneh, G., 2016. Multi-resolution-tract cnn with hybrid pretrained and skin-lesion trained layers. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 164–171.
Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T.P., 2016. On large–batch training for deep learning: generalization gap and sharp minima. arXiv preprint arXiv:1609.04836.
Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.
Liu, K., Li, Y., Xu, N., Natarajan, P., 2018. Learn to combine modalities in multimodal deep learning. arXiv preprint arXiv:1805.11730.
Liu, Y., Fan, Q., Zhang, S., Dong, H., Funkhouser, T., Yi, L., 2021. Contrastive multimodal fusion with tupleinfonce. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 754–763.
Ma, L., Staunton, R.C., 2013. Analysis of the contour structural irregularity of skin lesions using wavelet decomposition. Pattern Recognit 46 (1), 98–106.
Massone, C., Hofmann-Wellenhof, R., Ahlgrimm-Siess, V., Gabler, G., Ebner, C., Peter Soyer, H., 2007. Melanoma screening with cellular phones. PLoS ONE 2 (5), e483.
Mendoza, C.S., Serrano, C., Acha, B., 2009. Scale invariant descriptors in pattern analysis of melanocytic lesions. In: 2009 16th IEEE International Conference on Image Processing (ICIP). IEEE, pp. 4193–4196.
Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C., 2021. Attention bottlenecks for multimodal fusion. Adv Neural Inf Process Syst 34.

Pereira, P.M., Thomaz, L.A., Tavora, L.M., Assuncao, P.A., Fonseca-Pinto, R.M., Paiva, R.P., de Faria, S.M., 2021. Melanoma classification using light-fields with morlet scattering transform and cnn: surface depth as a valuable tool to increase detection rate. Med Image Anal 102254.

Pérez, E., Reyes, O., Ventura, S., 2021. Convolutional neural networks for the automatic diagnosis of melanoma: an extensive experimental study. Med Image Anal 67, 101858.

Polat, K., Koc, K.O., 2020. Detection of skin diseases from dermoscopy image using the combination of convolutional neural network and one-versus-all. Journal of Artificial Intelligence and Systems 2 (1), 80–97.

Pomponiu, V., Nejati, H., Cheung, N.-M., 2016. Deepmole: Deep neural networks for skin mole lesion classification. In: 2016 IEEE International Conference on Image Processing (ICIP). IEEE, pp. 2623–2627.

Rigel, D.S., Friedman, R.J., Kopf, A.W., 1996. The incidence of malignant melanoma in the united states: issues as we approach the 21st century. J. Am. Acad. Dermatol. 34 (5), 839–847.

Roy, A.G., Navab, N., Wachinger, C., 2018. Concurrent spatial and channel squeeze & excitation in fully convolutional networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 421–429.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet large scale visual recognition challenge. International Journal of Computer Vision (IJCV) 115 (3), 211–252. doi:10.1007/s11263-015-0816-y.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626.

Song, L., Lin, J., Wang, Z.J., Wang, H., 2020. An end-to-end multi-task deep learning framework for skin lesion analysis. IEEE J Biomed Health Inform 24 (10), 2912–2921.

Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F., 2021. Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 71 (3), 209–249.

Thomas, S.M., Lefevre, J.G., Baxter, G., Hamilton, N.A., 2021. Interpretable deep learning systems for multi-class segmentation and classification of non-melanoma skin cancer. Med Image Anal 68, 101915.

Wang, Y., Huang, W., Sun, F., Xu, T., Rong, Y., Huang, J., 2020. Deep multimodal fusion by channel exchanging. Adv Neural Inf Process Syst 33, 4835–4845.

Xu, X., Wang, C., Guo, J., Gan, Y., Wang, J., Bai, H., Zhang, L., Li, W., Yi, Z., 2020. Mscs-deepln: evaluating lung nodule malignancy using multi-scale cost-sensitive neural networks. Med Image Anal 65, 101772.

Yap, J., Yolland, W., Tschandl, P., 2018. Multimodal skin lesion classification using deep learning. Exp. Dermatol. 27 (11), 1261–1267.

Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.-A., 2016. Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Trans Med Imaging 36 (4), 994–1004.

Yu, Z., Jiang, X., Zhou, F., Qin, J., Ni, D., Chen, S., Lei, B., Wang, T., 2018. Melanoma recognition in dermoscopy images via aggregated deep convolutional features. IEEE Trans. Biomed. Eng. 66 (4), 1006–1016.

Zhang, J., Xie, Y., Xia, Y., Shen, C., 2019. Attention residual learning for skin lesion classification. IEEE Trans Med Imaging 38 (9), 2092–2103.

Zhou, H., Schaefer, G., Sadka, A.H., Celebi, M.E., 2009. Anisotropic mean shift based fuzzy c-means segmentation of dermoscopy images. IEEE J Sel Top Signal Process 3 (1), 26–34.