

# Deep Learning-based Multi-Modal Fusion Method for Skin Lesion Classification

Peng Tang

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung eines

**Doktors der Naturwissenschaften (Dr. rer. nat.)**

genehmigten Dissertation.

**Vorsitz:**

apl. Prof. Dr. Georg Groh

**Prüfende der Dissertation:**

1. apl. Prof. Dr. Tobias Lasser
2. apl. Prof. Dr. Alexander Zink

Die Dissertation wurde am 17.04.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 16.09.2024 angenommen.





# Abstract

Skin disease is one of the rapidly spreading diseases globally. Among them, skin cancer, such as melanoma, is a hazardous form of cancer. Early detection of skin cancer can significantly increase the 5-year survival rate of patients. However, accurate diagnosis remains challenging and relies heavily on appropriate training and experience, which takes considerable time to acquire. Additionally, even for experienced dermatologists, diagnosis can be influenced by factors such as stress, fatigue, or other human elements, making consistent high-accuracy diagnosis elusive. Therefore, a computer-aided diagnosis system for skin cancer is anticipated to alleviate the burden on dermatologists and enhance overall diagnostic efficiency.

Over the past decade, with the limitations of deep learning methods being overcome by the availability of computing capacity and large-scale datasets, deep learning has become predominant in various visual and textual tasks. Consequently, the research community has increasingly explored applying deep learning methods for automatic skin disease recognition to assist dermatologists. In clinical practice, clinicians make diagnoses by considering multiple complementary modalities, including clinical images, dermoscopy images, and metadata such as age, sex, and location of the lesion, among others. However, most current methods are based solely on single-modality images. Therefore, we aim to fill this gap by investigating the use of multi-modal deep learning methods in skin lesion classification.

This dissertation investigates various multi-modal scenarios, ranging from two modalities, such as [Clinical Images \(CI\)](#)-[Dermoscopy Images \(DI\)](#) or images-metadata, to three modalities encompassing [Clinical Images \(CI\)](#), [Dermoscopy Images \(DI\)](#) and metadata. More specifically, we present a joint-individual fusion structure with a fusion attention module to effectively fuse images and metadata for skin disease recognition. To efficiently conduct the modality interaction between [Clinical Images \(CI\)](#) and [Dermoscopy Images \(DI\)](#), we present two methods: a prior-inspired asymmetric fusion method and a single-shared network method. Finally, to fully use the information on all three modalities, we present a multi-stage method that progressively hierarchically integrates the three modalities' data.





# Zusammenfassung

Hautkrankheiten gehören zu den weltweit schnell verbreitenden Krankheiten. Unter ihnen ist Hautkrebs, wie das Melanom, eine gefährliche Krebsart. Eine frühzeitige Erkennung von Hautkrebs kann die 5-Jahres-Überlebensrate der Patienten erheblich steigern. Die genaue Diagnose bleibt jedoch eine Herausforderung und hängt stark von angemessener Ausbildung und Erfahrung ab, die viel Zeit erfordern. Darüber hinaus kann die Diagnose selbst bei erfahrenen Dermatologen durch Faktoren wie Stress, Müdigkeit oder andere menschliche Einflüsse beeinträchtigt werden, wodurch eine durchgehend präzise Diagnose schwer zu erreichen ist. Daher wird ein computergestütztes Diagnosesystem für Hautkrebs erwartet, das die Belastung für Dermatologen verringern und die allgemeine diagnostische Effizienz steigern soll.

In den letzten zehn Jahren, da die Einschränkungen von Deep-Learning-Methoden durch die Verfügbarkeit von Rechenkapazität und groß angelegten Datensätzen überwunden wurden, hat sich Deep Learning in verschiedenen visuellen und textuellen Aufgaben durchgesetzt. Infolgedessen hat die Forschungsgemeinschaft zunehmend die Anwendung von Deep-Learning-Methoden zur automatischen Erkennung von Hautkrankheiten untersucht, um Dermatologen zu unterstützen. In der klinischen Praxis stellen Ärzte Diagnosen, indem sie mehrere sich ergänzende Modalitäten berücksichtigen, darunter klinische Bilder, Dermatoskopiebilder und Metadaten wie Alter, Geschlecht und Ort der Läsion, unter anderem. Die meisten aktuellen Methoden basieren jedoch ausschließlich auf einmodalen Bildern. Daher beabsichtigen wir, diese Lücke zu schließen, indem wir den Einsatz multimodaler Deep-Learning-Methoden zur Klassifizierung von Hautläsionen untersuchen.

Diese Dissertation untersucht verschiedene multimodale Szenarien, die von zwei Modalitäten wie klinischen Bildern (CI) und Dermatoskopiebildern (DI) oder Bildern und Metadaten bis hin zu drei Modalitäten reichen, einschließlich klinischer Bilder, Dermatoskopiebilder und Metadaten. Insbesondere präsentieren wir eine Joint-Individual-Fusionsstruktur mit einem Fusionsaufmerksamkeitsmodul, um Bilder und Metadaten effektiv für die Erkennung von Hautkrankheiten zu fusionieren. Um die Interaktion zwischen klinischen Bildern und Dermatoskopiebildern effizient durch-

---

zuführen, stellen wir zwei Methoden vor: eine asymmetrische, auf Vorwissen basierende Fusionsmethode und eine einheitliche Netzwerk-Methode. Schließlich präsentieren wir eine mehrstufige Methode, um alle Informationen der drei Modalitäten vollständig zu nutzen, indem die Daten der drei Modalitäten schrittweise und hierarchisch integriert werden.



# Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor, Tobias Lasser. He has always supported my decisions and provided invaluable advice and guidance whenever I faced challenges during my PhD journey. It has been an honor to work with him. I am grateful for his sharing of personal experiences in research and for providing a flexible research environment. I enjoyed a relaxed yet productive academic life at the CIIP group at the Technical University of Munich.

Secondly, I want to thank my girlfriend, YAN Xintong. Studying abroad sometimes made me feel lonely, but she consistently kept in touch and encouraged me, enabling me to persevere in completing my doctoral studies. Before the advent of ChatGPT, she helped to revise the sentence logic and correct the grammar errors in my papers.

Thirdly, I want to express my gratitude to my friends in Munich, Gen Li, Wangyang Song, Shengming Zhang, Zhihao Zhang, and Hanqin Bao, for their support in my personal life. Without you all, my life in Munich would have been much less enjoyable. Additionally, I would like to thank my colleagues at CIIP, Theo, Alessandro, Josue, Jonas, and David, for answering my numerous questions and making my time in the office more efficient. I am also thankful to Sebastian Krammer at LMU Klinikum for introducing the knowledge of dermatology, Yan Nan at Imperial College London, and Xiaobin Hu at Youtu Lab, Tencent China, for their guidance regarding my future and for their assistance in academic research.

Finally, I am deeply grateful for the support from the China Scholarship Council and my family, I would not have had the chance and confidence to study abroad and gain an excellent doctoral experience at TUM.







# Contents

<b>Abstract</b> . . . . .	<b>i</b>
<b>Zusammenfassung</b> . . . . .	<b>iii</b>
<b>Acknowledgements</b> . . . . .	<b>v</b>
<b>Contents</b> . . . . .	<b>vii</b>
<b>List of Figures</b> . . . . .	<b>xi</b>
<b>1 Introduction</b> . . . . .	<b>1</b>
1.1 Outline . . . . .	2
<b>2 Background</b> . . . . .	<b>5</b>
2.1 Feedforward Neural Network . . . . .	5
2.2 Convolutional Neural Networks . . . . .	7
2.2.1 Convolutional layer . . . . .	8
2.2.2 Pooling layer . . . . .	9
2.2.3 Batch normalization . . . . .	9
2.2.4 Skip connection . . . . .	9
2.3 Transformer . . . . .	11
2.3.1 Self attention . . . . .	11
2.3.2 Vision Transformer . . . . .	11
2.4 Loss function . . . . .	12
2.5 Data Augmentation . . . . .	12
2.5.1 Training . . . . .	12
2.5.2 Test Time Augmentation . . . . .	13
<b>3 Background: Deep learning for skin lesion classification</b> . . . . .	<b>15</b>

3.1	Single image based skin lesion classification methods . . . . .	15
3.1.1	Dataset . . . . .	15
3.1.2	Related works . . . . .	17
<b>4</b>	<b>Fusing single-image modality and patient’s metadata for skin lesion classification . . . . .</b>	<b>19</b>
4.1	Introduction . . . . .	19
4.2	Method . . . . .	21
4.2.1	Notation . . . . .	21
4.2.2	Joint-Individual Fusion (JIF) structure . . . . .	23
4.2.3	Multi-Modal Fusion Attention (MMFA) module . . . . .	24
4.3	Experiments . . . . .	27
4.3.1	Datasets . . . . .	27
4.3.2	Implementation Details . . . . .	28
4.3.3	Experiments results . . . . .	29
4.4	Discussion . . . . .	39
4.4.1	Effectiveness of using patient’s metadata . . . . .	39
4.4.2	Performance comparison between our JIF-MMFA method and other fusion methods . . . . .	39
4.4.3	Effectiveness of the Multi-Modal Fusion Attention (MMFA) Module . . . . .	41
4.5	Conclusion . . . . .	41
<b>5</b>	<b>Fusing clinical and dermoscopy images for skin lesion classification <b>43</b></b>	
5.1	Introduction . . . . .	43
5.2	Related works of multi modal-based skin lesion classification . . . . .	44
5.2.1	Seven-point checklist features . . . . .	45
5.3	Pay Less On Clinical Images: Asymmetric Multi-Modal Fusion Method For Efficient Multi-Label Skin Lesion Classification . . . . .	46
5.3.1	Related works about asymmetric fusion structure . . . . .	48
5.3.2	Asymmetrical Multi-Modal Fusion Method . . . . .	49
5.3.3	Loss Function and Final Prediction . . . . .	51
5.3.4	Experiments and Discussion . . . . .	51
5.3.5	Conclusion . . . . .	59
5.4	Single Shared Network with Prior-Inspired Loss for Efficient Multi-Modal Skin Lesion Classification . . . . .	61
5.4.1	Related works about parameter-sharing network . . . . .	63
5.4.2	Method: Parameter-Efficient Multi-Modal (PEMM) framework . . . . .	63

---

5.4.3	Experiments and Discussion . . . . .	67
5.4.4	Conclusion . . . . .	73
<b>6</b>	<b>Fusing clinical images, dermoscopy images and patient’s metadata for skin lesion classification . . . . .</b>	<b>75</b>
6.1	Introduction . . . . .	75
6.1.1	Our Works: FusionM4Net . . . . .	76
6.2	Methods . . . . .	79
6.2.1	Overall Structure of FusionM4Net . . . . .	79
6.2.2	FusionNet . . . . .	79
6.2.3	Fusion Scheme 1 . . . . .	81
6.2.4	Fusion Scheme 2 . . . . .	82
6.3	Experiments and Results . . . . .	83
6.3.1	Dataset . . . . .	83
6.3.2	Evaluated Metrics . . . . .	84
6.3.3	Implementation Details . . . . .	84
6.4	Experiments and Results . . . . .	85
6.4.1	Evaluation of FusionM4Net . . . . .	87
6.4.2	Comparison between our algorithm and other state-of-the-art methods . . . . .	92
6.5	Discussion . . . . .	93
6.5.1	Evaluation of FusionM4Net . . . . .	93
6.5.2	Comparison between our algorithm and other state-of-the-art methods . . . . .	96
6.5.3	Potential future work . . . . .	97
6.6	Conclusion . . . . .	98
<b>7</b>	<b>Concluding Remarks . . . . .</b>	<b>99</b>
7.1	Outlook . . . . .	99
7.1.1	Large Scale Dataset . . . . .	100
7.1.2	Extracting information from clinical report . . . . .	100
7.1.3	Large Language Models . . . . .	100
	<b>Appendices . . . . .</b>	<b>103</b>
<b>A</b>	<b>List of Publications . . . . .</b>	<b>103</b>
A.1	Journal Publications . . . . .	103
A.2	Conference Publications . . . . .	103

CONTENTS

---

A.3 Submitted Journal Publications . . . . . 104

**Bibliography . . . . . 105**



# List of Figures

2.1	A typical architecture of Multi-Layer Preceptron (MLP). It consists of an input layer, hidden layers, and an output layer. These layers are composed of a number of connected computational units called neurons. . . . .	6
2.2	A sample convolutional neural network for classification. It consists of convolutional layers, pooling layers, and densely connected layers. . . . .	7
2.3	A simple convolution operation without bias and whose stride is 1 on the 2D data (can be seen on the gray scale image). . . . .	8
2.4	Two kinds of pooling layers, i.e., AveragePooling and Maxpooling. . . . .	10
2.5	Comparison between the basic convolutional blocks (A) and residual convolution blocks . . . . .	10
2.6	A simple of vision transformer structure. . . . .	11
4.1	Overview of joint fusion structure (a) and joint-individual fusion structure (b), see also sections 2.1 and 2.2. In this figure, the dermatological image branch is marked in blue, the patient metadata branch is marked in green, and the fusion branch is marked in yellow. The corresponding forward and gradient flows of these three branches are also marked in the corresponding color. $M_I$ is the model to extract image features; $M_M$ is the method to extract patient metadata features; $f_I$ , $f_M$ , and $f_{IM}$ are the extracted image features, the extracted metadata features, and the features integrated by $f_I$ and $f_M$ , respectively. $C_I$ , $C_M$ , and $C_{IM}$ are the corresponding classifiers of $f_I$ , $f_M$ , and $f_{IM}$ , respectively. $P_I$ , $P_M$ , and $P_{IM}$ are the predictions obtained from $C_I$ , $C_M$ , and $C_{IM}$ , respectively. $L_I$ , $L_M$ , and $L_{IM}$ are the corresponding loss functions for $C_I$ , $C_M$ , and $C_{IM}$ , respectively. In this workflow, the inputs are the dermatological image and the patient metadata, and the outputs are the predictions $P_I$ , $P_M$ , and $P_{IM}$ . . . . .	22

4.2 Overview of (a) Metablock and MetaNet, in which the metadata ( $f_M$ ) is used to enhance the image features ( $f_I$ ), and (b) our proposed multi-modal fusion attention module, in which both image features ( $f_I$ ) and metadata features ( $f_M$ ) are enhanced by the features of other modality data and its own features. TF indicates the transformation operation, a single-layer neural network, in our experiment. AP refers to the attention operations, such as element-wise multiplication and summation, and self-attention. C is the concatenation operation.  $f_{IM}$  is the enhanced feature representation after the fusion module fuses  $f_I$  and  $f_M$ . . . . . 24

4.3 The structure of the Multi-Modal Fusion Attention (MMFA) module. The MMFA module learns how to enhance both image features ( $f_I$ ) and metadata features ( $f_M$ ) based on their own features and other modality features simultaneously. The length of output feature  $f_{IM}$  is the sum of the length of  $f_I$  and  $f_M$  in our MMFA module.  $qkv$  is a single-layer neural network,  $f_{IM}$  is the enhanced feature.  $f_{Mk}, f_{Mq}, f_{Mv}, f_{Ik}, f_{Iq}, f_{Iv}, f_K, f_Q, f_V, f_K^h, f_Q^h, f_V^h$  are the intermediate feature vectors in the attention mechanism, the details about them can be seen in the literature [5, 32] . 25

4.4 The confusion matrix of the methods in Table 4.1 considering DenseNet-121 on the PAD-UFES-20 dataset. BCC: Basal Cell Carcinoma; ACK: Actinic Keratosis; NEV: Nevus; SEK: Seborrheic Keratosis; MEL: Melanoma; SCC: Squamous Cell Carcinoma. See also sections 3.3.2 and 4.2. . . . . 32

4.5 The T-SNE figures of the methods in Table 4.1 considering DenseNet-121 on the PAD-UFES-20 dataset. Here, 0-BCC: Basal Cell Carcinoma, 1-ACK: Actinic Keratosis, 2-NEV: Nevus, 3-SEK: Seborrheic Keratosis, 4-SCC: Squamous Cell Carcinoma and 5-MEL: Melanoma. See also sections 3.3.2 and 4.2. . . . . 33

5.1 The comparison between former MSLA methods and our method. The height of each rectangle denotes its relative computational size. CE, DE, and MI are short for clinical embedder, dermoscopy embedder, and modality interaction. . . . . 46

5.2 The overview of our Asymmetric Multi-Modal Fusion Method. Clinical and dermoscopy blocks are used to extract the features from clinical and dermoscopy images, respectively.  $P_C, P_D,$  and  $P_{FU}$  are the predictions from the clinical branch (green), dermoscopy branch (blur), and fusion branch (yellow). . . . . 48

5.3 The computational graph of the former bidirectional attention block (BAB) and our asymmetrical attention block (AAB). . . . . 49

---

5.4	The overview structure of former methods and our PEMM framework. . . . .	62
5.5	The detailed pipeline of our PEMM framework. . . . .	64
5.6	The detailed pipeline of shared cross-attention module. . . . .	65
6.1	Example of one patient’s case, including the clinical image, dermoscopy image, meta-data, seven-points checklist criteria label, and diagnostic label.	75
6.2	The overview of different multi-modal CNN architectures for skin disease recognition. (a) one-stage multi-modal CNN, (b) our proposed Fusion-M4Net. In this figure, $F_1$ is used to fuse information from two-modality images at the decision level in the first stage. $F_2$ is adopted to integrate non-image modality data and image modality data in the second stage. (Abbreviations: Clin image: clinical image; Derm image: dermoscopy image). . . . .	77
6.3	The flowchart of our proposed FusionM4Net algorithm. . . . .	78
6.4	The description of the second stage of Fusion-M4Net. . . . .	83
6.5	The detailed metrics of <i>Diag</i> labels of the results in Table. 6.3. . . . .	87
6.6	The detailed metrics of <i>Diag</i> labels of the results in Table. 6.4. . . . .	89
6.7	The detailed metrics of <i>Diag</i> labels in Table. 6.10. . . . .	91





# 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 of the body [58, 63]. Skin disease is a prevalent and frequently occurring condition that typically causes symptoms such as skin itching, pain, numbness, and other discomfort, significantly impacting the lives of patients, such as vitiligo that can significantly affect people's appearance [121, 37]. In more severe cases, it can even be life-threatening. Patients with ulcerated melanoma exceeding 4 mm thickness experience a 15% five-year survival rate, whereas those with melanoma thinner than 1mm exhibit a 95% 5-year survival rate [7, 117]. In Europe, more than 100,000 new cases of melanoma and 22,000 melanoma-related deaths are reported each year [16, 20]. In the US, there are over 910,000 new cases and 9,000 deaths reported annually. The US alone spends about 3 billion dollars annually on melanoma treatment [20]. Other types of skin cancer, such as keratinocyte cancer and basal cell carcinomas, are more common compared to melanoma. While these diseases often do not result in fatalities, the medical care costs for patients with these conditions are very high [48, 129]. Especially for basal cell carcinomas, costs can rise significantly if they are diagnosed and treated in an advanced stage [129, 78]. Early detection of skin diseases can substantially improve the five-year survival rate and decrease the associated costs [9]. However, experienced dermatologists are often in short supply, especially in rural areas [35, 69]. The diagnostic accuracy of non-specialists, including primary care physicians, nurse practitioners, and physician assistants, is suboptimal, ranging from 24% to 70% [69, 34, 79]. Given that **Deep Learning (DL)**-based methods consistently demonstrate superiority in lots of visual tasks, an **Deep Learning (DL)**-based system is highly anticipated for decision support in dermatological diagnosis.

There have been many works [33, 116, 21] proving that **Artificial Intelligence (AI)** or **Deep Learning (DL)**-based methods can assist dermatologist's diagnosis and has the potential to alleviate the burden. [33] proved a **Convolutional Neural Network (CNN)** trained on 129,450 **Clinical Images (CI)** can achieve dermatologist-

level classification of skin cancer. [116] demonstrated that physicians with [Artificial Intelligence \(AI\)](#)-based decision support perform better than either AI or physicians alone in skin lesion classification. [21] illustrated that the dermatologist’s confidence in their diagnosis could significantly increase by explainable [Artificial Intelligence \(AI\)](#). However, most current methods are solely based on single-modality images and need more explorations of multi-modal data for decision support systems.

For [Deep Learning \(DL\)](#)-based methods, using multiple sources typically results in better performance than using a single input. On the other hand, in clinical practice, dermatologists typically examine patients in person over one or multiple visits rather than relying solely on one imaging modality. Consequently, physicians can integrate [Dermoscopy Images \(DI\)](#), [Clinical Images \(CI\)](#), and meta-data when analyzing each lesion. [Dermoscopy Images \(DI\)](#) are obtained by dermoscopy, a popular non-invasive imaging technique that enlarges and illuminates the skin image to improve the clarity of skin spots. By eliminating surface reflections, the visual effect of deeper skin layers can be enhanced, thereby providing more detailed information about skin lesions [13, 117]. [Clinical Images \(CI\)](#) are taken with a standard digital camera or smartphone to capture a macro view of the lesion, exhibiting more variations in terms of view, angle, and lighting [39]. Meta-data indicates the personal information of the patients, providing information beyond visual features, including age, gender, location of the lesion, history of skin cancer, parent’s history of skin cancer, and others. Also, this multi-source feature availability holds true for the majority of teledermatology evaluations [14].

Therefore, this dissertation primarily focuses on leveraging [Deep Learning \(DL\)](#) methods to integrate multi-modal data for a more precise and robust decision support system in dermatology. While previous methods based on single-modality aimed to enhance diagnoses and clinical practice, the main incentive for adopting multi-modal decision support systems in dermatology is the worldwide shortage of dermatologists.

## 1.1 Outline

This dissertation focuses on applying [Deep Learning \(DL\)](#)-based methods in different multi-modal scenarios for skin disease diagnosis. Chapter 2 introduces the concept and knowledge of [Deep Learning \(DL\)](#), and Chapter 3 introduces the application of [Deep Learning \(DL\)](#) in skin disease recognition.

Chapter 4 is a new method to effectively fuse single-image and metadata for more accurate skin lesion classification. Chapters 5 proposes two multi-modal parameter-efficient methods to combine two image modalities for multi-label skin lesion classification from different perspectives. Chapter 6 proposes a multi-stage method to

integrate [Clinical Images \(CI\)](#), [Dermoscopy Images \(DI\)](#), and metadata progressively, improving the accuracy of multi-label skin lesion classification.



# Background

Deep Learning (DL) is the primary technique in all of our proposed multi-modal fusion methods. So, in this chapter, I just briefly introduce the key concepts of Deep Learning (DL) and notations used in my dissertation, including the components of Multi-Layer Preceptron (MLP), Convolutional Neural Network (CNN), and Transformer (TS), and the backbones. More systematic review of Deep Learning (DL)'s knowledge, please refer to [40, 66, 137].

## 2.1 Feedforward Neural Network

The simplest artificial neural network is the Feedforward Neural Network (FNN), which is inspired by the neuron connections in the human brain. An Multi-Layer Preceptron (MLP), also known as fully or dense connected layers, is a type of Feedforward Neural Network (FNN). As illustrated in Fig. 2.1, an Multi-Layer Preceptron (MLP) consists of numerous computational units, or neurons, and includes various layers: an input layer, hidden layers, and an output layer. The operation within each unit can be defined as follows:

$$\hat{f}(\mathbf{x}) = h(\mathbf{w}^T \mathbf{x} + b) \quad (2.1)$$

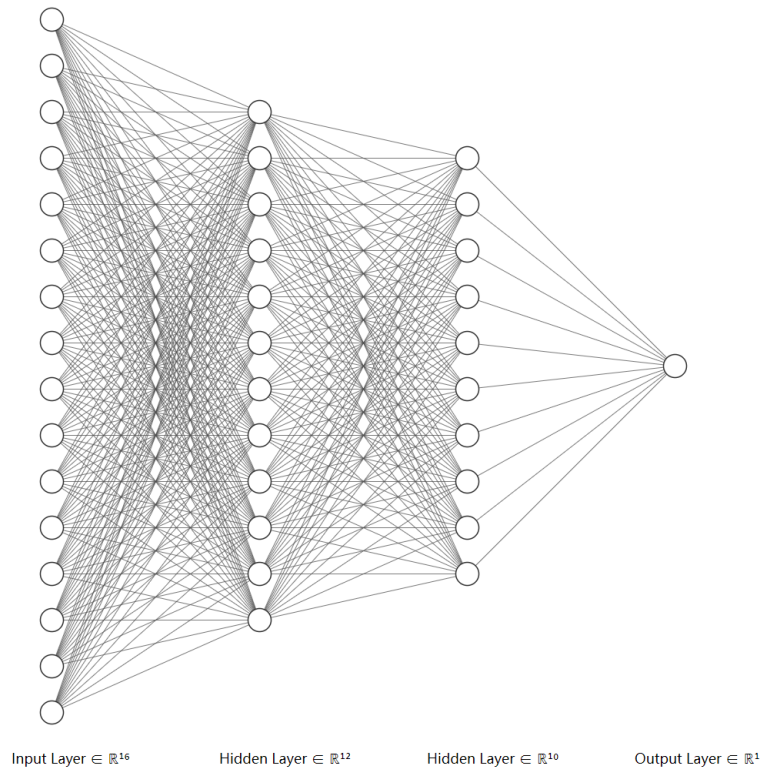
where  $\mathbf{x}$  indicates the input vectors,  $\mathbf{w} = (w_1, \dots, w_n)$  is the weight vector and  $b$  is the bias.  $h(\cdot)$  is the activation function that provides the non-linear mapping of the vectors. Mostly used activation functions in visual tasks include:

(1) Sigmoid

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (2.2)$$

(2) Softmax

$$\sigma(x) = \frac{e^{x_i}}{\sum_{j=1}^J e^{x_j}} \quad (2.3)$$



**Figure 2.1:** A typical architecture of [Multi-Layer Perceptron \(MLP\)](#). It consists of an input layer, hidden layers, and an output layer. These layers are composed of a number of connected computational units called neurons.

(3) Rectified Linear Unit (ReLU)

$$\sigma(x) = \max(0, x) \tag{2.4}$$

(4) Leaky ReLU

$$\sigma(x) = \max(0.1x, x) \tag{2.5}$$

(5) Tanh

$$\sigma(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{2.6}$$

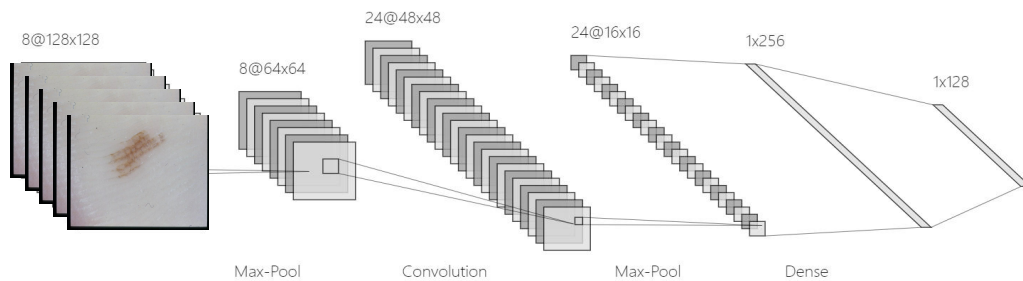
To enhance the ability of non-linear representation of [Multi-Layer Perceptron \(MLP\)](#), we can embed more hidden layers into the neural network, so it can be formulated as follows:

$$\begin{aligned} \hat{f}(\mathbf{x}; \Theta) &= (f_n \circ \dots \circ f_1)(\mathbf{x}) \\ &= h^n (h^{n-1} (\dots (h^2 (h^1(\mathbf{w}_1^T \mathbf{x} + b_1) + b_2) + b_{n-1}) + b_n) \end{aligned} \quad (2.7)$$

where  $\Theta = \{\mathbf{w}_1, \dots, \mathbf{w}_n, b_1, \dots, b_n\}$  is the model's parameters, which can be optimized by back-propagation gradient descent [93]. The gradient is generated by minimizing the loss function between the output (or called prediction) of the model and ground truth. The model's parameters  $\Theta$  are optimized as follows:

$$\Theta^{(i+1)} = \Theta^{(i)} - \eta \nabla E(\Theta^{(i)}). \quad (2.8)$$

where  $\eta$  indicates the learning rate and  $i$  is the iteration index.



**Figure 2.2:** A sample convolutional neural network for classification. It consists of convolutional layers, pooling layers, and densely connected layers.

## 2.2 Convolutional Neural Networks

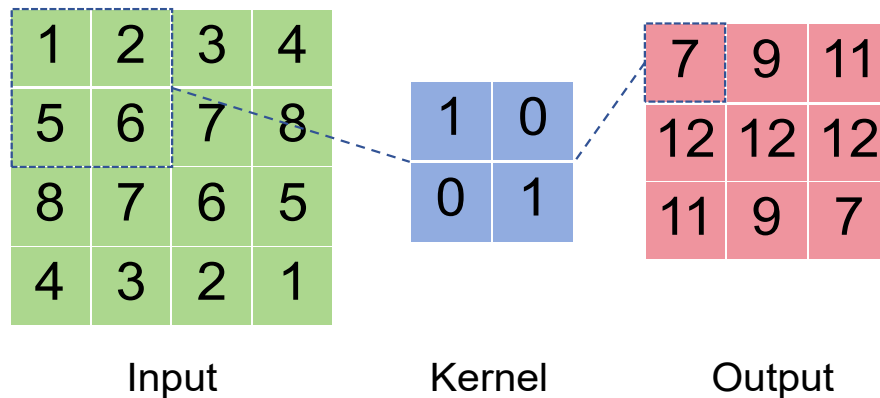
**Convolutional Neural Network (CNN)** is designed for extracting a representative feature from the imaging modality, inspired by the human visual system and the field of neuroscience. The deep structure of most **Convolutional Neural Network (CNN)** is based on the hierarchical processing of visual information in human perception, starting from low-level visual features like edges and textures to high-level semantic features like shapes and objects. The mechanisms of the local receptive field, weight sharing, and sub-sampling contribute to a powerful representation ability of **Convolutional Neural Network (CNN)** [98].

Generally, **Convolutional Neural Network (CNN)** for image classification tasks, such as AlexNet [64] and VGGNet [101] mainly contains three types of layers, i.e.,

convolutional layers, pooling layers, dense/fully connected layers (See Fig. 2.2). After ResNet [45], the skip connection structure becomes one of the basic components of following **Convolutional Neural Network (CNN)** works, including DenseNet [52], Xception [24], Mobilenet [95], EfficientNet [108] and ConNext [72].

### 2.2.1 Convolutional layer

Since the **Multi-Layer Perceptron (MLP)** is designed for processing one-dimensional (1D) data, it struggles to extract representative features from image data. Inspired by sliding windows in image processing techniques, LeCun introduced a convolutional layer to more efficiently capture image features [65]. Fig. 2.3 demonstrates how the kernel of convolutional layers operates on two-dimensional (2D) data. Specifically, a kernel with randomly initialized parameters is used to sweep over the 2D data step by step to produce an output. Specifically, the kernel values and the values of the swept area undergo element-wise multiplication to yield the corresponding output value. For instance, as shown in the area of blue bounding in Fig. 2.3, the top-left corner of the input data is swept by the kernel, resulting in the output value being computed as  $7 = 1 \times 1 + 0 \times 2 + 0 \times 5 + 1 \times 6$ . The convolutional layer can be trained to extract specific image characteristics, such as lines, textures, edges, and objects, through a weight-sharing scheme and a local receptive field.



**Figure 2.3:** A simple convolution operation without bias and whose stride is 1 on the 2D data (can be seen on the gray scale image).

Generally, a convolutional layer contains multiple kernels, so we can assume there is a group of  $N^l$  kernels in the  $l^{th}$  layer. The output of  $l^{th}$  layer  $Y^{(l)}$  will contain  $N^l$



feature maps, where the  $i^{\text{th}}$  feature map  $Y_i^{(l)}$  can be formulated mathematically as:

$$Y_i^{(l)} = h \left( \sum_{j=1}^{N^{(l-1)}} K_{i,j}^{(l)} * Y_j^{(l-1)} + B_i^{(l)} \right) \quad (2.9)$$

There are more advanced convolutional layers [28, 135, 130] available, such as the one proposed by Yu et al. [130], which increases the dilation rate of convolution to enlarge the receptive field of the model, and the deformable convolution introduced by Dai et al. [28], which incorporates an offset to extract deformation information from objects. However, all these advancements are built upon the basic convolutional layer.

### 2.2.2 Pooling layer

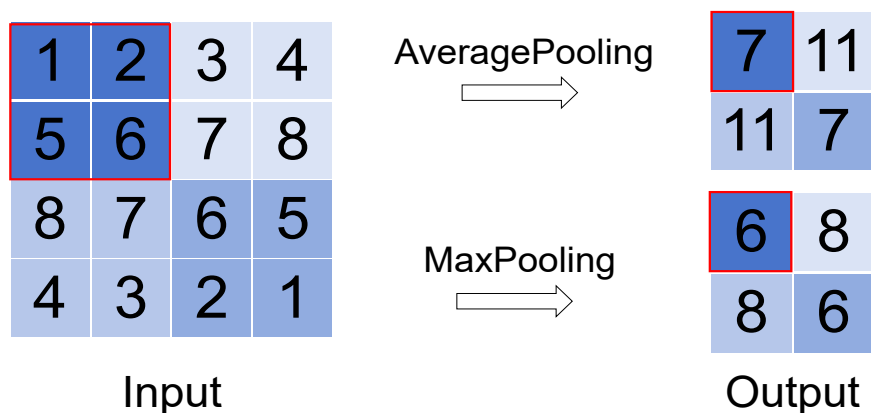
The pooling layer is common in [Convolutional Neural Network \(CNN\)](#), aiming to downscale the dimensionality of feature vectors and thus reduce the computational cost. Also, the downscale operations make the feature vector represent more advanced semantic information by increasing the receptive field. The pooling layer is similar to the convolution layer applied over a part of the image via a sliding window. Two types of pooling layers, i.e., AveragePooling and MaxPooling, are commonly used in [Convolutional Neural Network \(CNN\)](#). As shown in Fig 2.4, for the area in the red bounding box, the Averagepooling returns  $(1 + 2 + 5 + 6)/4 = 7$  and MaxPooling returns  $\max([1, 2, 5, 6]) = 6$

### 2.2.3 Batch normalization

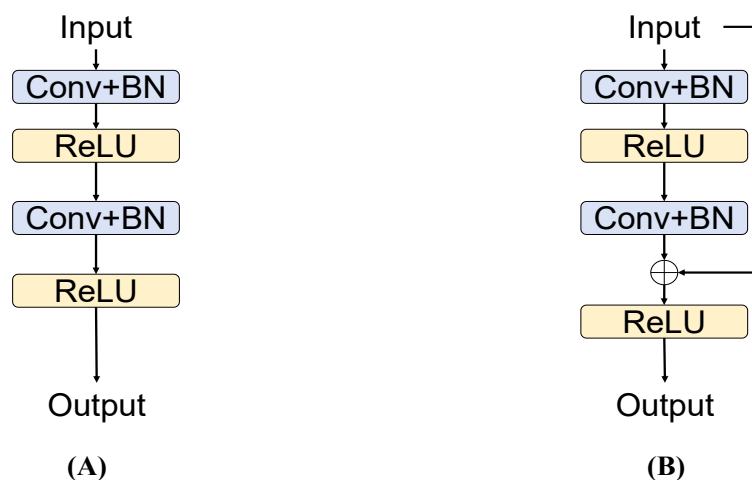
Batch normalization was proposed by Ioffe and Szegedy [55] to stabilize and accelerate the training process to ensure the distribution of input data in each network layer remains relatively stable. Specifically, the output is normalized to have zero mean and unit variance after each batch to counteract the distribution shift of data passing through a layer, allowing the weights to be updated more effectively over time.

### 2.2.4 Skip connection

Currently, the most widely-used structure for classification tasks is ResNet [45], according to Google Citations, as the proposed skip connection allows the network to be deeper. More specifically, the skip connection was introduced by He et al. [45] to address the problem of information flow to deep layers, thereby enabling the entire network to become deeper and possess stronger representational ability. As illustrated in Fig. 2.5, compared to the basic convolutional blocks in AlexNet [64] and VGGNet



**Figure 2.4:** Two kinds of pooling layers, i.e., Average Pooling and Maxpooling.



**Figure 2.5:** Comparison between the basic convolutional blocks (A) and residual convolution blocks

[101] (see Fig. 2.5 (A)), the residual convolutional blocks [45] include an additional skip connection that maps the input directly to the output (see Fig. 2.5 (B)). Skip connections are also popular in encoder-decoder segmentation structures to combine low-level spatial information with high-level semantic information for prediction, as seen in architectures like UNet [91].

## 2.3 Transformer

In recent years, Transformer (TS)-based [5] methods have dominated tasks of Natural Language Processing (NLP), exemplified by the GPT-series products [89, 90, 17, 1].

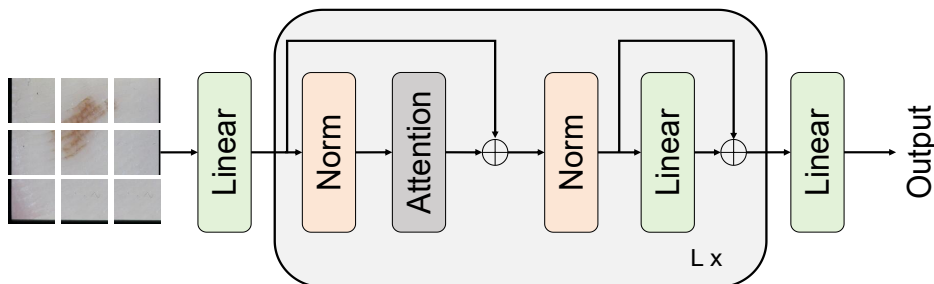
### 2.3.1 Self attention

The core component of the Transformer is the self-attention mechanism, which can be defined as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2.10)$$

where  $Q$ ,  $K$ , and  $V$  indicate the query, key, and value vectors, respectively, and  $d_k$  is the scaling factor.  $K^T$  indicate the transposed key vectors. *softmax* is usually-used activation function in attention mechanism. Actually,  $Q$ ,  $K$ , and  $V$  are different representations mapping from the same input. Why does self-attention do this? The intuitive reason is that the softmax function produces a feature distribution between 0 and 1 of the input, which can be considered as "attention or importance," indicating the importance of each word (key) relative to a word (query). This importance is then multiplied with the corresponding word (value).<sup>1</sup>

### 2.3.2 Vision Transformer



**Figure 2.6:** A simple of vision transformer structure.

Since the significant success of transformer-based methods in the NLP field, researchers have also begun to explore their potential in computer vision tasks. [Vision](#)

<sup>1</sup>An illustrated, detailed explanation of the transformer can be found at <https://jalammar.github.io/illustrated-transformer/>

**Transformer (ViT)** [32] was the first work to apply a transformer-based structure to process image data. As illustrated in Fig. 2.6, the input images are split into multiple patches and fed into a linear layer,  $L$  x transformer blocks, and another linear layer in sequence to get the final output.

## 2.4 Loss function

The loss function is the objective function to supervise the network in the training stage. By minimizing the loss error between the prediction from the network and ground truth, the network's parameters will be optimized. For different tasks and datasets, there are also different loss functions. For the multi-classes classification task, the cross-entropy loss is mainly used, which can be defined as follows:

$$L_{CE} = -\frac{1}{B} \sum_{i=1}^B \sum_{c=1}^C \delta(y_i = c) \log(P(y_i = c)) \quad (2.11)$$

where  $B$  represents the batch size and  $C$  represents the number of classes.  $\delta(y_i = c)$  is the indicator function and  $P(y_i = c)$  is the prediction from network. For the classification task on the unbalanced dataset, Focal loss [68] was introduced, and for the image segmentation task, DCE loss was the most frequently used; other segmentation loss can be found in [57] and [61].

## 2.5 Data Augmentation

### 2.5.1 Training

During the training stage of **Convolutional Neural Network (CNN)**, data augmentation is usually used to generate more data to enhance the model's generalization ability and thus increase the accuracy. However, not all the data augmentation techniques are useful for skin lesion classification tasks. In our experiments, only geometric data augmentations, including horizontal and vertical flipping, rotation, shift, scaling, and random brightness and contrast are implemented. Thanks for the Albumentations library that provides a convenient data augmentation pipeline [18]. More details about data augmentation can be seen in [99, 99]

### 2.5.2 Test Time Augmentation

During the testing stage, data augmentation can be used to create multiple transformed versions of the input image, and thus, a more accurate and robust prediction can be obtained by averaging the multiple predictions [97]. In our experiments, we just applied the commonly used version of Test Time augmentation, i.e., geometric transformation, including rotation and flipping.



# Background: Deep learning for skin lesion classification

In this chapter, we primarily introduce the background of single-image-based [Skin Lesion Classification \(SLC\)](#) methods, including datasets and related works, as well as the skin lesion classification tasks, encompassing both the diagnosis task and the seven-point checklist feature classification task. Given that there are different datasets and related works for various multi-modal scenarios, we will discuss them in the corresponding chapters accordingly.

## 3.1 Single image based skin lesion classification methods

### 3.1.1 Dataset

[Dermoscopy Images \(DI\)](#)-based skin lesion classification methods are typically optimized using a supervised learning scheme, where the prerequisite for training a [Convolutional Neural Network \(CNN\)](#) is the dataset. Here, we introduce the commonly-used public datasets for single image-based SLC, including PH2 <sup>1</sup> [76], [International Skin Imaging Collaboration \(ISIC\)](#) challenges <sup>2</sup> [42, 42, 26, 25, 115, 27, 92], SD-198, and SD-260 [104, 127]. More dataset information about skin lesions can be seen in <sup>3</sup>

---

<sup>1</sup><https://www.fc.up.pt/addi/ph2%20database.html>

<sup>2</sup><https://challenge.isic-archive.com/data/>

<sup>3</sup><https://github.com/sfu-mial/awesome-skin-image-analysis-datasets>

## CHAPTER 3. BACKGROUND: DEEP LEARNING FOR SKIN LESION CLASSIFICATION

---

### PH2

To the best of our knowledge, PH2 [76] is the first publicly available dermoscopic image database for melanoma detection. PH2 dataset was collected from the Universidade do Porto, Técnico Lisboa, and the Dermatology Service of Hospital Pedro Hispano in Matosinhos, Portugal. The images are with the size of  $768 \times 500$  and obtained through 20 times magnification of Tuebinger Mole Analyzer system. It was released in early 2013 and contains only 200 images (120 cases of nevus and 80 cases of melanoma) in total. Despite its small size, it is a high-quality dataset that includes manual segmentation labels, clinical diagnosis labels, and the identification of several dermoscopic features. As a result, many current studies still use it to evaluate various skin analysis tasks, including classification [102], segmentation [31, 10], and explainable AI [12].

### ISIC 2016

The ISIC 2016 Challenge comprises three tasks: lesion segmentation, detection and localization of visual dermoscopic features/patterns, and lesion classification. For the SLC task dataset, a total of 900 images are provided for training, with an additional 374 cases reserved for performance evaluation. All the images are categorized into two classes: melanoma and benign. These images are acquired from various devices across multiple leading clinical centers.

### ISIC 2017

The dataset of ISIC 2017 SLC Challenge [26] is the expansion of ISIC 2016 counterpart, mainly reflected in the data scale and more classes. In the 2017 challenge, participants were required to classify the lesion into three classes: Melanoma, Seborrheic Keratosis, and Benign Nevus. The dataset is split into three parts: training (2000 images), validation (150 images), and testing sets (500 images).

### ISIC 2018

The ISIC 2018 SLC dataset expands on the ISIC 2017 dataset by incorporating the HAM10000 dataset [115, 25]. The ISIC 2018 dataset consists of 10,015 training images, 193 validation images, and 1,500 test images. All images in this dataset are categorized into seven classes, i.e., Melanoma, Melanocytic nevus, Basal cell carcinoma, Actinic keratosis / Bowen's disease (intraepithelial carcinoma), Benign keratosis (solar lentigo / seborrheic keratosis/lichen planus-like keratosis). Dermatofibroma and Vascular lesion.



### 3.1. SINGLE IMAGE BASED SKIN LESION CLASSIFICATION METHODS

---

More advanced versions of ISIC datasets, such as ISIC 2019 and ISIC 2020 datasets, also exist, but the primary difference between them and the ISIC 2018 dataset lies in the number of data. Therefore, we will not delve into the details of these datasets here. For more information about these datasets, please refer to [27, 92].

#### SD-198

All of the above datasets are the dermoscopic images dataset, while the SD-198 [104] is a clinical images dataset (obtained by standard camera or smartphone) for skin disease recognition. The dataset is collected from DermQuest<sup>4</sup> and then annotated by two professional experts. The dataset contains 6,584 images from 198 different categories, with each category having 60 images at most to maintain the balance of categories. SD-260 is the large and unbalanced version of SD-260; more details can be seen in [127].

#### 3.1.2 Related works

Most current SLC methods are based solely on single-image modalities, i.e., *Dermoscopy Images (DI)* or *Clinical Images (CI)*. Therefore, to understand the application of deep learning in skin lesion classification, it is essential to be familiar with single image-based methods.

However, SLC using clinical images is rare. To facilitate research in this area, [104] released a large clinical skin diseases dataset, SD-198, as a benchmark for comparing convolutional neural networks (CNNs) and hand-crafted features. [126] 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 [127]. DL-based methods have significantly improved clinical-image (CI) based skin lesion classification compared to hand-crafted methods. However, there still exists a considerable gap between CI-based and dermoscopy image-based methods [113, 29]. For instance, in a comparative study by [29], dermoscopy image (DI)-based CNN models significantly increased the accuracy of skin cancer diagnosis from 75% to 88% compared to smartphone images.

More researches [8, 94, 74, 33, 41, 111, 128, 73, 37, 59, 77, 131, 134, 109, 132] is directed towards dermoscopy images (DI) rather than clinical images (CI) due to two main factors. Firstly, as mentioned earlier, dermoscopy images offer higher

---

<sup>4</sup><https://www.dermquest.com/>

diagnostic accuracy compared to clinical images [29]. Secondly, the availability and high quality of numerous dermoscopy image datasets in challenges organized by the [International Skin Imaging Collaboration \(ISIC\)](#) also plays a significant role. Automatic SLC methods, generally based on supervised learning and dermoscopy images, can be divided into traditional methods and Convolutional Neural Network (CNN)-based methods. Traditional methods [8, 94, 74] typically consist of two main steps: 1. Denoising and feature extraction involves using a color constancy algorithm to eliminate illumination noise and extracting a scale-invariant feature transform (SIFT) descriptor. 2. Building strong classifiers like AdaBoost, random forest, and SVM. However, these methods heavily rely on various feature engineering algorithms and could be more robust against dermoscopy images obtained from different imaging devices and lighting conditions.

Because of the flexibility of CNN structures and its huge success in the ImageNet challenge [100, 105], CNN-based methods and their variants have been applied in many medical image tasks, including skin image analysis. Recently published CNN-based SLC methods [59, 77, 131, 134, 109, 132] mainly focus on transfer-learning with fine-tuned techniques, segmentation-classification models, self-attention modules, and models based on the combination of deep features and handcrafted features. [59] and [77] fine-tuned a CNN model with the pre-trained weight on ImageNet, and these two methods outperformed the traditional methods by a large margin. [131] used a fully convolutional network (FCN) to segment the region of interest (ROI) from dermoscopy images and then directly classified the cropped images based on the ROI. Several publications [134] also reported that the segmentation-classification methods achieve a higher accuracy than other single-classification methods. [134] and [109] replaced the segmentation model with a self-attention module to prioritize the skin lesion area. This attention module enhances the classification performance without extra segmentation labels. [132] adopted the Fisher vector and a CNN model for melanoma classification. Some other methods [75, 43] use deep ensemble learning advantageously in the skin lesion classification challenge that was organized by the International Skin Imaging Collaboration (ISIC). Additionally, [41] 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. [128] 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. [37] explored and integrated information from different views, including RGB, HSL, and YCbCr, rather than only the RGB view, thereby enhancing skin lesion classification.

# Fusing single-image modality and patient’s metadata for skin lesion classification

## 4.1 Introduction

Multi-modal information fusion encompasses the integration of data from diverse sources, aiming to augment machine learning algorithms by capturing complementary and comprehensive information beyond what single-modality data can offer [6, 54]. In recent years, multi-modal deep learning models have demonstrated considerable success across various domains beyond medical image analysis [54]. For instance, [114] devised a multi-modal pipeline merging visual and textual features for social media video classification. This approach elevated the classification accuracy from 76.4% using a single modality CNN to 88%, showcasing the efficacy of multi-modal fusion techniques. Similarly, [87] engineered a detection system that amalgamates image data with Light Detection and Ranging (LiDAR) sensor data for autonomous driving. This fusion system achieved a 3.7% higher accuracy compared to models trained solely on single modality data. The success stories of multi-modal information fusion in non-medical domains have piqued the interest of researchers in the medical field. The adoption of multi-modal fusion schemes holds the promise of providing complementary insights and overcoming the limitations of single-modality models. Notably, recent literature reviews [54] suggest a growing trend towards integrating image data with electronic health records to tackle challenges that single-modality models struggle to robustly address, particularly in medical image analysis, including dermatological image analysis.

Patient demographics represent critical clinical information during dermatological examinations, especially in scenarios where visual characteristics of skin lesions

exhibit inter-class similarity and intra-class variation. Clinical metadata, including age, gender, lesion location, parental background, and skin cancer history, among others, emerge as pivotal factors in dermatological diagnosis [82].

Numerous studies have explored the fusion of dermoscopy images with patient metadata [129, 60, 69, 82, 67]. To our knowledge, [129] pioneered the integration of deep learning models for combining two-modal dermatological images and patient metadata for skin lesion classification. Subsequent works such as [60] introduced multi-modal learning networks that fuse image data and patient metadata for multi-label skin lesion classification. Likewise, [69] developed a deep learning system that combined multi-view images and metadata to differentiate skin diseases, achieving performance comparable to dermatologists while outperforming primary care physicians and nurse practitioners in validation. Despite their noteworthy outcomes, these methodologies relied on simple feature concatenation to integrate data from two modalities, potentially overlooking latent relationships between dermatological images and metadata [84, 67]. Recent research has suggested that concatenation alone may not fully exploit multi-modal data. Thus, approaches such as Metablock, Metanet, and Mutual Attention [82, 69, 19] have emerged to extract relevant image features through attention-based mechanisms, surpassing the performance achieved by concatenation operations. However, these approaches mentioned above generally used the joint fusion structure to fuse images and patient metadata. This means that these methods only learn a joint feature representation of multi-modality data and neglect to retain the specific characteristics of each modality that has been verified to be crucial for the multi-modal task [47, 51]. Also, most of the current fusion modules (fusion operations) only used metadata to enhance the most relevant image features and did not explore the possibility of using both image and metadata to enhance the most related features of these two-modality data. Therefore, in our opinion, there still exists great potential to get more accurate results by designing an improved fusion approach regarding the overall structure and a multi-modal attention module.

In this study, we propose a novel joint-individual fusion (JIF) framework complemented by a multi-modal fusion attention (MMFA) module to seamlessly integrate dermatological images and patient metadata. Initially, the JIF structure concurrently learns an optimized shared multi-modal feature representation while preserving modality-specific features. This approach aims to enhance the overall representation capacity of the data by capturing both shared and distinctive characteristics. Subsequently, the MMFA attention module is devised to accentuate the most pertinent image and metadata features. It achieves this by highlighting the most relevant features of each modality, leveraging insights from both the modality itself and the complementary information provided by the other modality. Our proposed method

underwent rigorous evaluation on three public datasets: namely, the PAD-UFES-20 dataset [83], Seven-Point Check (SPC) dataset [60], and ISIC-2019 dataset [115, 26, 27]. We conducted a comparative analysis with other state-of-the-art fusion techniques, including the Joint Fusion (JF) structure with Concatenation, Metanet, Metablocks, and Mutual Attention [84, 67, 82, 19]. Our experimental findings underscore the effectiveness of the JIF-MMFA method in consistently enhancing the performance of various CNN architectures. Across the different datasets, our method generally outperformed alternative fusion techniques, highlighting its robustness and superiority in skin lesion classification tasks.

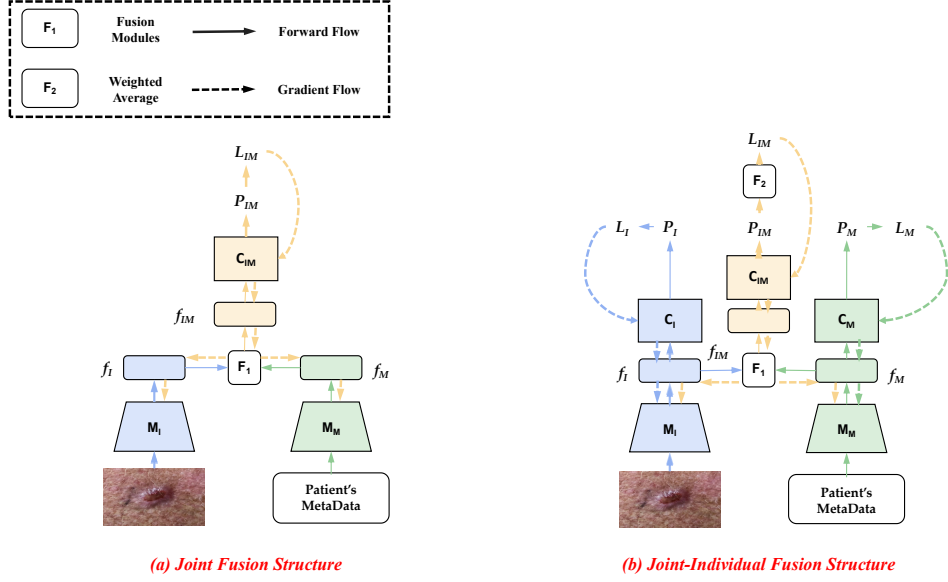
The method contribution can be summarized as follows:

1. Compared to previous methods that only focus on developing a new fusion module, we add an idea to improve the performance by exploring more efficient fusion structures.
2. A new joint-individual fusion (JIF) structure that learns modal-shared and modal-specific features simultaneously, which can consistently improve the classification performance of different fusion modules, different backbones, and different datasets.
3. A new multi-modal fusion attention (MMFA) module that enhances the most relevant image features and metadata features, where the most relevant features of a modality will be highlighted by the features from the other modality and itself.
4. Taking advantage of the JIF structure and MMFA module, we contribute a JIF-MMFA method, which achieves state-of-the-art performance on multiple skin disease datasets.

## 4.2 Method

### 4.2.1 Notation

For the purpose of skin lesion classification, we frame the fusion of dermatological images and patient metadata as a multi-class classification task. Each case comprises an image *Image*, a set of patient metadata *Meta*, and a ground truth label  $GT \in \{1, 2, 3, \dots, N\}$ , where  $N$  represents the number of labels. To process the raw image data *Image*, we employ a Convolutional Neural Network (CNN) denoted as  $M_I$  to extract image features  $f_I \in \mathbb{R}^{D_I}$ , which represent the last-layer feature maps of the



**Figure 4.1:** Overview of joint fusion structure (a) and joint-individual fusion structure (b), see also sections 2.1 and 2.2. In this figure, the dermatological image branch is marked in blue, the patient metadata branch is marked in green, and the fusion branch is marked in yellow. The corresponding forward and gradient flows of these three branches are also marked in the corresponding color.  $M_I$  is the model to extract image features;  $M_M$  is the method to extract patient metadata features;  $f_I$ ,  $f_M$ , and  $f_{IM}$  are the extracted image features, the extracted metadata features, and the features integrated by  $f_I$  and  $f_M$ , respectively.  $C_I$ ,  $C_M$ , and  $C_{IM}$  are the corresponding classifiers of  $f_I$ ,  $f_M$ , and  $f_{IM}$ , respectively.  $P_I$ ,  $P_M$ , and  $P_{IM}$  are the predictions obtained from  $C_I$ ,  $C_M$ , and  $C_{IM}$ , respectively.  $L_I$ ,  $L_M$ , and  $L_{IM}$  are the corresponding loss functions for  $C_I$ ,  $C_M$ , and  $C_{IM}$ , respectively. In this workflow, the inputs are the dermatological image and the patient metadata, and the outputs are the predictions  $P_I$ ,  $P_M$ , and  $P_{IM}$ .

CNN. Here,  $D_I$  signifies the dimensionality of the image features  $f_I$ . For the patient metadata  $Meta$ , we utilize one-hot encoding along with multiple Fully Connected Layers (FCLs) as  $M_M$  to convert the raw data into nonlinear metadata features  $f_M \in \mathbb{R}^{D_M}$ , where  $D_M$  denotes the dimensionality of the metadata features  $f_M$ . These two feature extraction processes can be formally expressed as follows:

$$f_I = M_I(Image) \quad (4.1)$$

$$f_M = M_M(Meta) \quad (4.2)$$

Therefore, our objective is to introduce a method denoted as ME, which aims to predict the probability  $P$  of the ground truth label  $GT$  belonging to a class  $c \in \{1, 2, 3, \dots, N\}$ , given the image  $Image$  and the metadata  $Meta$ :

$$P_{GT} = \text{ME}( GT = c \mid f_I, f_M ) \quad (4.3)$$

### 4.2.2 Joint-Individual Fusion (JIF) structure

To describe the former methods based on the Joint Fusion structure (see Fig. 4.1), Eq. (4.3) is modified as follows:

$$P_{IM} := C_{IM}( GT = c \mid F_1( f_I, f_M ) ) \quad (4.4)$$

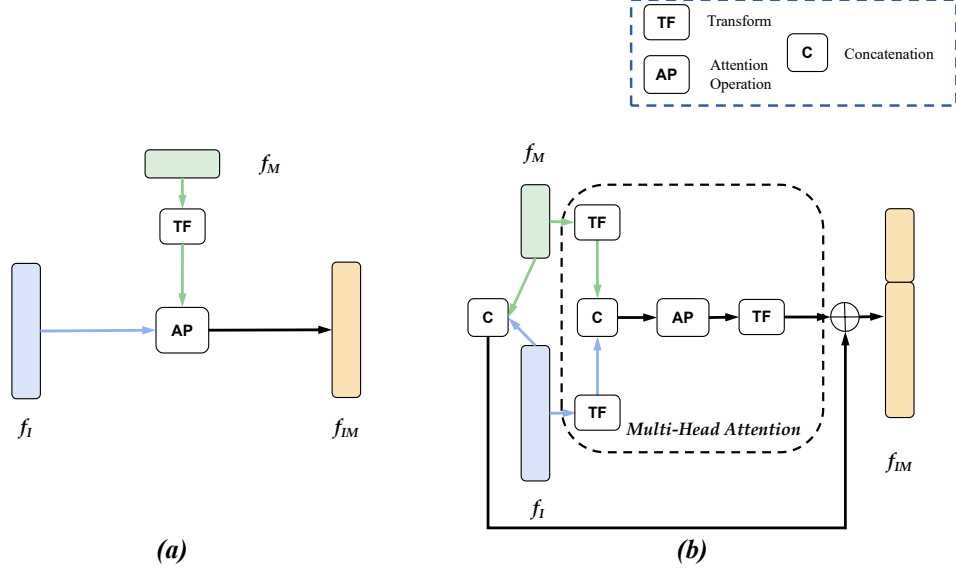
where  $C_{IM}$  represents the classifier corresponding to the fused features  $F_{IM}$ , where  $F_{IM} = F_1(f_I, f_M)$ , and  $F_1$  denotes the fusion module.  $P_{IM}$  denotes the prediction obtained from  $C_{IM}$ . For the proposed Joint-Individual Fusion structure, Eq. (4.3) is further derived as follows:

$$\begin{aligned} P_{IM} &:= C_{IM}( GT = c \mid F_1( f_I, f_M ) ) \\ P_I &:= C_I( GT = c \mid f_I ) \\ P_M &:= C_M( GT = c \mid f_M ) \\ P_{GT} &:= F_2( P_{IM}, P_I, P_M ) \end{aligned} \quad (4.5)$$

where  $C_I$  and  $C_M$  are the corresponding classifiers of the image features  $f_I$  and metadata features  $f_M$ .  $P_I$  and  $P_M$  are the predictions of  $C_I$  and  $C_M$ .  $C_{IM}$  is a fully connected layer that is commonly used as a classifier by CNNs to predict the last feature maps. From Eq. (5.4) and Eq. (4.5), it can be seen that the main differences between the proposed JIF structure and the JF structure are in  $F_2$ ,  $P_I$ , and  $P_M$ . These differences can be further differentiated by two aspects: training and testing.

During the training phase, we employ an intuitive approach where two loss functions,  $L_I$ , and  $L_M$ , are incorporated to individually supervise the image branch ( $C_I$  and  $FM_I$ ) and the metadata branch ( $C_M$  and  $FM_M$ ). This modification alters the entire gradient flow, facilitating the model to acquire a joint feature representation while preserving the distinctive features of each modality. As depicted in Fig. 4.1(b), the gradients from  $L_I$  (blue) and  $L_M$  (green) direct the image and metadata branches to maintain their specific representations,  $f_I$  and  $f_M$ , respectively. The loss function  $L_{IM}$  optimizes the entire structure, thereby attaining the joint feature representation  $F_{IM}$ .

During the testing phase, given the presence of three classifiers in the JIF structure, we naturally amalgamate these three predictions at the decision level to enhance accuracy.



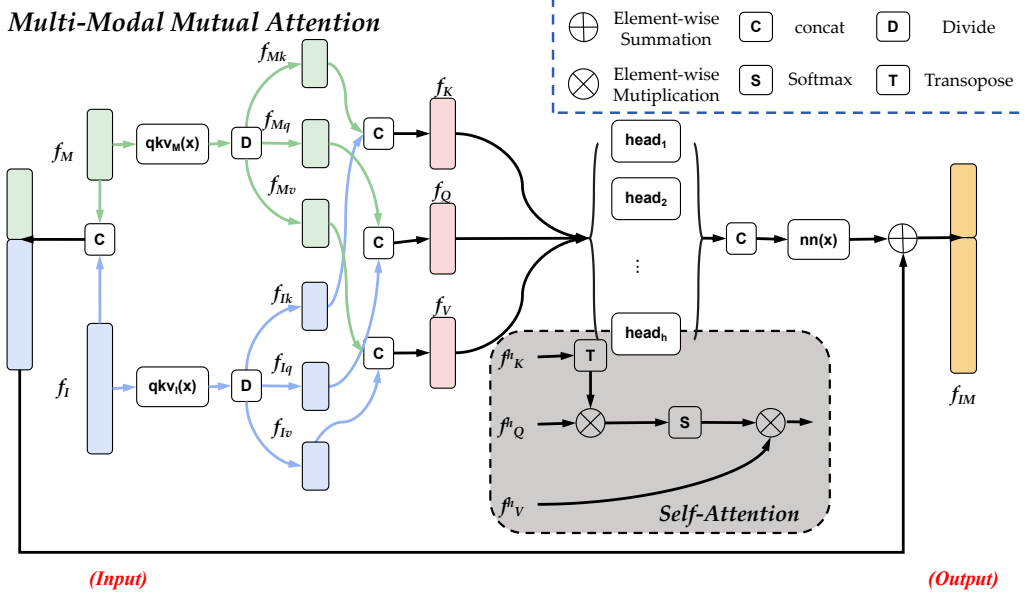
**Figure 4.2:** Overview of (a) Metablock and MetaNet, in which the metadata ( $f_M$ ) is used to enhance the image features ( $f_I$ ), and (b) our proposed multi-modal fusion attention module, in which both image features ( $f_I$ ) and metadata features ( $f_M$ ) are enhanced by the features of other modality data and its own features. TF indicates the transformation operation, a single-layer neural network, in our experiment. AP refers to the attention operations, such as element-wise multiplication and summation, and self-attention. C is the concatenation operation.  $f_{IM}$  is the enhanced feature representation after the fusion module fuses  $f_I$  and  $f_M$ .

### 4.2.3 Multi-Modal Fusion Attention (MMFA) module

The proposed fusion module aims to enrich both patient metadata features and image features by leveraging information from both modalities. For instance, image features can be enhanced not only by metadata features but also by the image features themselves simultaneously. This approach is grounded in the perspective of data-driven learning, where integrating more information into the attention/fusion operation enables the selection of more task-related features, thereby improving the performance of skin lesion classification.

The MMFA module integrates both mutual attention and self-attention mechanisms to enhance the two-modality data and obtain improved fusion feature vectors. Subsequently, it incorporates a skip connection to include  $f_I$ ,  $f_M$ , and the enhanced feature vectors, thereby constructing the final fused feature vector  $f_{IM}$ . This skip





**Figure 4.3:** The structure of the Multi-Modal Fusion Attention (MMFA) module. The MMFA module learns how to enhance both image features ( $f_I$ ) and metadata features ( $f_M$ ) based on their own features and other modality features simultaneously. The length of output feature  $f_{IM}$  is the sum of the length of  $f_I$  and  $f_M$  in our MMFA module.  $qkv$  is a single-layer neural network,  $f_{IM}$  is the enhanced feature.  $f_{Mk}, f_{Mq}, f_{Mv}, f_{Ik}, f_{Iq}, f_{Iv}, f_K, f_Q, f_V, f_K^h, f_Q^h, f_V^h$  are the intermediate feature vectors in the attention mechanism, the details about them can be seen in the literature [5, 32]

connection addresses the vanishing gradient problem [45] and harnesses valuable information from the original feature vectors.

A structural overview of the proposed MMFA module is illustrated in Fig. 4.3, and its key characteristics can be summarized as follows:

$$F_{IM} := F_1(f_i, f_M) = \text{MMFA}(f_i, f_M) = \text{MHA}(f_k, f_Q, f_V) \oplus \text{Concat}(f_I, f_M) \quad (4.6)$$

where  $\text{Concat}$  presents the concatenation operation, which is used to link  $f_I$  and  $f_M$ , and  $\oplus$  represents the element-wise summation operation.

We follow the paper of [5] and build a multi-head attention block MHA to implement the self-attention mechanism in the MMFA fusion module, as the effectiveness of this attention block in processing different modality data (such as sequence data [5] and vision data [32]) has been shown. .

$$\text{MHA}(f_K, f_Q, f_V) = f(\text{Concat}(\text{head}_1(f_K, f_Q, f_V), \dots, \text{head}_h(f_K, f_Q, f_V)))$$

$$head_i(f_K, f_Q, f_V) = \frac{Softmax((f_K^i)^T \otimes (f_Q^i))}{\sqrt{s}} \otimes v_I((f_V^i)) \quad (4.7)$$

where  $\sqrt{s}$  represents the scaling factor,  $\otimes$  denotes the element-wise multiplication operation, and  $(f_K^i)^T$  signifies the transpose of  $f_K^i$ . Here, the vectors  $f_K^i \in \mathbb{R}^{d_k}$ ,  $f_Q^i \in \mathbb{R}^{d_q}$ , and  $f_V^i \in \mathbb{R}^{d_v}$  correspond to the *key*, *query*, and *value* vectors in  $head_i$ , respectively. Additionally,  $d_k = d_q = d_i = s = D_T/h$ , where  $d_k$ ,  $d_q$ , and  $d_v$  denote the dimensions of the key, query, and value vectors, respectively, and  $h$  represents the number of heads.

$$\begin{aligned} f_K &= Concat(f_{Mk}, f_{Ik}) \\ f_Q &= Concat(f_{Mq}, f_{Iq}) \\ f_V &= Concat(f_{Mv}, f_{Iv}) \end{aligned} \quad (4.8)$$

Where  $k$ ,  $q$ , and  $v$  represent a type of single-layer neural network, serving as an intuitive means to conduct non-linear transformations on feature maps within deep learning architectures. The transformations  $k_M$ ,  $q_M$ , and  $v_M$  are employed to convert the metadata features to possess identical structures (i.e., the same input and output feature numbers) but with distinct parameters. Conversely,  $k_I$ ,  $q_I$ , and  $v_I$  share the same structure and are utilized to transform the image features.

$$\begin{aligned} f_{Iq}, f_{Ik}, f_{Iv} &:= D(qkv_I(f_I)) = D(BN(f_I \otimes W_I + b_I)) \\ f_{Mq}, f_{Mk}, f_{Mv} &:= D(qkv_M(f_M)) = D(BN(f_M \otimes W_M + b_M)) \end{aligned} \quad (4.9)$$

Where  $W_M \in \mathbb{R}^{L_M \times d_{meta}}$  and  $b_M \in \mathbb{R}^{d_{meta}}$  are the weights and biases of  $k_M$ , while  $W_I \in \mathbb{R}^{L_I \times d_{img}}$  and  $b_I \in \mathbb{R}^{d_{img}}$  are the weights and biases of  $k_I$ .  $D$  means divide operation that equally divides the output of  $qkv$  into thirds, i.e., key, value, and query features.  $BN$  indicates the batch normalization operation [55].  $f_K \in \mathbb{R}^{D_k}$ ,  $f_Q \in \mathbb{R}^{D_q}$  and  $f_V \in \mathbb{R}^{D_v}$  are the *query*, *key* and *value* feature vectors in the self-attention mechanism, where  $D_k = D_q = D_v = d_{img} + d_{meta}$ .

$nn(x)$  is a single-layer neural network like  $k$ ,  $q$ , and  $v$ , but with a different structure and parameters.  $nn(x)$  is defined as:

$$nn(x) = BN(x \otimes W^{nn} + b^{nn}) \quad (4.10)$$

where  $W^f \in \mathbb{R}^{D_T \times L_I + L_M}$  and  $b^f \in \mathbb{R}^{L_I + L_M}$  are the weights and biases.

## 4.3 Experiments

In this section, we evaluate the performance of our joint-individual fusion structure and multi-modal attention module. We conduct experiments using five CNN architectures and evaluate them on three established datasets. The section will cover the datasets used, implementation details, experimental results, and subsequent discussions. We will proceed by introducing each aspect in sequence.

### 4.3.1 Datasets

Three public skin lesion classification datasets with both dermatological images and patient metadata, PAD-UFES-20 [83], Seven-Point Checklist (SPC) [60], and ISIC-2019 [115, 26, 27], are used for the performance evaluation:

**PAD-UFES-20** dataset comprises 2298 patient cases, encompassing clinical images captured via smartphone devices and accompanied by 21 metadata entries. These metadata entries include age, gender, skin history, parent’s background, among others. This dataset is used to classify six classes of skin lesions: Seborrheic Keratosis (SEK), Melanoma (MEL), Nevus (NEV), Basal Cell Carcinoma (BCC), Squamous Cell Carcinoma (SCC), and Actinic Keratosis (ACK).

**Seven-Point Checklist (SPC)** dataset consists of 413 training cases, 203 validation cases, and 395 testing cases. Each case includes dermatological image data along with 14 metadata entries, encompassing information such as gender, location of the skin lesion, management, and features of the seven-point checklist. The SPC dataset mainly has five types of skin lesions, including MEL, NEV, SEK, BCC, and Miscellaneous (MISC).

**ISIC-2019** dataset comprises 25,331 dermoscopy images, with each image linked to three clinical features: location, gender, and age. This dataset has eight types of skin lesions: MEL, NEV, BCC, ACK, Vascular Lesion (VAL), Benign Keratosis (BK), Dermatofibroma (DF), and Squamous Cell Carcinoma (SCC).

For the PAD-UFES-20 dataset and the ISIC-2019 dataset, we adhere to the methodology outlined in the paper by [82], employing a five-fold cross-validation approach stratified by the frequency of classes to evaluate our method. Regarding the SPC dataset, we utilize the division provided by the dataset creator, which includes predefined training, validation, and testing subsets. For a more robust analysis of the SPC dataset, we train all the models five times and get the average and standard deviation values for the comparisons.

### 4.3.2 Implementation Details

In this paper, we evaluated the proposed method on the three datasets and conducted four performance comparisons on each dataset. First, to demonstrate the effectiveness of patient metadata, we compared methods using both images and metadata with those using only image data. Then, to illustrate the superiority of our method, we compared the proposed JIF-MMFA method with other current fusion methods. Finally, to highlight the effectiveness of the JIF structure and the MMFA module, we separately compared the JIF structure with the Joint Fusion (JF) structure and the MMFA module with other fusion modules.

For the comparisons, we employed five CNN backbones: Mobilenet-v2 [95], Efficientnet-B3 [107], Resnet-50 [45], Densenet-121 [52], and Xception [24] as  $M_I$  to evaluate the generalization ability of the fusion methods for those CNNs. We utilized commonly used metrics such as balanced accuracy (BAC), accuracy (ACC), and area under the curve (AUC) for performance evaluation. The BAC value was chosen as the ranking metric [82], and the main paper tables display the performance based on the BAC value. Other metrics are available in the supplementary materials.

All models were initialized with pre-trained weights from ImageNet [30] and fine-tuned on the three skin lesion classification datasets for 150 epochs. We used an SGD optimizer with an initial learning rate of 0.005 and a CosineAnnealingLR schedule in PyTorch. Training stopped early if the model’s validation BAC value did not improve for 30 consecutive epochs. Given the imbalanced nature of the dataset, we follow the paper of [82] that employed class-weighted cross-entropy as the loss function.

In our JIF structure, we utilized three-branch loss functions:  $L_I$ ,  $L_M$ , and  $L_{IM}$ . During training, the goal of the JIF structure was to minimize the total loss function  $L_{total} = \beta * L_I + (1 - \beta) * L_M + L_{IM}$ , where  $\beta$  represents the weight of each modality in the training process. We set  $\beta$  to 0.5 to consider the two-modality data equally important.

All images were resized to  $224 \times 224 \times 3$  and with a batch size of 128 before training, and common data augmentations such as horizontal and vertical flipping, shifting, rotation, and scaling were applied to expand the datasets. Test-time augmentation, including flipping and rotation, was also used to improve the performance. The Python libraries Pytorch [85], Sklearn [86], Numpy [44], and Albumentation [18], were used to build our workflow, including model design, data loader, training and testing flows.

### 4.3.3 Experiments results

To simplify the description in the following text, we use the following abbreviations: **JF**: Joint Fusion; **JIF**: Joint-Individual Fusion; **OFB**: Only fusion branch; **FS**: Fusion structure; **Cat**: Concatenation; **MB**: Metablock; **MN**: MetaNet; **MMFA**: Multi-Modal Fusion Attention. Also, we concatenate these abbreviations to name the employed fusion methods. For example, we abbreviate our Joint-Individual Fusion (JIF) structure with multi-modal fusion attention (MMFA) as **JIF-MMFA**, and the Joint Fusion structure (JF) with metablock (MB) as **JF-MB**. Additionally, **JIF-MMFA (OFB)** denotes the result only from the  $P_{IM}$  of the JIF-MMFA method; while **JIF-MMFA (All)** denotes the result by averaging the three predictions  $P_I$ ,  $P_M$ , and  $P_{IM}$  of the JIF-MMFA method (see Fig. 4.1(b)).

In Table 4.1, Table 4.2, and Table 4.3, we present the performance comparisons between our method and other existing methods. We showcase the mean value and standard deviation of the BAC metric, aiming to illustrate the effectiveness of utilizing metadata, the superiority of our proposed method, and ablation studies of our JIF-MMFA. Then, Table 4.4, Table 4.5, and Table 4.6 provide the results of the Wilcoxon test for the methods outlined in Table 4.1, Table 4.2, and Table 4.3, respectively. This analysis offers a further comparison of these methods in terms of their statistical differences. Moreover, Table 4.7, Table 4.8, and Table 4.9 individually depict the experimental outcomes of the JIF structure and the JF structure with different fusion modules. These tables analyze the effectiveness of the JIF structure and the MMFA Module.

**Table 4.1:** Performance comparisons between JIF-MMFA and other methods on the PAD-UFES-20 dataset in terms of BAC. *FS*: Fusion structure; *Cat*: Concatenation; *MB*: Metablock; *MN*: MetaNet; *MA*: Mutual Attention, *MMFA* Multi-Modal Fusion Attention. *JF*: Joint Fusion Structure, *JIF*: Joint-Individual Fusion Structure.

<i>FS</i>	<i>Image</i>	<i>JF</i>					<i>JIF</i>	
<i>CNN</i>		<i>CAT</i>	<i>MB</i>	<i>MN</i>	<i>MA</i>	<i>MMFA</i>	<i>MMFA (OFB)</i>	<i>MMFA (All)</i>
<i>densenet</i>	68.9±2.6	73.8±1.4	72.4±2.1	68.6±2.2	76.0±2.3	75.6±1.7	<b>78.0±2.0</b>	77.7±1.8
<i>mobilenet</i>	67.1±1.5	73.7±1.2	70.1±3.7	69.1±3.0	75.0±1.6	75.2±1.6	74.7±1.4	<b>75.6±0.7</b>
<i>resnet</i>	66.1±1.5	72.9±1.7	72.1±1.6	68.8±3.0	73.3±1.6	73.6±2.4	76.0±1.2	<b>76.4±1.5</b>
<i>effnet</i>	64.6±1.4	76.8±1.4	71.4±2.2	65.4±2.0	74.8±2.0	76.0±1.8	78.8±1.6	<b>79.8±1.4</b>
<i>xception</i>	68.3±1.5	73.8±1.9	70.1±1.6	66.8±1.3	73.5±2.9	74.1±3.0	75.9±1.4	<b>76.3±1.2</b>
<i>Average</i>	67.0±2.3	74.2±2.0	71.2±2.6	67.8±2.8	74.5±2.4	74.9±2.3	76.7±2.2	<b>77.2±2.0</b>

**Table 4.2:** Performance comparisons between JIF-MMFA and other methods on the SPC dataset in terms of BAC. *FS*: Fusion structure; *Cat*: Concatenation; *MB*: Metablock; *MN*: MetaNet; *MA*: Mutual Attention, *MMFA* Multi-Modal Fusion Attention. *JF*: Joint Fusion Structure, *JIF*: Joint-Individual Fusion Structure.

<i>FS</i>	<i>Image</i>	<i>JF</i>					<i>JIF</i>	
<i>CNN</i>		<i>CAT</i>	<i>MB</i>	<i>MN</i>	<i>MA</i>	<i>MMFA</i>	<i>MMFA(OFB)</i>	<i>MMFA(AU)</i>
<i>densenet</i>	54.9±2.9	61.1±2.2	67.4±0.6	57.5±1.9	69.5±3.3	72.3±2.6	70.9±2.3	<b>73.1±2.6</b>
<i>mobilenet</i>	57.4±4.8	70.3±1.2	69.3±0.9	60.2±4.0	70.4±0.9	69.4±3.7	72.1±4.9	<b>73.1±3.9</b>
<i>resnet</i>	53.7±4.2	62.8±5.1	67.8±1.8	55.0±2.2	65.7±3.0	67.5±2.6	70.0±2.7	<b>70.4±2.6</b>
<i>effnet</i>	55.0±1.4	73.2±2.3	68.2±2.3	55.1±2.4	70.0±1.9	70.8±1.2	71.2±2.0	<b>74.0±1.1</b>
<i>xception</i>	55.7±3.7	<b>72.8±2.2</b>	67.0±1.4	57.4±3.0	68.9±3.4	68.1±2.8	70.6±2.0	71.5±2.7
<i>Average</i>	55.4±3.8	68.1±5.9	68.0±1.7	57.1±3.4	68.9±3.1	69.6±3.2	71.0±3.1	<b>72.4±3.0</b>

**Table 4.3:** Performance comparisons between JIF-MMFA and other methods on the ISIC-2019 dataset in terms of BAC. *FS*: Fusion structure; *Cat*: Concatenation; *MB*: Metablock; *MN*: MetaNet; *MA*: Mutual Attention, *MMFA* Multi-Modal Fusion Attention. *JF*: Joint Fusion Structure, *JIF*: Joint-Individual Fusion Structure.

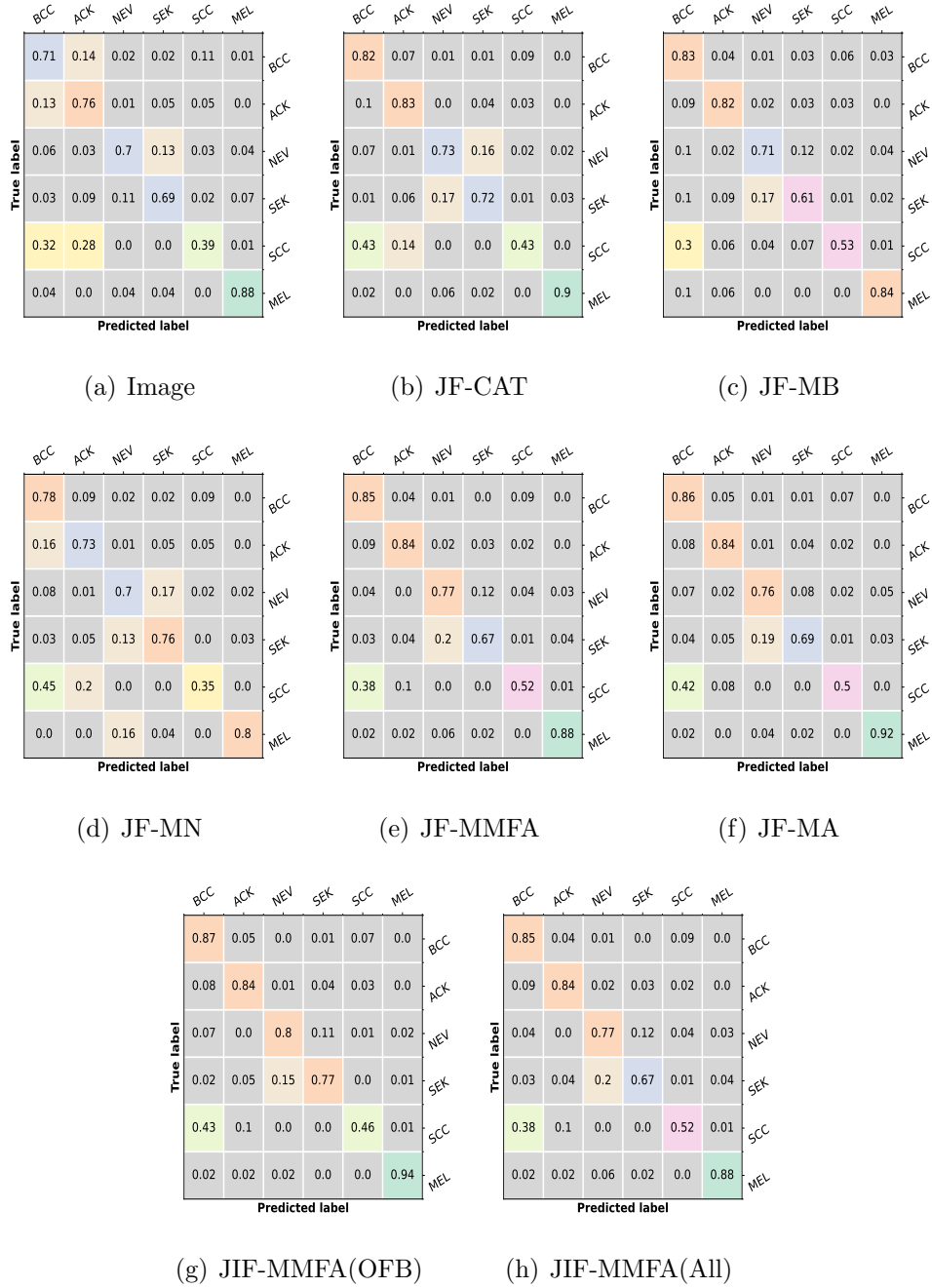
<i>FS</i>	<i>Image</i>	<i>JF</i>					<i>JIF</i>	
<i>CNN</i>		<i>CAT</i>	<i>MB</i>	<i>MN</i>	<i>MA</i>	<i>MMFA</i>	<i>MMFA(OFB)</i>	<i>MMFA(AU)</i>
<i>densenet</i>	81.8±0.5	83.3±1.0	82.9±0.5	82.9±1.6	82.4±1.1	82.4±0.7	<b>84.8±1.1</b>	84.6±0.9
<i>mobilenet</i>	80.3±1.7	83.0±0.7	82.9±1.0	83.4±0.2	81.8±0.6	81.6±1.2	<b>85.0±1.5</b>	84.8±1.4
<i>resnet</i>	81.5±0.4	82.7±1.1	83.4±0.4	83.4±0.8	65.5±9.4	68.8±5.3	83.7±0.5	<b>83.7±0.3</b>
<i>effnet</i>	79.4±0.7	80.2±0.5	79.3±1.7	79.6±0.7	81.9±1.2	80.8±1.5	<b>82.6±0.6</b>	82.5±0.7
<i>xception</i>	79.2±1.4	79.8±0.9	78.2±0.6	79.0±0.4	82.1±1.4	81.1±1.5	82.5±0.3	<b>82.7±0.3</b>
<i>Average</i>	80.4±1.5	81.8±1.7	81.3±2.3	81.7±2.1	78.7±7.9	78.9±5.8	<b>83.8±1.4</b>	83.7±1.3

**Table 4.4:** The results of the statistical test (Wilcoxon pair-wise test) for all the methods on the PAD-UFES-20 dataset. The  $P_{value} > 0.05$  is highlighted in bold.

<i>Model-Pairs</i>	<i>P_value</i>	<i>Model-Pairs</i>	<i>P_value</i>
Image - JF-CAT	5.96E-08	JF-MB - JF-MA	1.01E-05
Image - JF-MB	2.56E-06	JF-MB - JF-MMFA	1.13E-06
Image - JF-MN	<b>0.2635</b>	JF-MB - JIF-MMFA (OFB)	1.19E-07
Image - JF-MA	1.19E-07	JF-MB - JIF-MMFA (All)	1.19E-07
Image - JF-MMFA	1.19E-07	JF-MN - JF-MA	5.96E-08
Image - JIF-MMFA (OFB)	5.96E-08	JF-MN - JF-MMFA	1.79E-07
Image - JIF-MMFA (All)	5.96E-08	JF-MN - JIF-MMFA (OFB)	1.19E-07
JF-CAT - JF-MB	6.37E-05	JF-MN - JIF-MMFA (All)	5.96E-08
JF-CAT - JF-MN	1.19E-07	JF-MA - JF-MMFA	<b>0.5249</b>
JF-CAT - JF-MA	<b>0.3957</b>	JF-MA - JIF-MMFA (OFB)	0.000714958
JF-CAT - JF-MMFA	<b>0.2411</b>	JF-MA - JIF-MMFA (All)	2.21E-05
JF-CAT - JIF-MMFA (OFB)	2.21E-05	JF-MMFA - JIF-MMFA (OFB)	0.004175186
JF-CAT - JIF-MMFA (All)	1.19E-07	JF-MMFA - JIF-MMFA (All)	0.000216901
JF-MB - JF-MN	8.80E-05	JIF-MMFA (OFB) - JIF-MMFA (All)	0.001815677

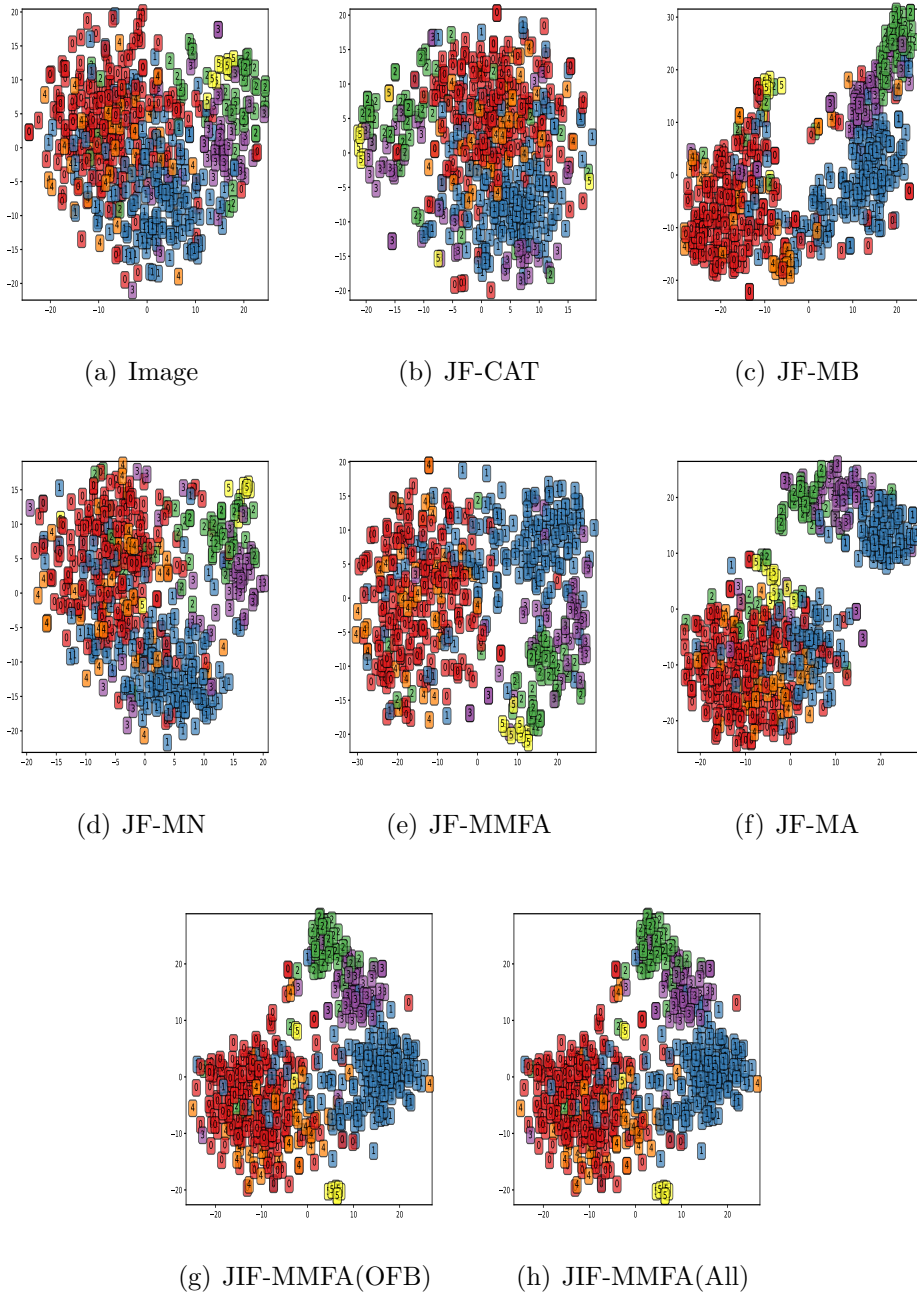
**Table 4.5:** The results of the statistical test (Wilcoxon pair-wise test) for all the methods on the SPC dataset. The  $P_{value} > 0.05$  is highlighted in bold.

<i>Model-Pairs</i>	<i>P_value</i>	<i>Model-Pairs</i>	<i>P_value</i>
Image - JF-CAT	4.17E-07	JF-MB - JF-MA	<b>0.1073</b>
Image - JF-MB	5.96E-08	JF-MB - JF-MMFA	0.0236
Image - JF-MN	<b>0.1135</b>	JF-MB - JIF-MMFA (OFB)	1.83E-05
Image - JF-MA	5.96E-08	JF-MB - JIF-MMFA (All)	1.79E-07
Image - JF-MMFA	5.96E-08	JF-MN - JF-MA	5.96E-08
Image - JIF-MMFA (OFB)	5.96E-08	JF-MN - JF-MMFA	5.96E-08
Image - JIF-MMFA (All)	5.96E-08	JF-MN - JIF-MMFA (OFB)	5.96E-08
JF-CAT - JF-MB	0.832509398	JF-MN - JIF-MMFA (All)	5.96E-08
JF-CAT - JF-MN	1.79E-07	JF-MA - JF-MMFA	<b>0.3123</b>
JF-CAT - JF-MA	<b>0.6915</b>	JF-MA - JIF-MMFA (OFB)	0.0088
JF-CAT - JF-MMFA	<b>0.4578</b>	JF-MA - JIF-MMFA (All)	1.23E-05
JF-CAT - JIF-MMFA (OFB)	0.0255	JF-MMFA - JIF-MMFA (OFB)	<b>0.1073</b>
JF-CAT - JIF-MMFA (All)	0.0022	JF-MMFA - JIF-MMFA (All)	7.50E-05
JF-MB - JF-MN	5.96E-08	JIF-MMFA (OFB) - JIF-MMFA (All)	4.54E-05



**Figure 4.4:** The confusion matrix of the methods in Table 4.1 considering DenseNet-121 on the PAD-UFES-20 dataset. BCC: Basal Cell Carcinoma; ACK: Actinic Keratosis; NEV: Nevus; SEK: Seborrheic Keratosis; MEL: Melanoma; SCC: Squamous Cell Carcinoma. See also sections 3.3.2 and 4.2.





**Figure 4.5:** The T-SNE figures of the methods in Table 4.1 considering DenseNet-121 on the PAD-UFES-20 dataset. Here, 0-BCC: Basal Cell Carcinoma, 1-ACK: Actinic Keratosis, 2-NEV: Nevus, 3-SEK: Seborrheic Keratosis, 4-SCC: Squamous Cell Carcinoma and 5-MEL: Melanoma. See also sections 3.3.2 and 4.2.

**Table 4.6:** The results of the statistical test (Wilcoxon pair-wise test) for all the methods on the ISIC-2019 dataset. The  $P_{value} > 0.05$  is highlighted in bold.

<i>Model-Pairs</i>	<i>P_value</i>	<i>Model-Pairs</i>	<i>P_value</i>
Image - JF-CAT	6.56E-06	JF-MB - JF-MA	<b>0.8119</b>
Image - JF-MB	0.0309	JF-MB - JF-MMFA	<b>0.4742</b>
Image - JF-MN	0.0025	JF-MB - JIF-MMFA (OFB)	2.56E-06
Image - JF-MA	<b>0.3254</b>	JF-MB - JIF-MMFA (All)	3.28E-06
Image - JF-MMFA	<b>0.7712</b>	JF-MN - JF-MA	<b>0.6528</b>
Image - JIF-MMFA (OFB)	5.96E-08	JF-MN - JF-MMFA	<b>0.2411</b>
Image - JIF-MMFA (All)	5.96E-08	JF-MN - JIF-MMFA (OFB)	1.23E-05
JF-CAT - JF-MB	<b>0.1730</b>	JF-MN - JIF-MMFA (All)	1.23E-05
JF-CAT - JF-MN	<b>0.4578</b>	JF-MA - JF-MMFA	<b>0.6915</b>
JF-CAT - JF-MA	<b>0.4108</b>	JF-MA - JIF-MMFA (OFB)	5.39E-05
JF-CAT - JF-MMFA	<b>0.0957</b>	JF-MA - JIF-MMFA (All)	1.83E-05
JF-CAT - JIF-MMFA (OFB)	5.25E-06	JF-MMFA - JIF-MMFA (OFB)	1.13E-06
JF-CAT - JIF-MMFA (All)	2.56E-06	JF-MMFA - JIF-MMFA (All)	5.96E-07
JF-MB - JF-MN	<b>0.1485</b>	JIF-MMFA (OFB) - JIF-MMFA (All)	<b>0.6073</b>

### Effectiveness of using patient’s metadata

The experiments in this section demonstrate the effectiveness of using patient metadata in addition to image data. As depicted in Table 4.1, Table 4.2, and Table 4.3, models that incorporate metadata achieve higher BAC values compared to those that rely solely on image data across all three datasets. Particularly, the proposed JIF-MMFA (All) method leads to a significant improvement for all five CNN backbones and multiple datasets.

When comparing models using only images to those incorporating metadata, the JIF-MMFA (All) method demonstrates substantial enhancements. Specifically, it elevates the average BAC value from  $67.0\% \pm 2.3\%$  to  $77.2\% \pm 2.0\%$  on the PAD-UFES-20 dataset,  $55.4\% \pm 3.8\%$  to  $72.4\% \pm 3.0\%$  on the SPC dataset, and  $80.4\% \pm 1.5\%$  to  $83.7\% \pm 1.3\%$  on the ISIC-2019 dataset. However, JF-MN only gets a slight increase of 0.8% on the PAD-UFES-20 dataset.

These results underscore that fusing patient metadata with images can enhance performance, with the degree of improvement varying based on the fusion methods employed. Our proposed JIF-MMFA method demonstrates the most significant improvement, highlighting its effectiveness in leveraging both image and metadata information for skin lesion classification.

**Table 4.7:** Performance comparison of different fusion structures (FS) with different fusion modules and CNN backbones on the PAD-UFES-20 dataset in terms of the BAC value. The highest are highlighted in bold for each row. *FS*: Fusion structure; *Cat*: Concatenation; *MB*: Metablock; *MN*: MetaNet; *MA*: Mutual Attention, *MMFA* Multi-Modal Fusion Attention. JIF (OFB): the result only from the  $P_{IM}$  of the JIF structure; JIF (All): the result by averaging these three predictions  $P_I$ ,  $P_M$  and  $P_{IM}$  of the JIF structure. (numbers in %)

<i>FS</i>	<i>JIF (OFB)</i>					<i>JIF (ALL)</i>				
<i>CNN</i>	<i>CAT</i>	<i>MB</i>	<i>MN</i>	<i>MA</i>	<i>MMFA</i>	<i>CAT</i>	<i>MB</i>	<i>MN</i>	<i>MA</i>	<i>MMFA</i>
<i>densenet</i>	73.4±1.4	72.7±2.1	67.3±2.8	75.7±1.3	<b>78.0±2.0</b>	74.0±1.2	74.8±1.3	69.7±2.0	76.3±1.5	77.7±1.8
<i>mobilenet</i>	74.7±2.5	71.7±3.8	68.5±1.9	75.7±0.8	74.7±1.4	75.7±2.8	73.6±2.1	71.1±1.2	<b>76.9±1.2</b>	75.6±0.7
<i>resnet</i>	73.2±1.4	71.3±2.9	69.7±1.8	75.1±1.9	76.0±1.2	73.9±1.6	73.4±0.7	71.5±2.2	75.4±1.8	<b>76.4±1.5</b>
<i>effnet</i>	76.6±1.9	69.8±2.3	65.7±0.4	75.0±1.8	78.8±1.6	77.3±1.5	72.7±1.4	71.1±2.3	77.0±1.3	<b>79.8±1.4</b>
<i>xception</i>	74.2±1.4	70.5±0.4	65.2±2.8	75.3±2.2	75.9±1.4	74.8±1.6	72.9±0.9	68.8±2.7	75.6±1.8	<b>76.3±1.2</b>
<i>Average</i>	74.4±2.2	71.2±2.7	67.3±2.7	75.4±1.7	76.7±2.2	75.1±2.2	73.5±1.5	70.4±2.4	76.3±1.7	<b>77.2±2.0</b>

## Performance comparison between our JIF-MMFA method and other fusion methods

In this part, we examine the performance enhancement achieved by the JIF structure and the MMFA module in our JIF-MMFA method through an ablation study. We compare the performance of JF-MMFA, JIF-MMFA (OFB), and JIF-MMFA (All) to assess the impact of different components. Subsequently, we compare the proposed JIF-MMFA approach with other fusion methods (JF-CAT, JF-MB, and JF-MN) across the three datasets.

Firstly, in contrast to the JF structure, our JIF structure conserves the distinctive properties of each modality to facilitate a more refined joint feature presentation, seamlessly integrating multi-modal information at the decision-making stage. To ascertain the performance enhancements attributable to these dual factors, we juxtapose the outcomes derived from the singular fusion branch (OFB)  $P_{IM}$  of the JIF structure (JIF-MMFA (OFB)), against the collective average of all three predictions of the JIF structure (JIF-MMFA (All)):  $P_{IM}$ ,  $P_I$  and  $P_M$  (see Fig. 4.1(b)), see Table 4.1, Table 4.2, and Table 4.3. Compared with JF-MMFA, JIF-MMFA (OFB) elevates the averaged BAC value from  $74.9\% \pm 2.3\%$  to  $76.7\% \pm 2.2\%$  on the PAD-UFES-20 dataset, from  $69.6\% \pm 3.2\%$  to  $71.0\% \pm 3.1\%$  on the SPC dataset, and from  $78.9\% \pm 5.9\%$  to  $83.7\% \pm 1.3\%$  on the ISIC-2019 dataset. Notably, JIF-MMFA (All) marginally

**Table 4.8:** Performance comparison of different fusion structures (FS) with different fusion modules and CNN backbones on the SPC dataset in terms of the BAC value. The highest BAC values are highlighted in bold for each row. *FS*: Fusion structure; *Cat*: Concatenation; *MB*: Metablock; *MN*: MetaNet; *MA*: Mutual Attention, *MMFA* Multi-Modal Fusion Attention. JIF (OFB): the result only from the  $P_{IM}$  of the JIF structure; JIF (All): the result by averaging these three predictions  $P_I$ ,  $P_M$  and  $P_{IM}$  of the JIF structure. (numbers in %)

<i>FS</i>	<i>JIF (OFB)</i>					<i>JIF (ALL)</i>				
	<i>CAT</i>	<i>MB</i>	<i>MN</i>	<i>MA</i>	<i>MMFA</i>	<i>CAT</i>	<i>MB</i>	<i>MN</i>	<i>MA</i>	<i>MMFA</i>
<i>densenet</i>	64.4±4.1	69.8±2.9	56.5±4.4	69.4±4.3	70.9±2.3	66.1±3.9	73.0±2.1	60.9±4.3	70.8±3.5	<b>73.1±2.6</b>
<i>mobilenet</i>	68.0±2.4	68.9±1.8	62.5±2.9	74.5±2.3	72.1±4.9	68.9±1.9	72.2±1.9	65.6±2.9	<b>75.0±2.8</b>	73.1±3.9
<i>resnet</i>	61.1±1.8	68.7±2.5	53.9±3.2	68.3±3.9	70.0±2.7	63.6±2.4	<b>71.7±3.0</b>	60.4±2.7	69.1±3.4	70.4±2.6
<i>effnet</i>	74.7±1.3	69.7±2.1	53.4±4.0	70.8±2.2	71.2±2.0	<b>75.1±1.4</b>	71.6±1.3	65.9±2.0	72.2±1.8	74.0±1.1
<i>xception</i>	70.6±1.6	67.5±1.1	58.0±1.8	72.5±1.6	70.6±2.0	71.3±1.8	68.9±0.9	66.9±1.4	<b>73.0±1.9</b>	71.5±2.7
<i>Average</i>	67.7±5.3	68.9±2.3	56.9±4.7	71.1±3.8	70.9±3.1	69.0±4.7	71.5±2.4	63.9±3.9	72.1±3.4	<b>72.4±3.0</b>

surpasses JIF-MMFA (OFB) in average BAC across these three datasets. These findings underscore that the enhancements in JF-MMFA and JIF-MMFA predominantly stem from the retention of modality-specific features, thereby facilitating a superior joint feature representation with minimal influence from the decision-level fusion of multi-modal data.

Secondly, we juxtapose JIF-MMFA against other fusion methodologies: Joint Fusion structures employing Concatenation (JF-CAT), Metablock (JF-MB), Metanet (JF-MN), and Mutual Attention (JF-MA) [67, 84, 82, 19]. Our proposed JIF-MMFA (All) method surpasses all other methods across all datasets based on the average BAC value. In comparison to prior techniques, JIF-MMFA (All) notably demonstrates significant enhancements on both the PAD-UFES-20 and SPC datasets, exhibiting an increase of 2.7% in averaged BAC value compared with the second-best method (JF-MA) on the PAD-UFES-20 dataset (see Table 4.1), and an increase of 3.5% on the SPC dataset (see Table 4.2), Also, concerning the ISIC-2019 dataset, JIF-MMFA (OFB) ( $83.7\% \pm 1.3\%$ ) achieves an increase of 1.9% in averaged BAC value compared with the second-best method (JF-CAT  $81.8\% \pm 1.3\%$ ), underscoring the advantageous nature of our approach.

The Friedman test and subsequent Wilcoxon test were conducted for statistical analysis, employing a significance level of  $p = 0.05$ . The Friedman test yielded  $p$  values of approximately  $1.99 \times 10^{-24}$ ,  $8.71 \times 10^{-23}$ , and  $1.06 \times 10^{-12}$  on the PAD-UFES-20

**Table 4.9:** Performance comparison of different fusion structures (FS) with different fusion modules and CNN backbones on the ISIC-2019 dataset in terms of the BAC value. The highest are highlighted in bold for each row. *FS*: Fusion structure; *Cat*: Concatenation; *MB*: Metablock; *MN*: MetaNet; *MA*: Mutual Attention, *MMFA* Multi-Modal Fusion Attention. JIF (OFB): the result only from the  $P_{IM}$  of the JIF structure; JIF (All): the result by averaging these three predictions  $P_I$ ,  $P_M$  and  $P_{IM}$  of the JIF structure. (numbers in %)

<i>FS</i>	<i>JIF (OFB)</i>					<i>JIF (ALL)</i>				
	<i>CAT</i>	<i>MB</i>	<i>MN</i>	<i>MA</i>	<i>MMFA</i>	<i>CAT</i>	<i>MB</i>	<i>MN</i>	<i>MA</i>	<i>MMFA</i>
<i>densenet</i>	82.4±0.6	82.8±1.1	81.3±1.7	84.3±0.3	<b>84.8±1.1</b>	82.4±0.7	82.4±1.0	81.6±1.7	84.5±0.4	84.6±0.9
<i>mobilenet</i>	81.8±0.7	81.7±1.5	82.0±0.3	85.6±0.3	85.0±1.5	81.9±0.8	81.4±1.2	82.0±0.6	<b>85.7±0.2</b>	84.8±1.4
<i>resnet</i>	82.8±0.2	82.0±0.7	81.7±1.0	<b>84.3±0.7</b>	83.7±0.5	82.8±0.4	81.5±0.7	82.0±1.1	84.1±0.5	83.7±0.3
<i>effnet</i>	79.0±1.6	79.6±1.0	78.6±1.0	82.5±0.5	<b>82.6±0.6</b>	79.0±1.7	79.9±0.7	79.1±1.0	<b>82.7±0.9</b>	82.5±0.7
<i>xception</i>	78.6±1.3	79.0±0.9	78.0±0.7	82.3±0.3	82.5±0.3	78.7±1.3	79.0±0.8	78.5±0.5	82.2±0.4	<b>82.7±0.3</b>
<i>Average</i>	80.9±2.0	81.0±1.8	80.3±2.0	<b>83.8±1.3</b>	<b>83.8±1.4</b>	80.9±2.1	80.9±1.5	80.6±1.9	<b>83.8±1.3</b>	83.7±1.3

dataset, the SPC dataset, and the ISIC-2019 dataset, respectively. Consequently, the Wilcoxon test (two-sided) was carried out, and the results are presented in Table 4.4, Table 4.5, and Table 4.6. Examination of these tables reveals that on the PAD-UFES-20 and SPC datasets, the model pairings involving our JIF-MMFA (ALL) against previous methods (JF-CAT, JF-MB, JF-MN, JF-MA) all returned values exceeding 0.05, indicating superior performance of JIF-MMFA (ALL) relative to these methods. Illustrations of the confusion matrix and T-SNE plot for various fusion methods are depicted in Fig.4.4 and Fig.4.5. Given the abundance of 15 confusion matrices and T-SNE figures, we opted to showcase the results specifically for DenseNet-121 on the PAD-UFES-20 dataset due to its lightweight nature and widespread use as a CNN backbone in deep learning, presenting a fair performance in our experiments. An observation can be made regarding the correlation between Fig.4.4 and Fig.4.5: a higher misclassification rate between two types of skin diseases in Fig.4.4 corresponds to a closer distance between the corresponding two types in Fig.4.5. For instance, with our JIF-MMFA (All) model, 38% of SCC cases are predicted as BCC in Fig.4.4 (g). Consequently, the cluster between 0 (BCC) and 4 (SCC) is difficult to distinguish in Fig.4.5. Conversely, no BCC cases were predicted as NEV and SEK in Fig.4.4 (g), resulting in a clear separation between the clusters representing 0 (BCC) and 2 (NEV), 3 (SEK) in Fig.4.5. This distinction arises from the features utilized for T-SNE analysis, which comprise the final feature vector of the fusion method, which

is directly utilized for prediction.

### **Effectiveness of Joint-Individual Fusion (JIF) structure**

To further assess the efficacy of our JIF structure, we conduct a comparative analysis between the JIF structure and the JF structure utilizing four distinct fusion modules, as outlined in Table 4.7, Table 4.8, and Table 4.9. Upon examining the results on the PAD-UFES-20 dataset (refer to Table 4.7), it is evident that the JIF (All) structure enhances the performance of all four fusion modules in terms of the average BAC value. Additionally, the JIF (OFB) structure demonstrates improvement in 3 out of 4 fusion modules, except for the Metanet, compared with the JF structure. Similar trends are observed for the SPC dataset, as depicted in Table 4.8. Both the JIF (All) and JIF (OFB) structures exhibit enhancements across all four fusion modules compared to the JF structure. On the ISIC-2019 dataset (see Table 4.9), JIF (All) and JIF (OFB) improve the performance of the MA fusion module from  $78.7\% \pm 7.9\%$  to  $83.8\% \pm 1.3\%$  and to  $83.8\% \pm 1.3\%$ , respectively, and MMFA fusion module from  $78.9\% \pm 5.5\%$  to  $83.6\% \pm 1.3\%$  and to  $83.8\% \pm 1.2\%$ , respectively, while degenerating the performance of MN fusion module from  $81.7\% \pm 2.1\%$  to  $80.3\% \pm 2.0$  and to  $80.6\% \pm 1.9\%$ , respectively, and the CAT fusion module from  $81.8\% \pm 1.7\%$  to  $80.9\% \pm 2.0$  and  $80.9\% \pm 2.1\%$ , respectively.

In conclusion, when compared to the JF structure, the JIF (All) structure consistently improves all fusion modules across the PAD-UFES-20 and SPC datasets, which encompass a wider array of metadata types, as evidenced by the increased averaged BAC value. However, its impact on the ISIC-2019 dataset, which features fewer types of metadata, is varied, except for the MA and MMFA modules. This observation underscores the JIF structure’s capacity for generalization across all fusion modules on datasets rich in metadata. Furthermore, these findings suggest that our JIF structure may exhibit diminished effectiveness for certain fusion modules—particularly those relying solely on metadata to enhance image features or perform basic transformations—when applied to datasets with limited metadata.

### **Effectiveness of Multi-Modal Fusion Attention (MMFA) Module**

To demonstrate the efficacy of the proposed MMFA module, we conduct comparisons with three other fusion modules (CAT, MB, MN, and MA) across various fusion structures and datasets. As illustrated in Table 4.1 and Table 4.2, as well as Table 4.7 and Table 4.8, our MMFA module consistently achieves the highest average BAC value across all three fusion structures: JF, JIF (OFB), and JIF (All). Specifically, on the PAD-UFES-20 dataset, the BAC values are the BAC values are  $74.9\% \pm 2.3\%$ ,

76.7%  $\pm$  2.2%, and 77.2%  $\pm$  2.0%, respectively, while on the SPC dataset, they are 69.6%  $\pm$  3.2%, 71.0%  $\pm$  3.1%, and 72.4%  $\pm$  3.0%, respectively. As depicted in Table 4.3, when combined with the JF structure on the ISIC-2019 dataset, MA exhibits the lowest BAC value of 78.7%  $\pm$  7.9% and MMFA achieves the second-lowest BAC value of 78.9%  $\pm$  5.8%. However, Table 4.9 reveals a shift in performance, with both MA and MMFA securing the top two rankings in terms of BAC value when integrated with the JIF (OFB) (83.8%  $\pm$  1.3% and 83.8%  $\pm$  1.4%) and JIF (All) (83.8%  $\pm$  1.3% and 83.7%  $\pm$  1.3%) structures.

## 4.4 Discussion

### 4.4.1 Effectiveness of using patient’s metadata

The comparative analysis presented in Table 4.1, Table 4.2, and Table 4.3 reveals notable disparities in performance between models utilizing metadata and those that do not. Specifically, JIF-MMFA (All) exhibits substantial improvements on the PAD-UFES-20 dataset and the SPC dataset while showcasing only marginal enhancements on the ISIC-2019 dataset. We attribute these differences to the varying richness of metadata across the datasets. The PAD-UFES-20 and SPC datasets boast 21 and 14 metadata features, respectively, which likely contribute significantly to model performance. Conversely, the ISIC-2019 dataset offers limited patient metadata, such as age, location, and gender, which may not carry as much predictive value.

### 4.4.2 Performance comparison between our JIF-MMFA method and other fusion methods

Indeed, JIF-MMFA (All) demonstrates remarkable performance across a majority of the CNN scenarios, achieving the highest BAC value in 12 out of 15 cases. Notably, in the remaining three scenarios where it does not attain the top spot (Xception on the SPC dataset, and Resnet-50 and Efficientnet-B3 on the ISIC-2019 dataset), JIF-MMFA (All) still showcases comparable performance with the best fusion methods. This underscores the generalization ability of our method for CNNs across diverse scenarios. For instance, consider the scenario involving Efficientnet-B3 on the SPC dataset. Despite not achieving the absolute highest BAC value, the performance gap between JF-CAT, the best-performing method, and our JIF-MMFA (All) is subtle. This further validates the robustness and effectiveness of our proposed method across different CNN architectures and datasets.



The statistical results presented in Table 4.4, Table 4.5, and Table 4.6 provide compelling evidence that JIF-MMFA (All) combined with other fusion methods (excluding JF-MMFA and JIF-MMFA (OFB)) consistently yields statistically significant outcomes ( $p < 0.05$ ). This indicates that JIF-MMFA generally outperforms other fusion methods in the evaluated scenarios.

Next, the confusion matrices displayed in Fig. 4.4 present an interesting result. Generally, the metadata assists the CNN model in increasing the diagnostic rate of all skin diseases. However, the misclassification rate between BCC and SCC is still considerable. This is because these two lesions have not only similar visual features but also many similar values in the metadata. In fact, classifying SCC and BCC is a challenging task, even for experienced dermatologists who use dermoscopy. Nevertheless, this confusion is not a big problem, as both are types of skin lesions and require biopsy for further evaluation. It is a real problem to confuse them with ACK, which is a minor skin disease that is treated without a surgical process [82]. What is more, it is worth noticing that the metadata helps distinguish NEV from MEL, which is quite helpful for the expert’s diagnosis since NEV is benign, circumscribed malformations of the skin, while MEL is one of the most malignant cancers. For the classification of BCC, SCC, and ACK, JIF-MMFA (All) and JF-MB achieve better performance (see Fig. 4.4). A similar phenomenon is also observed in the T-SNE figures (Fig. 4.5) that JIF-MMFA (All) and JF-MB improve the clustering of samples between BCC, SCC, and ACK. However, it is still hard to differentiate the lesions in the sub-clusters. It reflects the problem of inter-class similarity and intra-class variation for skin lesion classification. For MEL, our JIF-MMFA (All) method achieves the best performance according to the averaged BAC value.

Also, The analysis reveals that the Efficient-B3 CNN backbone consistently outperforms other models when utilized in conjunction with JIF-MMFA (All) across various scenarios. Moreover, it shows the most significant improvements compared to using image data alone on the PAD-UFES-20 and SPC datasets. These findings suggest that the Efficient-B3 model is well-suited as the image model ( $I_M$ ) for multi-modal skin disease classification tasks.

Finally, JIF-MMFA increases the parameters of the models that do not use metadata when applied to the CNN backbone, but the increase is not significant. We follow the paper of [82] and only consider the experiments on the PAD-UFES-20 dataset, in which the number of model parameters of Densenet-121, Mobilenet-v2, Resnet-50, Efficientnet-B3, and Xception are increased by 0.08, 0.22, 0.04, 0.05, 0.08 and 0.05. It seems that Mobilenet-v2 is the most impacted model, with an increase of 0.22. However, JIF-MMFA only increases the Mobilenet-v2’s parameters from  $3.6 \times 10^6$  to  $4.4 \times 10^6$ , which is insignificant in terms of training time.



### 4.4.3 Effectiveness of the Multi-Modal Fusion Attention (MMFA) Module

Some interesting phenomenon about MMFA in Table 4.3 and Table 4.9 shows that the MMFA module achieves the second-worst performance when combined with JF, but the second-best performance when combined with JIF on the ISIC-2019 dataset. Similar results also can be seen in MA, which suggest that the CNN with MMFA module cannot conduct the mutual attention mechanism well on image and metadata features when combined with JF structure on the dataset with little metadata (ISIC-2019 dataset) due to the characteristic of MMFA and MA (the fusion module that mutually enhances image and metadata features). Further considering the results of the JF and JIF structures in Table 4.9, we believe that this problem of the JF structure can be handled by the JIF structure that well preserves the modal-specific feature.

## 4.5 Conclusion

In this chapter, we introduce the Joint-Individual Fusion (JIF) structure coupled with the Multi-Modal Fusion Attention (MMFA) module for the classification of skin lesions. Firstly, our proposed MMFA module enhances image and metadata features concurrently through a multi-head self-attention mechanism, resulting in superior performance compared to alternative attention modules. Secondly, we conduct a comprehensive investigation into various fusion structures, contrasting with methods that overlook fusion structure exploration. Moreover, the Joint-Individual Fusion structure we propose facilitates the learning of shared features by preserving modal-specific characteristics, thereby enhancing classification performance across most scenarios. Experimental results across three public datasets demonstrate that our proposed JIF-MMFA achieves the highest averaged BAC value, underscoring the efficacy of both JIF and MMFA components. Furthermore, statistical analyses via the Friedman and Wilcoxon tests corroborate the superiority of our method across all datasets. Notably, experimental results on the ISIC-2019 dataset reveal that, compared to the JF structure, our JIF structure fails to enhance the performance of non-mutual attention fusion modules (CAT, MB, and MN) in datasets with limited metadata. Consequently, our future research endeavors will center on the development of adaptive fusion structures with robust generalization capabilities across diverse scenarios.



# Fusing clinical and dermoscopy images for skin lesion classification

In this chapter, we introduce two multi-modal imaging fusion methods for skin lesion classification. Below is the outline of this chapter to help readers better understand its structure:

Sec. 5.1 provides a gentle introduction to two modality images, a brief discussion on current multi-modal [Skin Lesion Classification \(SLC\)](#) methods, and an explanation of the seven-point checklist features. Sec. 5.2 covers Related works on multi-modal SLC. The above sections serve as foundational parts for understanding the proposed methods. The details of our proposed methods will be presented in Sec. 5.3 and Sec. 5.4 respectively.

## 5.1 Introduction

In addition to fusing single-image modality and patient metadata, there are also two imaging modalities ([Clinical Images \(CI\)](#) and [Dermoscopy Images \(DI\)](#)) that need to be efficiently fused for [Skin Lesion Classification \(SLC\)](#). For the presentation of localized visual features, DIs, are captured using a high-resolution magnifying (e.g. dermatoscopy and epiluminescence microscopy) imaging device [117] in direct contact with the skin. In contrast, CIs, taken with a standard digital camera or smartphone, exhibit more variations in terms of view and angle [39]. In contrast to single-modality-based SLC, multi-modal-based SLC harnesses complementary information from both modalities and leads to a more accurate and robust diagnosis, driving further exploration on this topic [136].

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. From the perspective of clinical diagnosis, dermatologists typically examine patients in person over one or multiple visits rather than rely solely on one imaging modality [129]. 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. [60, 129] were among the first to propose fusing multi-modal features using concatenation for skin lesion classification. Subsequent research of [110, 36] improved performance by integrating prediction information and feature fusion. To further enhance the diagnostic accuracy, [11, 46, 136] introduced more advanced fusion modules for the feature interaction of clinical and dermoscopy images. They argued that more than simple concatenation is needed to fully exploit the information from both modalities. However, the introduction of fusion modules requires significant computational costs, limiting their applications in real-world scenarios.

Hence, our primary objective is to explore an MM-SLC framework that achieves a favorable parameter/accuracy trade-off, i.e., significantly reducing the model’s parameters while only modestly affecting its accuracy. In this chapter, we introduce two novel fusion structures from different perspectives: one is constructed based on prior knowledge, and the other is developed using a deep learning scheme with a parameters-sharing network.

## 5.2 Related works of multi modal-based skin lesion classification

Despite the success of single modality-based methods for SLC, they tend to deviate from routine dermatologists’ examinations [136] and overlook the potential to enhance diagnostic accuracy by exploiting complementary information from both modalities. To fill this gap, increasing works about MM-SLC were presented [39, 129, 60, 110, 36, 46, 136].

[39] and [129] extracted the features from clinical images and dermoscopy images using VGG-16 [101] and Resnet-50 [45] and then fused the features by a simple concatenation to learn a joint representation for final prediction. The introduction of the Seven-Point Checklist (SPC) dataset by [60] marked a significant advancement in multi-modal skin lesion classification. They proposed a multi-modal framework based on Inception-V3 [106] for the simultaneous diagnosis classification and the seven-point checklist.

After the SPC dataset’s release, many multi-modal approaches have been proposed

for multi-label skin lesion classification. To enhance performance, [110] and [36] introduced weighted-fusion and graph-based fusion schemes, respectively. Both approaches combine CI and DI predictions in the model’s late stages.

More recently, [46] and [136] recognized the limitations of the simple concatenation operation used in former methods. They introduced multiple bidirectional attention blocks to mutually enhance CI and DI, facilitating efficient interaction between these modalities across multiple scales.

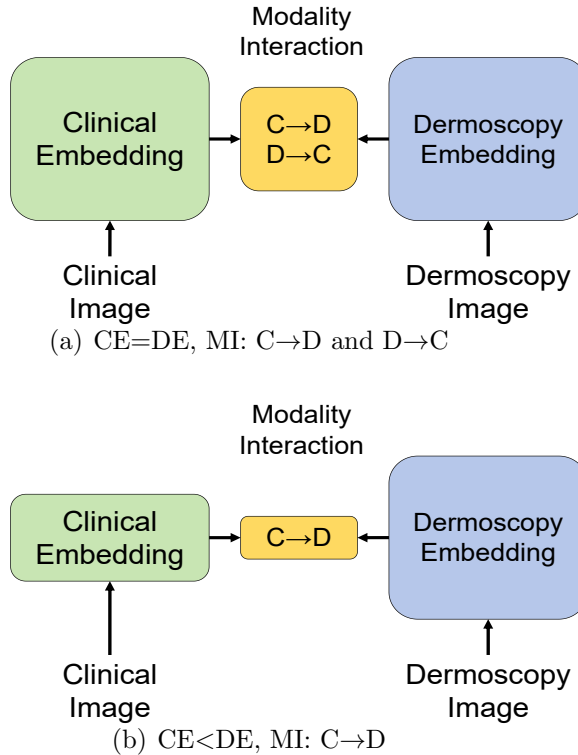
However, in these methods, the emphasis on clinical images (CI) may impact their performance, considering that CI is regarded as supplementary information to the DI. Also, the introduction of additional fusion branches incurs high computational costs, which may hinder their application in various scenarios, such as deploying on mobile devices or implementing local AI-enabled family doctor systems for skin care.

### 5.2.1 Seven-point checklist features

In the commonly-used dataset, Derm7pt or [Seven-Point Checklist \(SPC\)](#) <sup>1</sup>, the classification tasks include not only classifying skin lesions into several types of diseases, such as nevus, basal cell carcinoma, and melanoma but also detecting the seven melanoma-associated dermoscopic features (as shown in [Table 5.1](#)) [60].

Pattern analysis of multiple subtle skin lesion features is a common method for experienced dermatologists to differentiate between benign and malignant skin tumors. To simplify the diagnostic procedure, rule-based diagnostic algorithms such as the ABCD rule [80] and the 7-point checklist [3] have been proposed and are widely accepted [15]. Specifically, the seven-point checklist criteria assign seven labels to a skin lesion, each with a corresponding score. For example, irregular dots and globules are scored as one, while absent dots and globules are scored as zero. When the total score of a skin lesion exceeds a certain threshold, it is assessed as melanoma [3, 4].

It can be challenging for dermatologists to understand how DL-based methods make diagnoses and to explain them to patients. Detecting these criteria may assist in developing more interpretable diagnostic models [60].



**Figure 5.1:** The comparison between former MSLA methods and our method. The height of each rectangle denotes its relative computational size. CE, DE, and MI are short for clinical embedder, dermoscopy embedder, and modality interaction.

### 5.3 Pay Less On Clinical Images: Asymmetric Multi-Modal Fusion Method For Efficient Multi-Label Skin Lesion Classification

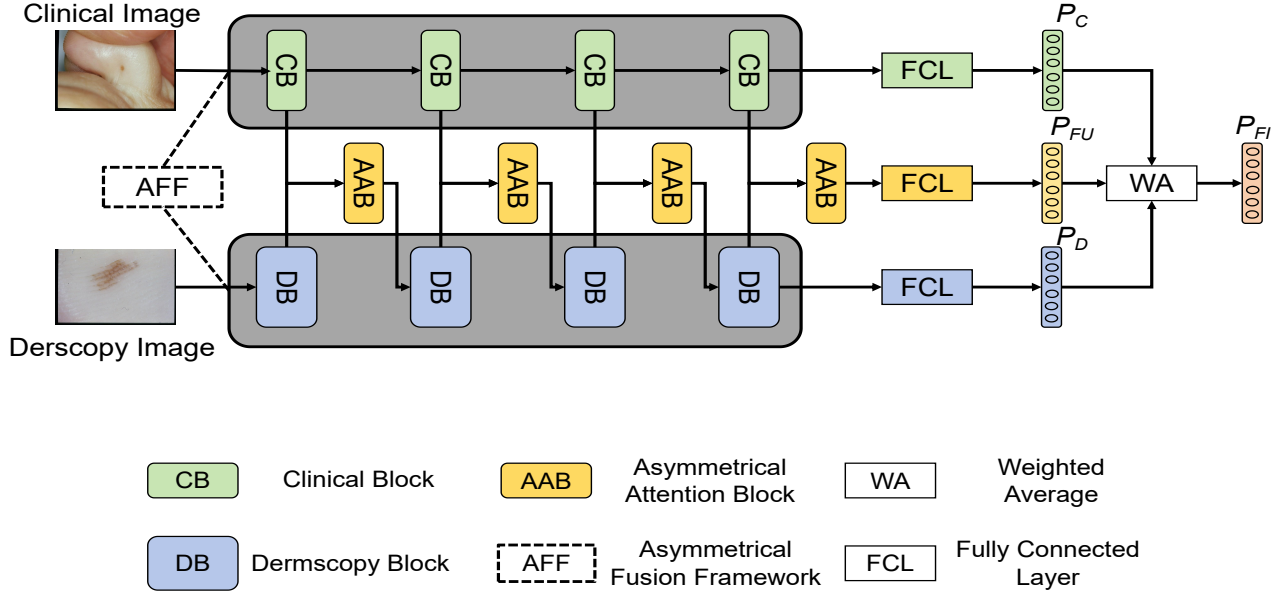
Our idea of this method is motivated by two observations: (1) Clinical Statistics: According to experienced dermatologists, the diagnostic accuracy of melanoma based on Dermoscopy Imaging (DI) is 25% higher than that of visual inspection with the naked eye. Visual features observed by the naked eye are akin to those captured by standard cameras and smartphones [113]. (2) Experimental Results of DL Algorithms: In diagnosis tasks using the SPC dataset, the majority of papers [60, 11, 110, 36, 46, 136] reported a 6% higher accuracy in Dermoscopy Images (DI)-based diagnosis

<sup>1</sup><http://derm.cs.sfu.ca/>

compared to [Clinical Images \(CI\)](#). Additionally, [29] demonstrated an increased diagnostic accuracy with DI (85%) compared to CI (75%). Considering both observations, it is evident that a significant amount of key diagnostic information comes from DI rather than CI. Therefore, employing two identical structures to extract information from DI and CI individually is unreasonable.

Inspired by that, in this paper, we propose a novel Asymmetrical Multi-Modal Fusion Method (AMMFM) for efficient multi-label skin lesion classification. Our approach differentiates itself from previous methods in two key aspects, i.e., Asymmetric Fusion Framework (AFF) and Asymmetric Attention Block (AAB): Firstly, differing from the commonly used Symmetrical Fusion Framework (SFF), our AFF incorporates the prior domain knowledge into the structure design. AFF utilizes an advanced model, e.g., ResNet, ConvNext, and SwinTransformer [45, 72, 71], for capturing the primary diagnostic information from DI, but a lightweight deep model, i.e., MobilenetV3 [49] for the supplementary information from CI. Compared to SFF, AFF significantly reduces the model’s parameters with only a subtle decrease in accuracy. Secondly, in contrast to previous methods utilizing bidirectional attention blocks (BAB) to mutually enhance DI and CI (Fig. 5.1.a), we believe that enhancing the supplementary information CI may lead to overfitting, affecting the final classification. Therefore, we propose an asymmetric attention block (AAB) that exclusively leverages the features of CI to enhance those of DI (Fig. 5.1.b), achieving superior performance to BAB with fewer model parameters. In total, our contributions can be summarized as follows:

1. Inspired by prior knowledge, we introduce a novel Asymmetrical Fusion Framework (AFF) that significantly reduces the model’s parameters while maintaining unchanged or slightly decreased classification accuracy compared to the currently used Symmetric Fusion Framework (SFF).
2. We present a new Asymmetrical Attention Block (AAB) that exclusively utilizes features extracted from clinical images (CI) to enhance those of dermoscopy images. This approach addresses potential accuracy impacts associated with focusing on supplementary information from CI. In comparison to the former Bidirectional Attention Block (BAB), our AAB demonstrates improved classification performance with fewer parameters.
3. Our proposed Asymmetrical Multi-Modal Fusion Method achieves state-of-the-art performance in both accuracy and model parameters. The extensive results confirm the effectiveness of our proposed AFF and AAB, demonstrating their applicability to various deep-learning algorithms, including both CNN and Transformer structures.



**Figure 5.2:** The overview of our Asymmetric Multi-Modal Fusion Method. Clinical and dermoscopy blocks are used to extract the features from clinical and dermoscopy images, respectively.  $P_C$ ,  $P_D$ , and  $P_{FU}$  are the predictions from the clinical branch (green), dermoscopy branch (blue), and fusion branch (yellow).

### 5.3.1 Related works about asymmetric fusion structure

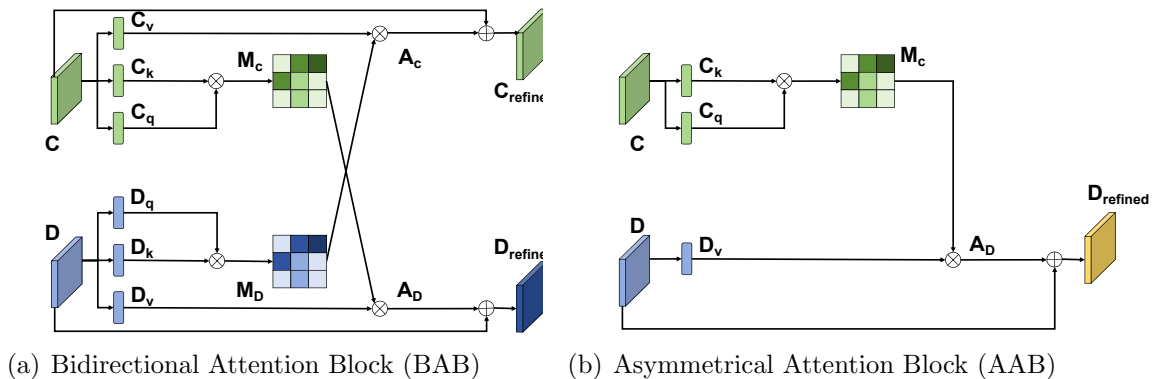
A limited number of works on the Asymmetrical Fusion Model (AFM) [124, 138, 38, 125, 122] have been proposed for various computer vision tasks. For example, [138] introduced an asymmetric non-local network to fuse multi-scale features for semantic segmentation. [38] presented an asymmetrical model to extract asymmetric relations between humans and objectives for action recognition. Additionally, [122] proposed an asymmetrical fusion framework focusing on the gallery side for image retrieval. In the medical domain, [123] employed an asymmetrical model to address issues in 3D slices for universal lesion detection.

These methods are tailored to specific modalities (e.g., 3D slices [38]) or tasks (e.g., semantic segmentation [138], action recognition [38], and image-text retrieval [122]). Their design can not directly apply to the Multi-Modal Skin Lesion Classification (MM-SLC) task. However, the successes of AFM in different fields encourage us to explore the potential of asymmetrical fusion methods for the MM-SLC task.



### 5.3.2 Asymmetrical Multi-Modal Fusion Method

As illustrated in Fig. 4.2, the proposed Asymmetric Multi-Modal Fusion Method (AMMFM) is constructed with four components: Asymmetric Fusion Framework (AFF), Asymmetric Attention Blocks (AABs), Fully Connected Layers (FCLs), and Weighted Averaging (WA) operation. AFF and AABs are pivotal components that extract information from two modalities and facilitate interaction. These two components are core to our methodology, and we will delve into their details in the following subsections. FCLs are used to classify the extracted features in three branches: clinical branch (highlighted in green), dermoscopy branch (highlighted in blue), and fusion branch (highlighted in yellow), as depicted in Fig. 4.2. WA is employed to fuse predictions from clinical, dermoscopy, and fusion branches, namely,  $P_C$ ,  $P_D$ , and  $P_{FU}$ , resulting in the final prediction  $P_{FI}$



**Figure 5.3:** The computational graph of the former bidirectional attention block (BAB) and our asymmetrical attention block (AAB).

#### 5.3.2.1 Asymmetric Fusion Framework

Before introducing the Asymmetric Fusion Framework (AFF), let us define easier understanding: in our article, the framework refers explicitly to the composing manner of two feature extractors, excluding the modality interaction modules (see Fig. 5.2).

Currently, most methods are based on a symmetrical fusion structure (SFF), utilizing two identical structures to extract features from clinical and dermoscopy images. However, relevant research has demonstrated that the accuracy based on dermoscopy is much higher than that based on naked eyes and clinical images captured by smartphones with a standard digital camera. The seven-point checklist criteria are also proposed based on features observed under dermoscopy. These phenomena

drive us to consider dermoscopy images primary, with clinical images regarded as supplementary for the multi-label skin lesion classification task. Therefore, we propose an asymmetrical fusion structure (AFF) for this task.

Specifically, as shown in Fig. 4.2, AFF utilizes two different models for the feature extractions from two modalities respectively, a lightweight model, MobilenetV3 [49], for clinical images (CI) and an advanced model requiring much more parameters, e.g., ResNet, ConvNext and SwinTransformer [45, 72, 70], for dermoscopy images (DI). Compared to that, our AFF can significantly reduce the model’s parameters while maintaining unchanged or subtly decreased classification performance. This is because replacing an advanced model with a lightweight model in the clinical branch may affect the information captured from Clinical Images (CI), so it is important to note that clinical images are considered supplementary in our pipeline. Therefore, this change can only slightly impact the final classification.

### 5.3.2.2 Asymmetric Attention Block

Building on the discussion in Sec.5.2 and 5.3, the information from the clinical branch is considered supplementary in this paper. Treating it as equal to dermoscopy information and enhancing the supplementary information is not reasonable in our pipeline. Therefore, we introduce an asymmetric attention block (AAB) for the modality interactions between clinical and dermoscopy images.

In contrast to the former bidirectional attention block (BAB) [46] that mutually enhances the features of both modalities (see Fig.5.3(a)), our AAB only adopts clinical features to generate an attention map for enhancing dermoscopy features. This design allows us to save approximately half of the parameters compared to BAB (See Fig.5.3(b)). Like BAB, AAB is embedded into different stages of deep learning models to facilitate the interaction of multi-scale features from the two modalities.

In our AAB, the inputs are the extracted clinical features  $C \in \mathbb{R}^{H \times W \times C}$  and dermoscopy features  $D \in \mathbb{R}^{H \times W \times C}$ , both of which have the same size ( $H$ ,  $W$ , and  $C$  indicate the height, width, and channel number of the features, respectively). As shown in Fig. 5.3(a), firstly, two  $1 \times 1$  convolutions are applied to  $C$  to generate  $C_k$  and  $C_q$ , and one  $1 \times 1$  convolution is employed on  $D$  to obtain  $D_v$ , where  $(C_k, C_q, D_v) \in \mathbb{R}^{H \times W \times C}$ . Then,  $C_k$  and  $C_q$  are reshaped to  $\mathbb{R}^{N \times C}$ , where  $N = H \times W$ . Next, a multiplication is conducted between the reshaped  $C_k$  and  $C_q$ , followed by a non-linear activation function *Softmax* to generate the attention map  $M_c \in \mathbb{R}^{N \times N}$ . Finally, the refined dermoscopy features are obtained based on Eq. 5.1,

$$D_{refined} = D_v \cdot M_c + D, \tag{5.1}$$

where  $\cdot$  indicates matrix dot product operation, and  $+$  indicates matrix summation.

### 5.3.3 Loss Function and Final Prediction

The total loss  $L_{total}$  used to optimize our model is as follows:

$$L_{total} = L_{derm} + L_{clinc} + L_{fusion}, \quad (5.2)$$

where  $L_{derm}$ ,  $L_{clinc}$  and  $L_{fusion}$  are the multi-label classifications losses for the dermoscopy image branch ( $P_D$  in Fig. 5.2), clinical image branch  $P_C$  in Fig. 5.2 and fusion image branch  $P_{FU}$  in Fig. 5.2, respectively. All the losses are computed as Eq. 5.3

$$L_K = \sum_j^X \sum_i^Y CE(D^j, C^j, G_i^j, P_i^j; \theta_K), \quad (5.3)$$

where  $X$  is the batch size in our training,  $Y=8$  indicates the number of the multi-label classification tasks (see Table 4.1),  $C^j$  and  $D^j$  represent input pairs of dermoscopy and clinical images respectively.  $G_i^j$  and  $P_i^j$  are the corresponding ground truths and predictions respectively and  $\theta_K$  is the parameters of our model.  $CE$  indicates the cross-entropy loss.

During the testing stage, we use a weighted average scheme to fuse  $P_D$ ,  $P_C$  and  $P_{FU}$  into the final prediction  $P_{FI}$  for the evaluation as follows:

$$P_{FI} = W_D * P_D + W_C * P_C + W_{FU} * P_{FU}, \quad (5.4)$$

where  $W_D$ ,  $W_C$  and  $W_{FU}$  are the corresponding weights for  $P_D$ ,  $P_C$  and  $P_{FU}$ , respectively, which are obtained by conducting a weight search scheme on validation dataset [110].

### 5.3.4 Experiments and Discussion

#### 5.3.4.1 Implementation Details

We use Adam [62] with a batch size of 24 to optimize our model for 250 epochs during training. Data augmentations, including flipping, shifting, scaling, rotating, and brightening operations, are randomly conducted to enhance the model’s generalization ability. Stochastic Weights Averaging (SWA) [56] scheme is used in the last 50 epochs to generate the final weights for evaluation. All images are resized to  $224 \times 224 \times 3$  for training and evaluating the model. Following [110], testing time augmentation is also used during the evaluation to improve classification performance. All of our experiments are conducted on a GPU of NVIDIA A100-PCIE-40GB. Without special instructions, our AMMFM is based on MobinetV3 for clinical images and Swin-Transformer for dermoscopy images, as it achieves the best classification performances in our experiments.

**Table 5.1:** Details of the SPC dataset.

Classification Task	Name	Abbrev.	Num.
Diag	Basal Cell Carcinoma	BCC	42
	Nevus	NEV	575
	Melanoma	MEL	252
	Miscellaneous	MISC	97
	Seborrheic Keratosis	SK	45
PN	Absent	ABS	400
	Typical	TYP	381
	Atypical	ATP	230
STR	Absent	ABS	653
	Regular	REG	107
	Irregular	IR	251
PIG	Absent	ABS	588
	Regular	REG	118
	Irregular	IR	305
RS	Absent	ABS	758
	Present	PRS	253
DaG	Absent	ABS	229
	Regular	REG	334
	Irregular	IR	448
BWV	Absent	ABS	816
	Present	PRS	195
VS	Absent	ABS	833
	Regular	REG	117
	Irregular	IR	71

### 5.3.4.2 Dataset and Metrics

The effectiveness of our AMMFM is evaluated on the well-recognized seven-point checklist (SPC) dataset [60], which contains 1011 patients’ cases. Each case includes a pair of dermoscopy and clinical images, a diagnosis (Diag) label, and labels of seven-point checklist (SPC) features. As shown in Table 5.1, Diag has five categories: BCC, NEV, MEL, MISC, and SK, and the SPC labels include Pigment Network (PN), Streaks (STR), Pigmentation (PIG), Regression Structures (RS), Dots and Globules (DaG), Blue Whitish Veil (BWV), and Vascular Structures (VS), which are divided into the following categories: ABS, PRS, TYP, ATP, REG, and IR.

Building on previous works [60, 36, 110, 46, 136], metrics including accuracy (ACC),

**Table 5.2:** The comparison between our method and other currently advanced methods is based on averaged AUC values. The highest and second highest values in each column are bolded 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	PRS		REG
Incep-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	<b>83.6</b>	81.3	81.9	77.7	82.6	89.8	87	<i>82.7</i>			84.3
FM-FS	95.3	<i>92.6</i>	89	<b>94</b>	89.2	<i>85.9</i>	83.9	<i>87.9</i>	81.4	80.9	83.5	81.7	<i>79.1</i>	80.1	90.6	<i>87.8</i>	78			86
GIIN	92.8	86.8	87.6	88.8	79.8	80.1	<b>87.5</b>	84.9	81.2	81.1	83.6	79	78.6	<i>83.1</i>	90.8	80.7	75.4			83.6
CAFNet	<b>97.1</b>	<b>92.7</b>	<b>92.2</b>	92.5	<i>91</i>	81.9	75.3	87.4	<b>85.4</b>	76.1	<i>85</i>	<b>85.4</b>	75.2	78.7	<b>94.7</b>	84.8	<b>83.5</b>			85.8
AMMFM	<i>95.8</i>	92.4	<i>89.3</i>	<i>93.7</i>	<b>91.7</b>	<b>87.1</b>	<i>86.7</i>	<b>91.5</b>	<i>85.2</i>	<i>82.5</i>	<b>86.2</b>	<i>83.6</i>	<b>79.8</b>	<b>86.0</b>	<i>94.1</i>	<b>90.2</b>	82.4			<b>88.1</b>

**Table 5.3:** The comparison between our method and other currently advanced methods based on averaged ACC values. 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	77.2	74.2	73.7
HcCNN	70.6	87.1	<b>84.8</b>	68.6	71.6	<b>65.6</b>	80.8	69.9	74.9
FM4-FS	70.9	86.8	81.8	<i>72.4</i>	74.4	61	<b>83</b>	74.9	75.7
CAFNet	70.1	87.8	84.3	<b>73.4</b>	<i>77</i>	61.5	<i>81.8</i>	<b>78.2</b>	<i>76.8</i>
TFormer	70.9	86.4	83.5	68.8	74	64.9	81.3	73	75.3
<b>AMMFM</b>	<b>72.7</b>	<b>89.1</b>	82.3	<i>72.4</i>	<b>78.7</b>	<b>65.8</b>	81	<i>75.2</i>	<b>77.2</b>

area under the curve (AUC), precision (Prec), specificity (SPE), and sensitivity (SEN) are used for the evaluation of our method.

### 5.3.4.3 Comparisons with Currently Existing Methods

In Tables 5.2 and 5.3, a comparative analysis was conducted to assess the performance of the proposed (AMMFM) against contemporary classification methodologies utilizing clinical and dermoscopy images. The evaluated methods encompass Inception-combined [60], HcCNN [11], FusionM4Net-FS [110], GIIN [36], CAFNet [46], and TFormer [136].

In Table 5.2, a comparison followed by [11, 36] was adopted to compare selected features, gauged by the AUC for each method. Table 5.3 presents a comprehensive comparison of all methods regarding accuracy. Note that results for all other methods

**Table 5.4:** The comparisons between our AMMFM and other methods in melanoma-related features (%).

Metric	Method	DIAG	PN	STR	PIG	RS	DaG	BWV	VS	AVG
		MEL	ATP	IR	IR	PRS	IR	PRS	IR	
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
	FM-FS	89	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	85.4	78.7	94.6	83.4	85
	<b>AMMFM</b>	89.3	86.7	85.2	86.2	83.6	86.0	94.1	82.4	<b>86.7</b>
PRE	Incep-com	65.3	61.6	52.7	57.8	56.5	70.5	63	30.8	57.3
	HcCNN	62.8	62.3	52.4	65.1	81.6	69.6	91.9	50	67
	FM-FS	65.7	82.2	56.2	67.6	82	67.2	64.9	42.9	68.5
	GIIN	65.6	48.4	50.4	82.3	73.5	74.9	67.4	100	<b>70.3</b>
	CAFNet	77.9	50.8	54.8	70.1	76.7	67.8	75.4	58.3	66.5
	<b>AMMFM</b>	59.4	57.0	56.4	63.7	51.9	77.4	66.7	16.7	56.1
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	59.7
	FM-FS	62.4	49.5	47.9	58.9	47.1	68.4	66.7	20	52.6
	GIIN	59	77.5	67	39.2	21.9	70.1	69.9	3.6	51
	CAFNet	75.3	65.9	67.1	60.3	42.7	74.1	68.8	45	62.4
	<b>AMMFM</b>	71.4	63.1	64.6	71.8	69.6	74.9	73.5	31.3	<b>65</b>
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	86.3	41.5	71.6	65.3	98.4	79.2
	FM-FS	88.8	90.1	88.4	88.1	96.2	72.9	91.6	97.8	89.2
	GIIN	89.5	79	80.3	95.8	96.8	78.8	91	100	88.9
	CAFNet	93.6	90.2	91.2	89.1	96.5	74	95.1	98.7	<b>91.1</b>
	<b>AMMFM</b>	86.8	87.1	86.9	84.2	83.9	81.1	92.4	93.4	87.0

were quoted from their respective publications or sourced from [46]. These results are assumed to represent the best performance, except for TFormer, where mean values and standard deviation were reported. Consequently, the comparison between AMMFM and TFormer is based on mean values, while comparisons with other methods are grounded on the highest value.

As shown in Table 5.2, AMMFM attains the highest performance, boasting an

**Table 5.5:** 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 (Mb)
Incep-com	83.7	73.7	> 57,4
HcCNN	84.3	74.9	> > 51.2
FM-FS	86	75.7	54.45
GIIN	83.6	-	>51.2
CAFNet	85.8	76.8	> > 51.2
TFormer	-	75.3	77.76
<b>AMMFM</b>	<b>88.1</b>	<b>77.2</b>	<b>33.06</b>

averaged (AVG) AUC of 88.1%. This outperforms the second-best method, FM-FS, with 86%, and the third-based method, CAFNet, with 85.8%, by 2.1% and 2.3% respectively. AMMFM demonstrates superiority by achieving the highest values in seven categories and the second-highest in eight categories, showcasing its excellence across all eight classification tasks. AMMFM outperforms FM-FS in the Diag task and most Seven-Point features tasks. In comparison to CAFNet, AMMFM achieves comparable performance in the Diag task (AVG AUC: CAFNet: 93.1%, AMMFM: 92.6%) and significantly better performance in other Seven-Point feature tasks (AVG AUC: CAFNet: 83.0%, AMMFM: 86.4%), establishing its overall superiority over CAFNet.

Similarly, in Table 5.3, AMMFM secures the highest averaged accuracy (AVG Acc) value of 77.2%, outperforming all other methods. It attains the highest values in four classification tasks (PN, BWV, STR, DaG) and the second-highest values in two tasks (PIG, Diag). CAFNet and FM-FS secure the second-best and third-best AVG Acc in Table 5.3. These results underscore the superior performance of AMMFM and the efficacy of the cross-attention modules in CAFNet and the late fusion scheme in FM-FS. For further insights, Table 5.4 details that AMMFM achieves the highest AUC and sensitivity values in melanoma-related features, substantiating its proficiency in melanoma detection. However, our AMMFM achieves the lowest precision value because of the bad performance in the categories of RS-PRE (51.9 %) and VS-IR (16.7%). In particular, VS-IR is hugely lower than other methods. We attribute this to the unbalanced categories in RS-PR (ABS: 758 PRE: 253) and VS-IR (ABS: 833, REG: 117 and IR: 91, see Table 5.1).

In Table 5.5, we compare all the methods mainly based on the model’s parameters, and for convenience, we also present the AVG AUC and AVG ACC in this table. Since only the source codes for FM-FS and TFormer were available, we estimated the model parameters for Incep-com, HcCNN, GIIN, and CAFNet, albeit with

**Table 5.6:** Ablation studies of our AMMFM in terms of AVG AUC, AVG ACC, and model parameters. (%)

AFF	AAB	AVG AUC	AVG ACC	Parameters(Mb)
	Baseline	87.3±0.4	76.6±0.5	58.49
✓		87.3±0.4	76.5±0.4	32.48
✓	✓	87.6±0.4	76.7±0.4	33.06

some approximations. In our estimation, we calculated the parameters based on the employed backbones, specifically, two InceptionV3 (57.4Mb) for Incep-com and two ResNet-50 (51.2Mb) for GIIN, HcCNN, and CAFNet. However, concerning Incep-com and GIIN, the model parameters are marginally higher than those of their utilized backbones, attributed to the absence of multiple blocks employed in constructing the third branch. Concerning HcCNN and CAFnet, which construct a third branch utilizing attention and ResNet blocks, the model parameters are expected to significantly exceed those of their backbones. As shown in the table. 5.5, our AMMFM achieves the highest values in both AVG AUC and AVG ACC with the least model’s parameters (33.06Mb), demonstrating the great accuracy/parameter trade-off of our AMMFM.

#### 5.3.4.4 Ablation Studies

In the following experiments, all the models are trained and tested ten times to obtain the mean value and standard deviation for a fair comparison.

In Table 5.6, ablation studies are conducted to analyze the two primary components of our AMMFM: the asymmetrical fusion framework (AFF) and the asymmetrical attention block (AAB). For comparative purposes, we establish a baseline utilizing a commonly used symmetrical fusion framework (SFF) based on two Swin-Transformer (ST) models and a concatenation operation, serving as a reference point in the ablation studies.

Compared to the baseline model(second row), the proposed AFF with concatenation operation (third row) can significantly reduce the parameters from 58.49M to 32.48M without compromising the performance metrics, as evidenced by the maintained the AVG AUC (Baseline: 87.3%, AFF: 87.3%) and AVG ACC (Baseline:76.6%, AFF: 76.5%). These outcomes substantiate our initial hypothesis that substituting an advanced model (ST) with a more lightweight model, MobileNetV3 (MN), for information extraction from clinical images would have a subtle or negligible impact on overall performance. This observation underscores the supplementary nature of



clinical images in the context of multi-label skin lesion classification tasks, where dermoscopy images remain the primary source of information. Subsequently, ABB contributes to further enhancements in AFF’s performance. This improvement is observed across both metrics, with AVG AUC increasing from 87.3% for AFF to 87.6% for AFF+AAB and AVG ACC from 76.5% to 76.7%. Remarkably, this performance boost is achieved with only a marginal increase in model size, rising from 32.48Mb for AFF to 32.06Mb for AFF+AAB, denoting the effectiveness of our AAB in accuracy/parameters trade-off.

**Table 5.7:** Comparison between single-modal, baseline multi-modal, and our proposed multi-modal methods. Clic: Clinical Images, Derm: Dermoscopy Images, MN: MobilenetV3, ST: Swin-Transformer, Param: Parameters and - indicates no model for extracting the information from dermoscopy and clinical images, i.e., single-modal methods. (%)

Method	Clic	Derm	AVG AUC	AVG ACC	Params(Mb)
Single-Modal	-	ST	86.8±0.3	76.2±0.6	28.82
	ST	-	77.7±0.5	68.7±0.7	
	-	MN	83.5±0.3	72.6±0.3	2.91
	MN	-	75.8±0.4	67.6±0.5	
Baseline	ST	ST	87.3±0.4	76.6±0.5	58.49
	MN	MN	84.9±0.3	73.4±0.2	6.48
AMMFM	ST	MN	85.0±0.3	73.8±0.3	32.62
	MN	ST	87.6±0.4	76.7±0.4	33.06

### 5.3.4.5 Comparison between single-modal, baseline multi-modal and our proposed multi-modal methods

In Table 5.7, we conduct a comprehensive comparison among single-modal, baseline multi-modal (SFF with concatenation), and our proposed multi-modal approach (AFF with AAB). Compared to clinical images, the results reveal a substantial performance improvement when utilizing dermoscopy images, regardless of whether they are based on MN or ST. Specifically, there is an increase of over 7% in AVG AUC and 5% in AVG ACC values, underscoring the pronounced significance of dermoscopy images in multi-label classification tasks. Furthermore, the baseline multi-modal methods exhibit an additional increase in accuracy compared to their Derm-based counterparts. This emphasizes the complementary nature of clinical images, which provide supplementary information to dermoscopy images. In the

single-modal approaches, we are substituting ST with MN for dermoscopy image processing, resulting in a 3.6% reduction in both metrics. Conversely, replacing ST with MN for clinical image processing shows a more modest 1.1% AVG ACC. Also, compared to another counterpart in AMMFM approaches, employing ST for the dermoscopy branch and MN for the clinical branch enhances both metrics by approximately 3%. These results demonstrate and support our AFF’s effectiveness and affirm the chosen architecture’s suitability for efficiently optimizing performance in multi-modal classification tasks.

In Table 5.8, we present detailed information about ST-based single-modal and multi-modal methods, facilitating a nuanced analysis of their impact on individual classification tasks. Examining the table reveals that Derm-based ST consistently outperforms Clic-based ST across all categories (CTs), a result aligned with expectations given that the seven-point checklist criteria are formulated based on observed features under dermoscopy [60]. Moreover, compared to Derm-based ST, the baseline and our AMMFM demonstrate performance improvements across nearly all CTs. This observation illustrates the complementary role of clinical images in enhancing the overall performance when combined with dermoscopy images.

#### 5.3.4.6 Comparison between bidirectional attention block (BAB) and asymmetrical attention block (AAB)

To delve deeper into the impact of the proposed asymmetrical attention block (AAB), a comparative analysis is conducted with other fusion blocks within different fusion frameworks, namely the symmetrical fusion framework (SFF) and the asymmetrical fusion framework (AFF). As illustrated in Table 5.9, the block attention block (BAB) demonstrates performance improvement over concatenation (CAT) for both fusion frameworks, highlighting the effectiveness of multi-modal interactions. Notably, our proposed AAB further enhances the accuracy achieved by BAB within both SFF and AFF. This observation supports our assumption that over-augmenting the importance of clinical supplementary information in the multi-modal pipeline may impact the classification task. At the same time, ABB also shows superiority in the model’s parameters compared to BAB.

#### 5.3.4.7 Generalization ability of AMMFM using different backbones

To assess the generalization capability of our AMMFM across various backbones, we employ ResNet-50, ConvnextTiny, and Swin-Transformer as backbone architectures for comparative analysis. As depicted in Table 5.10, our AMMFM consistently outperforms the baseline multi-modal method across all three backbones while maintaining

**Table 5.8:** A detailed comparison between ST-based single-modal and multi-modal in terms of AVG AUC is needed. CT: classification task, CG: category, Clic: clinical images, Derm: Dermoscopy Images. (%)

CT	CG	Clic (ST)	Derm (ST)	Baseline (two STs)	AMMFM (Clic: MN Derm: ST)
Diag	BCC	85.5±3.8	94.5±1.3	95.0±0.9	<b>95.0±0.8</b>
	NEV	85.8±0.7	91.8±0.5	<b>92.4±0.4</b>	<b>92.4±0.4</b>
	MEL	79.5±1.1	89.1±0.6	89.3±0.8	<b>89.5±0.7</b>
	MISC	86.5±1.8	93.7±0.8	93.9±0.9	<b>94.7±0.7</b>
	SK	74.5±4.2	87.5±3.2	88.3±2.0	<b>90.4±1.4</b>
PN	TYP	80.4±1.1	87.5±0.6	<b>88.1±0.7</b>	87.7±0.8
	ATP	73.4±1.6	85.1±0.7	85.7±0.7	<b>86.2±0.8</b>
STR	REG	79.9±2.2	89.3±1.1	88.6±1.3	<b>89.4±1.8</b>
	IR	71.0±1.6	83.9±1.0	84.2±0.9	<b>84.6±1.1</b>
PIG	REG	68.3±2.1	82.2±1.5	81.4±1.4	<b>83.5±1.1</b>
	IR	73.6±1.7	85.5±1.3	<b>85.5±0.9</b>	85.4±1.1
RS	PRS	72.9±1.3	82.5±0.9	<b>83.5±0.9</b>	83.4±0.8
DaG	REG	72.6±1.4	79.3±1.0	80.2±0.9	<b>80.3±0.6</b>
	IR	73.8±0.9	84.1±0.8	84.1±0.9	<b>84.4±0.7</b>
BWV	PRS	83.9±1.2	92.7±0.7	<b>93.8±0.5</b>	93.8±0.8
VS	REG	81.7±2.0	86.2±1.1	<b>88.3±0.8</b>	87.9±1.3
	IR	77.0±2.2	80.4±1.7	<b>82.2±1.5</b>	81.3±2.3
AVG		77.7±0.5	86.8±0.3	87.3±0.4	<b>87.6±0.4</b>

significantly fewer parameters. This substantiates the robustness of our AMMFM, showcasing its ability to deliver superior performance across diverse deep-learning backbones.

### 5.3.5 Conclusion

In this section, we introduced a novel Asymmetrical Multi-Modal Fusion Method (AMMFM) for efficient multi-label skin lesion classification, driven by the observation that dermoscopy images provide more crucial information than clinical images. Our

**Table 5.9:** The comparison between our proposed asymmetrical attention block (AAB) and other fusion blocks (FBs), including concatenation (CAT) and bidirectional attention block (BAN), in different fusion frameworks (FS). SFF: symmetrical fusion framework, AFF: asymmetrical fusion framework (%).

FS	FB			AVG AUC	AVG ACC	Params(Mb)
	CAT	BAB	AAB			
SFF	✓			87.3±0.4	76.6±0.5	58.49M
		✓		87.4±0.3	76.5±0.4	60.45M
			✓	<b>87.7±0.7</b>	<b>77.1±0.4</b>	59.09M
AFF	✓			87.3±0.4	76.5±0.4	32.48M
		✓		87.6±0.3	76.6±0.3	33.88M
			✓	<b>87.6±0.4</b>	<b>76.7±0.4</b>	33.06M

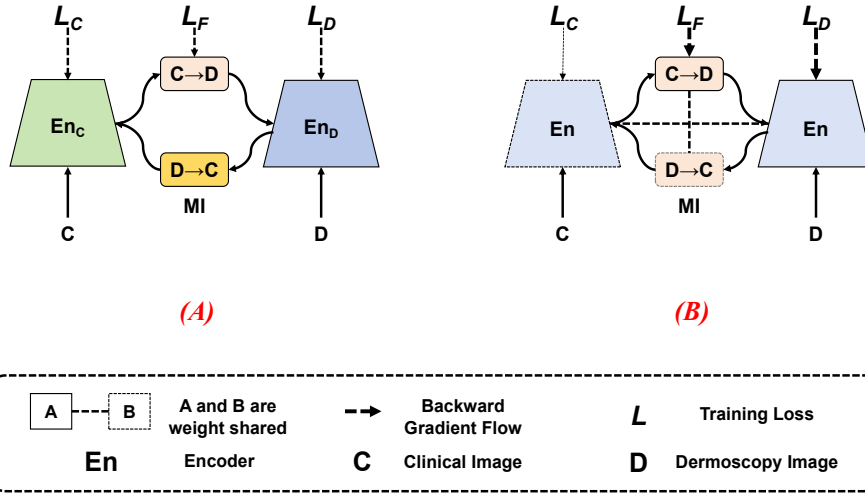
AMMFM comprises two key components: the asymmetrical fusion framework (AFF) and the asymmetrical attention block (AAB). To optimize efficiency by reducing parameters, AFF integrates one advanced model for feature extraction from dermoscopy images and one lightweight model for clinical images. This design is grounded in the assumption that affecting the ability to capture supplementary information from clinical images will subtly or not impact the overall multi-modal pipeline’s performance. In contrast to the bidirectional attention block (BAB), AAB focuses solely on enhancing dermoscopy features while excluding attention to clinical images due to our belief that directing attention to supplementary information may adversely impact the final classification performance. Extensive results demonstrate that, in comparison to the previous symmetrical fusion framework, AFF significantly reduces model parameters while maintaining accuracy. Additionally, AAB enhances the performance of BAB with fewer parameters, showcasing its efficacy in improving the overall classification task. Lastly, our AMMFM attains state-of-the-art performance with the fewest model parameters.

**Table 5.10:** Comparisons between single-modal, baseline multi-modal, and our AMMFM methods based on different backbones. ST: Swin-Transformer, Params: Parameters (%).

Backbone	Modal	AVG AUC	AVG ACC	Params(Mb)
ResNet	Derm	84.3±0.5	74.0±0.4	26.68
	Clic	76.8±0.5	67.5±0.4	
	Baseline	85.5±0.2	74.6±0.4	55.52
	<b>AMMFM</b>	<b>85.9±0.2</b>	<b>74.9±0.3</b>	36.36
ConvNext	Derm	86.7±0.4	75.9±0.3	29.05
	Clic	77.9±0.3	69.0±0.3	
	Baseline	87.2±0.3	76.5±0.3	58.96
	<b>AMMFM</b>	<b>87.4±0.5</b>	<b>76.7±0.4</b>	33.29
ST	Derm	86.8±0.3	76.2±0.6	28.82
	Clic	77.7±0.5	68.7±0.7	
	Baseline	87.2±0.4	76.6±0.5	58.49
	<b>AMMFM</b>	<b>87.6±0.4</b>	<b>76.7±0.4</b>	33.06

## 5.4 Single Shared Network with Prior-Inspired Loss for Efficient Multi-Modal Skin Lesion Classification

In this section, we propose a novel parameter-efficient multimodal (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 differences between the previous methods and our method. Firstly, unlike previous approaches that commonly employed ResNet as feature encoder, we conduct a comprehensive comparison between ResNet and more advanced backbones, i.e., DenseNet [52], ConvNext (CXT) [72], and SwinTransformer (ST) [70], 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. 5.4). Therefore, we explore and verify that multimodal features can be efficiently learned in a single-shared network with strong capacity by merely maintaining the modal-



**Figure 5.4:** The overview structure of former methods and our PEMM framework.

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 helpful 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 a few additional parameter increases on single-modal-based networks.

#### 5.4.1 Related works about parameter-sharing network

Parameter-sharing networks (PSNs) or weight-sharing networks (WSNs) 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 [53, 96, 112]. Additionally, some works utilize PSN to improve performance while achieving lower memory consumption [2, 118, 120]. For instance, [118] presented a parameter-sharing transformer block that captures scale-invariant information for 3D medical image segmentation. Similarly, [120] 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 multi-modal skin lesion classification, TFormer [136] employed a weight-sharing scheme to alleviate the overfitting problem. However, they needed to thoroughly explore the impact of weight-sharing schemes on reducing parameters, leading to a more precise conclusion. For instance, in their configuration, the parameters of the introduced fusion branch are nearly identical to those of the feature encoder. The weight-sharing scheme is likely 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 the encoder’s capacity and maintaining individual classifiers. We further explored its generalization ability across different backbones by conducting extensive experiments. Moreover, compared 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.

#### 5.4.2 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 5.18 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 5.5, into our Parameter-Efficient Multi-Modal (PEMM) framework.

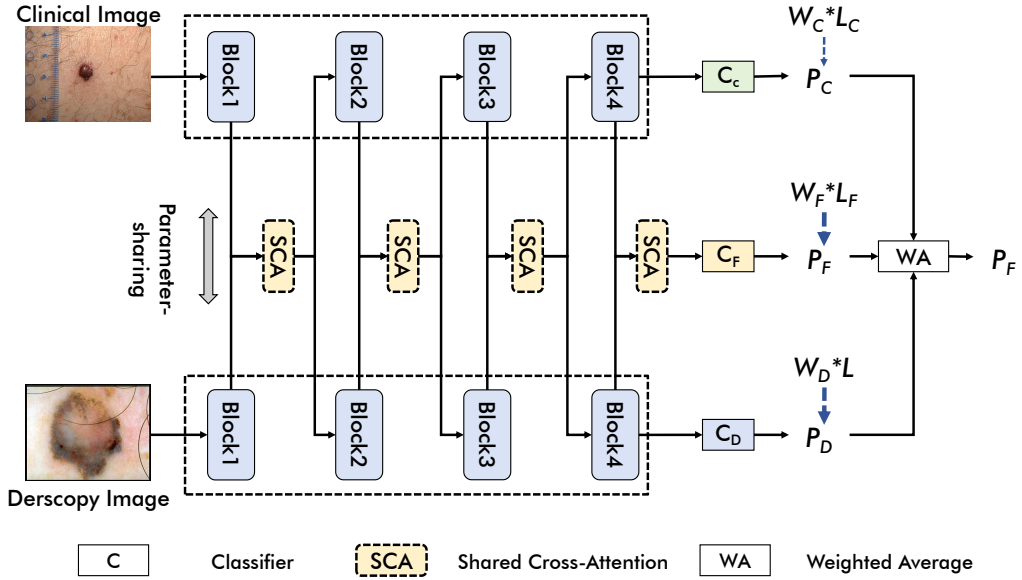


Figure 5.5: The detailed pipeline of our PEMM framework.

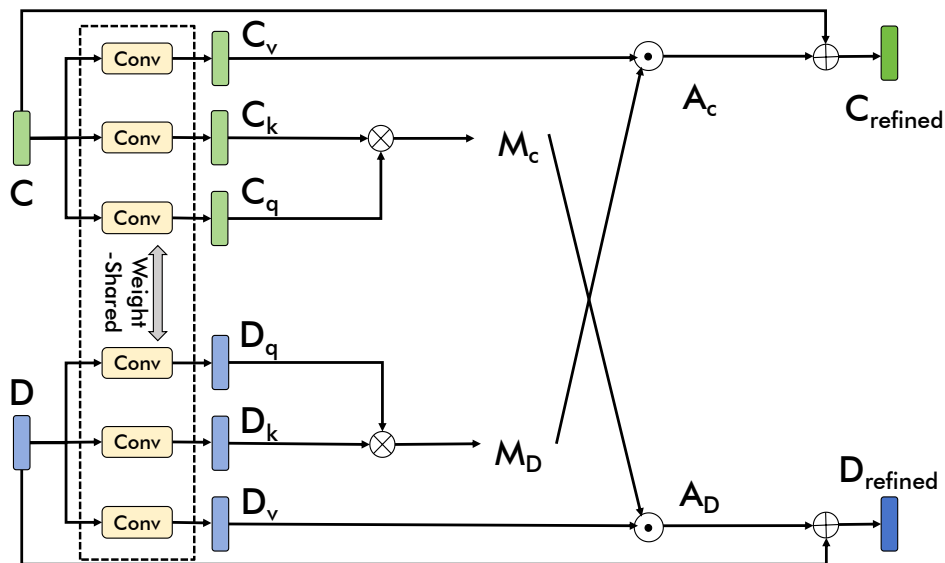
#### 5.4.2.1 Single-Shared Network

Following [110, 36, 46], 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 extract 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.5.4(b)), rather than using two individual encoders as commonly done in previous methods (See Fig.4.1(a)). More specifically, as depicted in Fig. 4.2, the Single-Shared Network (SSN) adopts weight-sharing encoders to extract multi-modal features. At the same time, individual classifiers are built upon fully connected layers to predict 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, attributed to the robustness of convolution layers and the sensitivity of fully connected layers (More details can be found in Table 5.16). 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.



**Figure 5.6:** The detailed pipeline of shared cross-attention module.

#### 5.4.2.2 Shared Cross-Attention Module

Following [110, 36, 46], we also see the effectiveness of the current advanced fusion module, i.e., cross-attention (CA) for multi-modal skin image fusion, has been demonstrated in [46]. As illustrated in Fig. 5.6, 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  $D_{refined}$ .

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. 5.6). Consequently, we can save half of the parameters of each CA module.

### 5.4.2.3 Biased Loss Function

In the training of previous methods, three branches are equally optimized, so their loss function can be formulated as Eq. 5.5:

$$L_{total} = (L_C + L_D + L_F)/3 \quad (5.5)$$

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. 5.6).

However, optimizing these three branches equally seems unreasonable based on prior knowledge, which demonstrated that the dermoscopy image-based model outperforms the clinical image-based model [29]. Inspired by the prior knowledge, we can have a hypothesis that dermoscopy information is more valuable than clinical one in the multi-modal task and thus proposed a new biased loss function, which is achieved by adjusting the corresponding weights of loss functions for different branches, as shown in Eq. 5.6.

$$L_{total} = W_C \cdot L_C + W_D \cdot L_D + W_F \cdot L_F \quad (5.6)$$

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  $W_F$  is the sum of  $W_C$  and  $W_D$  as the fusion information is the combination of clinical and dermoscopy information. So, Eq. 5.6 can be simplified into Eq. 5.7

$$L_{total} = W \cdot L_C + (0.5 - W) \cdot L_D + 0.5 \cdot L_F \quad (5.7)$$

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 the clinical branch.

### 5.4.3 Experiments and Discussion

#### 5.4.3.1 Implementation Details

During training, the Adam optimizer [62] 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 [56] is utilized to generate the final weight used for testing. All images are resized to  $224 \times 224 \times 3$  for both training and testing. During the testing, we followed [110] that searches the weights on the validation set and then forms the final predictions by a weighted averaging scheme. All the experiments are based on the backbone of SwinTransformer and are on the SPC dataset unless specified. The weight factor in Eq. 5.7 is set to 0.1, as it yields the best performance (See Table 5.17).

#### 5.4.3.2 Datasets and Metrics

The dataset and evaluation metrics are the same as that in Sec.5.3.4.2 and Table 5.1.

#### 5.4.3.3 Comparison with state-of-the-art methods

We undertake a comparative analysis of our PEMM model with several existing methodologies, including TFormer [136], GIIN [36], FusionM4Net-FS [110], AMFAM [119], HcCNN [11], and Inception-combination [60], on the SPC dataset. The comparative results concerning Averaged AUC and accuracy are presented in Tables 5.11 and 5.12, respectively. Notably, 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. Our model underwent training five times for the ensuing experiments, and the mean values alongside the standard deviation from these five iterations were employed for a more robust analysis of our model.

As demonstrated in Table 5.11, 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 following 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 almost all categories except

**Table 5.11:** 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 bolded 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	
Incep-com	92.9	89.7	86.3	88.3	<i>91</i>	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	<i>85.9</i>	78.3	87.8	77.6	<i>83.6</i>	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	<i>89.5</i>	80.7	<b>85.1</b>	83.4	<b>86.7</b>	77.7	81.9	91.1	<b>88.8</b>	80.9	85.7
FM-FS	<i>95.3</i>	92.6	89	<i>94</i>	89.2	<i>85.9</i>	<i>83.9</i>	87.9	81.4	80.9	83.5	81.7	<i>79.1</i>	80.1	90.6	87.8	78	86
GIIN	92.8	86.8	87.6	88.8	79.8	80.1	<b>87.5</b>	84.9	81.2	81.1	83.6	79	78.6	<i>83.1</i>	90.8	80.7	75.4	83.6
CAFNet	<b>97.1</b>	<i>92.7</i>	<b>92.2</b>	92.5	<i>91</i>	81.9	75.3	87.4	<b>85.4</b>	76.1	<i>85</i>	<i>85.4</i>	75.2	78.7	<b>94.7</b>	84.8	<i>83.5</i>	85.8
PEMM (Ours)	94.7	<b>93.0</b>	<i>90.8</i>	<b>94.9</b>	<b>91.7</b>	<b>86.7</b>	83.8	<b>90.1</b>	<i>84.4</i>	79.4	<b>86.1</b>	84.9	<b>80.7</b>	<b>84.0</b>	<i>93.9</i>	<i>88.5</i>	<b>85.4</b>	<b>87.6</b>

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 5.12, 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 % to 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.

**Table 5.12:** 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 bolded 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	<i>70.9</i>	87.1	79.7	66.1	74.2	60.0	77.2	74.2	73.7
HcCNN	70.6	87.1	<b>84.8</b>	68.6	71.6	<b>65.6</b>	80.8	69.9	74.9
AMFAM	70.6	<i>88.1</i>	83.3	70.9	74.7	63.8	<i>82.3</i>	75.4	76.0
FM-FS	<i>70.9</i>	86.8	81.8	<i>72.4</i>	74.4	61.0	<b>83.0</b>	74.9	75.7
CAFNet	70.1	87.8	<i>84.3</i>	<b>73.4</b>	<b>77.0</b>	61.5	<i>81.8</i>	<b>78.2</b>	<i>76.8</i>
TFormer	<i>70.9</i>	86.4	83.5	68.8	74.0	<i>64.9</i>	81.3	73	75.3
PEMM (ours)	<b>73.7</b>	<b>88.9</b>	82.5	71.9	<i>76.0</i>	<b>65.6</b>	<b>83.0</b>	<i>77.7</i>	<b>77.4</b>

For further analysis, we adhere to the methodology outlined in [11, 36] to present the results of melanoma-related features in Table 5.14. From this table, it is evident

**Table 5.13:** 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)	<b>87.6</b>	<b>77.4</b>	31.12M

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 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 to identify 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 5.13. 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

**Table 5.14:** Further comparison in melanoma-related features (%).

Metric	Method	DIAG	PN	STR	PIG	RS	DaG	BWV	VS	Avg
		MEL	ATP	IR	IR	PRS	IR	PRS	IR	
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	<b>86.7</b>
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	<b>70.3</b>
	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	<b>64.1</b>
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	<b>91.1</b>
	PEMM	88.5	87.1	85.5	84.2	84.5	80.6	93.7	93.4	87.2

(in terms of Avg AUC) and CAFNet (in terms of Avg ACC). This result substantiates the effectiveness of our method.

#### 5.4.3.4 Ablation studies

The ablation studies of our PEMM are shown in Table 5.15 to analyze the effect of three components, i.e., parameter-sharing (PS) encoder, shared cross-attention modules (SCA), and 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 both modalities and trained by equal optimization (See Eq. 5.5). From the data presented in the table,

**Table 5.15:** 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			87.0±0.3	76.5±0.6	58.49M
PS				86.9±0.4	76.4±0.4	30.14M
	✓			86.9±0.4	75.9±0.4	32.10M
		✓		87.2±0.1	76.6±0.3	31.12M
		✓	✓	87.3±0.3	76.8±0.7	31.12M

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 0.1% decrease in both metrics (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 87.2% in AUC value and from 76.4% 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, implementing biased loss (BL) further enhances the performance of the PS-SCA model to 87.3% 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 diagnostic accuracy with minimal or no increase in the model’s parameters. In our comparison between CA [46] 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 and resulting in poorer performance than concatenations, especially in smaller datasets. However, this issue can be mitigated by employing the PS scheme, akin to the phenomenon observed in [136].

#### 5.4.3.5 Other experiments

**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 5.16, compared to the Non-PS classifiers, the PS classifiers exhibit a significant reduction in AUC from 87.3% to 86.8% and ACC from 76.8% to 76.4%. This could be attributed to the sensitivity of fully connected layers to the input.

**Table 5.16:** Comparison between our model using parameter-sharing (PS) and non-PS classifiers. (%)

classifiers	AVG AUC	AVG ACC	Parameters(Mb)
PS	86.8±0.5	76.4±0.4	30.65
Non-PS	87.3±0.3	76.8±0.4	31.12

**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. 5.7). 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 5.17, 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 and 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.

**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 5.18, the proposed PEMM decreases diagnostic performance compared to the Baseline model when utilizing ResNet50 and DenseNet201 as the parameter-sharing encoder, specifically for ResNet50 and DenseNet201, the 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%



**Table 5.17:** The effect of different weight factor  $W$  in Eq. 4.3. EQ: Equally optimization that indicates the model is optimized by the loss function as shown in Eq. 5.5. (%)

W	AVG AUC	AVG ACC
0	<b>87.4±0.4</b>	76.6±0.5
0.1	87.3±0.3	<b>76.84±0.4</b>
0.2	87.3±0.4	76.78±0.1
0.3	87.2±0.2	76.3±0.4
0.4	86.7±0.2	76.1±0.3
0.5	84.6±0.2	73.9±0.4
EQ	87.2±0.2	76.6±0.3

vs. Derm: 74.3% for ResNet50; PEMM: 74.6% vs. Derm: 75.1% 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. This discrepancy may be attributed to differences in the capacity of the backbones. More advanced backbones possess parameter-sharing capacity, while traditional backbones may lack this capability.

#### 5.4.4 Conclusion


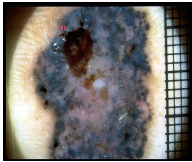
In this paper, we introduce a novel Parameter-Efficient Multi-Modal (PEMM) method for skin lesion classification. Our approach offers several key contributions: Firstly, by sharing the parameters of encoders with strong capacity while retaining individual classifiers, PEMM achieves approximately 50% 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 utilizing fewer parameters on the SPC dataset.

**Table 5.18:** 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	84.7±0.3	74.3±0.4	26.68
	Clic	76.9±0.3	67.6±0.3	
	Baseline	85.4±0.1	74.3±0.2	
	PEMM	84.7±0.1	73.3±0.6	
DenseNet201	Derm	85.3±0.1	75.1±0.1	20.17
	Clic	77.2±0.2	68.5±0.4	
	Baseline	86.5±0.4	75.7±0.5	
	PEMM	86.0±0.2	74.6±0.4	
Convnext	Derm	86.6±0.5	75.9±0.2	29.05
	Clic	78.1±0.2	69.1±0.2	
	Baseline	87.2±0.3	76.6±0.5	
	PEMM	87.2±0.2	76.4±0.5	
ST	Derm	86.9±0.3	76.3±0.4	28.82
	Clic	77.7±0.5	69.0±0.7	
	Baseline	87.2±0.3	76.5±0.6	
	PEMM	87.3±0.2	76.8±0.4	

# Fusing clinical images, dermoscopy images and patient's metadata for skin lesion classification

## 6.1 Introduction

		Clinical Image	Dermoscopy Image	Patient's Meta-Data	
		 ID:NEL035	 ID:NEL034	<b>Diagnostic Difficulty</b>	Low
				<b>Elevation</b>	Palpable
				<b>Sex</b>	Male
				<b>Management</b>	Excision
				<b>Location</b>	Lower Limbs
Seven - Point Checklist Label	<b>Pigment Network</b>	Absent			
	<b>Streak</b>	Absent			
	<b>Pigment</b>	Absent			
	<b>Regression Structures</b>	Present			
	<b>Dots and Globules</b>	Irregular			
	<b>Vascular Structures</b>	Present			
	<b>Blue whitish veil</b>	Present			
	<b>Diagnosis Label</b>	Melanoma			

**Figure 6.1:** Example of one patient's case, including the clinical image, dermoscopy image, meta-data, seven-points checklist criteria label, and diagnostic label.

In the last two chapters, we introduced a single image-meta fusion method and a clinical-dermoscopy fusion method, respectively. The experiments in these two

chapters demonstrate the superiority of two modal-based methods compared to a single modal-based one. However, how to efficiently fuse them for [Skin Lesion Classification \(SLC\)](#)? When we get all of these three modalities. In this chapter, we will answer the question. Before delving into our proposed fusion methods, let’s review the previous works on the fusion of three modalities for skin lesion classification. The literature on three-modalities-based SLC is limited; most existing works [36, 119, 11, 46] focus only on the fusion of two modality images. Therefore, we include this review in this section rather than creating a separate section.

Prior to the release of the [Seven-Point Checklist \(SPC\)](#) dataset by [60], there was only one study [129] that proposed the use of ResNet-50 [45] to extract semantic features from clinical and dermoscopy images, respectively. These features were then integrated with metadata encoded in a one-hot manner for skin disease diagnosis. [60] introduced the first publicly available multi-modal dataset for multi-label SLC. They proposed a unified framework that utilized two InceptionV3 [106] models to process multi-modality data for multi-label SLC (SPC criteria label and diagnosis label).

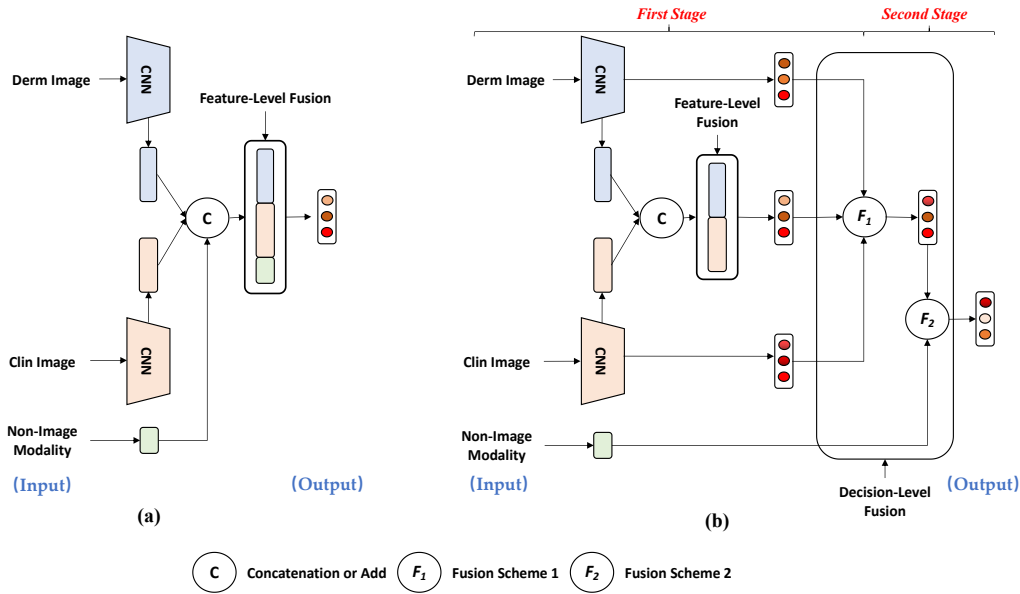
There are two main limitations to the methods mentioned above. Firstly, learning both shareable and individual representations has been proven to be important in many multi-modal fusion applications [50, 133]. Additionally, decision-level fusion approaches have also been shown to be preferable to avoid the overfitting problem of machine learning models when the training dataset is not large [88, 54, 103]. However, the above-mentioned methods only focus on training a shareable feature representation and ignore decision-level fusion.

Secondly, the patient’s meta-data (non-image modality data) is typically only embedded into the fully connected layer and is used to co-train the whole deep learning model with two-modality images at one stage, as shown in Fig. 6.2(a). Consequently, the entire model’s weight is heavily biased towards feature vectors extracted from multi-modality images. The impact of non-image modality data on the final prediction is minimal or even ignored.

### 6.1.1 Our Works: FusionM4Net

To address the limitations mentioned above, in this work, we propose a multi-stage, multi-modal learning algorithm (FusionM4Net), which gradually fuses two-modality images and patient’s meta-data information at different levels. FusionM4Net consists of two stages, as illustrated in Fig. 4.3.

In the first stage, our proposed algorithm utilizes a FusionNet model as the feature extractor, which employs two convolutional neural networks to learn both



**Figure 6.2:** The overview of different multi-modal CNN architectures for skin disease recognition. (a) one-stage multi-modal CNN, (b) our proposed Fusion-M4Net. In this figure,  $F_1$  is used to fuse information from two-modality images at the decision level in the first stage.  $F_2$  is adopted to integrate non-image modality data and image modality data in the second stage. (Abbreviations: Clin image: clinical image; Derm image: dermoscopy image).

single-modality specific representation and cross-modality common representation at the feature level. Subsequently, our algorithm employs Fusion Scheme 1 to fuse the predictions from the two CNN models at the decision level.

In the second stage, we introduce Fusion Scheme 2 to incorporate the patient’s meta-data information. With the decision information from the first stage, the patient’s meta-data information is first utilized to co-optimize an SVM cluster. Then, the decision from the SVM cluster and from the first stage is integrated into the final multi-label SLC result. We evaluate our multi-modal learning algorithm for a multi-label SLC task on a publicly available dataset [60]. In summary, the proposed FusionM4Net achieves significant improvements compared to single-modal algorithms and outperforms other multi-modal, multi-label SLC methods, as we will demonstrate in this work.

The contributions of this work are summarized into three points:

1. We build a FusionNet model and Fusion Scheme 1 to effectively learn and fuse

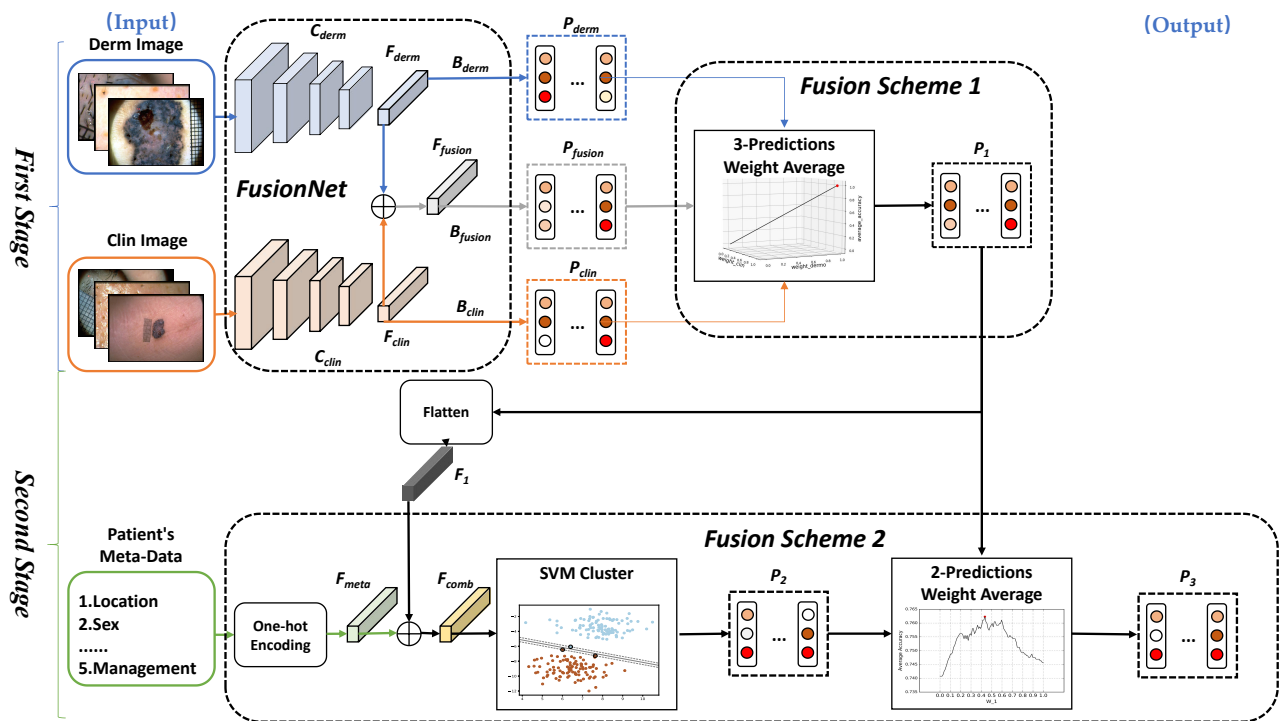


Figure 6.3: The flowchart of our proposed FusionM4Net algorithm.

the representation of two-modality images at both the feature level and decision level.

2. We propose a new two-stage multi-modal learning algorithm named FusionM4Net, which gradually integrates the discriminative information from this three-modality data in a hierarchical manner.
3. We conduct extensive experiments to evaluate our algorithm, the results of which show the advantage of our proposed method.

The code of our algorithm is publicly available at <sup>1</sup>

<sup>1</sup><https://github.com/pixixiaonaogou/MLSDR>

**Table 6.1:** The detailed data structure of the final prediction  $P_3$  in FusionM4Net, see Fig.3.  $P_3^i$  indicates  $i$ th prediction in  $P_3$ . pred type: the prediction for the type of label.

	pred type	shape
$P_3^1$	<i>Diag</i>	(5)
$P_3^2$	<i>Bwv</i>	(2)
$P_3^3$	<i>Str</i>	(3)
$P_3^4$	<i>Pig</i>	(3)
$P_3^5$	<i>Rs</i>	(2)
$P_3^6$	<i>Dag</i>	(3)
$P_3^7$	<i>Pn</i>	(3)
$P_3^8$	<i>Vs</i>	(3)

## 6.2 Methods

### 6.2.1 Overall Structure of FusionM4Net

Fig. 6.3 shows the flowchart of our proposed FusionM4Net, which consists of two main stages and gradually fuses clinical images, dermoscopy images, and patient’s meta-data for multi-label SLC. In this flowchart, we input  $x_{clin} \in \mathbb{R}^{W \times H \times C}$ ,  $x_{derm} \in \mathbb{R}^{W \times H \times C}$  and  $x_{meta} \in \mathbb{R}^L$  into our FusionM4Net, and obtain the final output  $P_3$ . Here,  $W$ ,  $H$ , and  $C$  indicate the image’s width, height, and channel number, respectively.  $L$  is the length of one-hot encoded features of the patient’s metadata.  $P_3$  is a list type and consists of eight predictions, as shown in Table 6.1. Other predictions in Fig. 6.3, including  $P_{derm}$ ,  $P_{clin}$ ,  $P_{fusion}$ ,  $P_1$ , and  $P_2$ , and the ground truth  $y$  have the same data structure as  $P_3$ .

In the following sub-sections, we will introduce the three main parts of our method, FusionNet, Fusion Scheme 1, and Fusion Scheme 2, sequentially. FusionNet and Fusion Scheme 1 belong to the first stage of FusionM4Net and are used to fuse the two-modality images. Fusion Scheme 2 belongs to the second stage of FusionM4Net and is adopted to integrate the patient’s metadata.

### 6.2.2 FusionNet

FusionNet is employed to integrate and learn the information from two-modality images at the feature level. The proposed FusionNet contains two ResNet-50 backbones and three branches. The two ResNet-50 backbones are adopted to extract features from clinical images and dermoscopy images. We denote these two ResNet-50 backbones as  $C_{clin}$  and  $C_{derm}$ , and the three branches  $B_{clin}$ ,  $B_{derm}$  and  $B_{fusion}$ , which

**Table 6.2:** The detailed structure of branch  $B_{clin}$  in FusionNet (which is also indicated to  $B_{derm}$  and  $B_{fusion}$ ), see Fig. 3.  $B_{clin}^i$  indicates  $i$ th multi-layer FCL in  $B_{clin}$ . FCL type: the FCL for the corresponding label of the prediction.

	FCL type	configuration
$B_{clin}^1$	<i>Diag</i>	(input_size=2048,output_size=5)
$B_{clin}^2$	<i>Bwv</i>	(input_size=2048,output_size=2)
$B_{clin}^3$	<i>Str</i>	(input_size=2048,output_size=3)
$B_{clin}^4$	<i>Pig</i>	(input_size=2048,output_size=3)
$B_{clin}^5$	<i>Rs</i>	(input_size=2048,output_size=2)
$B_{clin}^6$	<i>Dag</i>	(input_size=2048,output_size=3)
$B_{clin}^7$	<i>Pn</i>	(input_size=2048,output_size=3)
$B_{clin}^8$	<i>Vs</i>	(input_size=2048,output_size=3)

are adopted for optimizing  $P_{clin}$  and  $P_{derm}$ ,  $P_{fusion}$ , respectively. The detail of the ResNet-50 structure can be seen in [he2016], and the structure of  $B_{clin}$ ,  $B_{derm}$  and  $B_{fusion}$  are the same, so we only use  $B_{clin}$  for descriptions.  $B_{clin}$  contains eight fully connected layers (FCL) for obtaining eight predictions, and its structure is as shown in Table 6.2.

During the training,  $x_{clin}$  and  $x_{derm}$  are fed into the ResNet-50s  $C_{clin}$  and  $C_{derm}$  to get  $F_{clin} \in \mathbb{R}^N$  and  $F_{derm} \in \mathbb{R}^N$ , respectively. Then,  $F_{clin}$  and  $F_{derm}$  are combined by element-wise addition to constructing a shared feature vector  $F_{fusion} \in \mathbb{R}^N$ , where  $N$  is the length of the feature vector. Next, three feature vectors  $F_{clin}$ ,  $F_{derm}$  and  $F_{fusion}$  are input into the FCLs  $B_{clin}$ ,  $B_{derm}$  and  $B_{fusion}$  to obtain  $P_{clin}$ ,  $P_{derm}$  and  $P_{fusion}$ , respectively, in which  $B_{clin}$  and  $B_{derm}$  are employed for learning single-modality feature representation, while  $B_{fusion}$  is used for learning cross-modality feature representation. Our FusionNet is co-trained by two-modality images and their corresponding labels in pairs to minimize the overall cross-entropy loss between multi-label predictions and ground truths. The loss function that is used to optimize the three predictions  $P_{derm}$ ,  $P_{clin}$ , and  $P_{fusion}$  simultaneously can be defined as follows:

$$L_{clin} = \sum_{i=1}^N \text{CE}(P_{clin}^i, y^i) \quad (6.1)$$

$$L_{derm} = \sum_{i=1}^N \text{CE}(P_{derm}^i, y^i) \quad (6.2)$$

$$L_{fusion} = \sum_{i=1}^N \text{CE}(P_{fusion}^i, y^i) \quad (6.3)$$



$$L_{overall} = L_{clin} + L_{derm} + L_{fusion} \quad (6.4)$$

where  $L_{clin}$ ,  $L_{derm}$  and  $L_{fusion}$  indicate the loss function of  $P_{clin}$ ,  $P_{derm}$ , and  $P_{fusion}$ , respectively.  $y$  is the multi-label ground truth and has the same data structure as  $P_3$ , as shown in Table 6.1.  $y^i$  is the  $i$ th label in  $y$ .  $CE()$  is the cross-entropy loss function. During our training process, the above three loss functions combine to an overall loss function  $L_{overall}$  to train the clinical branch, dermoscopy branch, and fusion branch together. With the co-optimization scheme, our FusionNet can capture single-modality information and learn cross-modality feature representation concurrently.

---

**Algorithm 1** 3-predictions weights searching algorithm

---

**Input:** The predictions  $P_{derm}$ ,  $P_{clin}$ ,  $P_{fusion}$  and corresponding ground truth  $y$  on validation dataset

**Output:** The weights ( $W_{clin}$ ,  $W_{derm}$  and  $W_{fusion}$ ) with best accuracy on validation dataset

```

1: Initialize  $ACC_{best} = 0$  and array  $B_1$  (50 equally spaced entries).
2: for  $(i, j, z) \in B_1 \times B_1 \times B_1$  do
3:    $i_{norm} = \frac{i}{i+j+z}$ ,  $j_{norm} = \frac{j}{i+j+z}$ ,  $z_{norm} = \frac{z}{i+j+z}$ 
4:   for each  $i \in [1, 2, 3, 4, 5, 6, 7, 8]$  do
5:      $P_{temp}^i = i_{norm} \times P_{derm}^i + j_{norm} \times P_{clin}^i + z_{norm} \times P_{fusion}^i$ 
6:   end for
7:    $P_{temp} = [P_{temp}^1, P_{temp}^2, P_{temp}^3, P_{temp}^4, P_{temp}^5, P_{temp}^6, P_{temp}^7, P_{temp}^8]$ 
8:    $ACC_{temp} = \text{Compare}(P_{temp}, y)$ 
9:   if  $ACC_{temp} > ACC_{best}$  then
10:     $ACC_{best} \leftarrow ACC_{temp}$ 
11:     $W_{derm} \leftarrow i$ .
12:     $W_{clin} \leftarrow j$ ;
13:     $W_{fusion} \leftarrow 1 - i - j$ ;
14:   end if
15: end for
return  $W_{derm}, W_{clin}, W_{fusion}$ 

```

---

### 6.2.3 Fusion Scheme 1

Fusion Scheme 1 effectively fuses the output information from the three branches of FusionNet at the decision level, as shown in Fig. 6.3. Specifically, Fusion Scheme 1 is

a 3-predictions weighted average scheme, which integrates  $P_{derm}$ ,  $P_{clin}$  and  $P_{fusion}$  to form  $P_1$ . It mainly comprises two steps: weights searching and prediction fusing.

First, a 3-predictions weights searching algorithm is employed to search for the weights, making the fused prediction  $P_1$  get the best average accuracy on the validation dataset. The pseudo-code of the 3-predictions weights searching algorithm is shown in Algorithm 1.

Secondly, according to eq. 4.5, we can calculate  $P_1 = [P_1^1, P_1^2, P_1^3, P_1^4, P_1^5, P_1^6, P_1^7, P_1^8]$  as follows:

$$P_1^i = W_{clin} * P_{clin}^i + W_{derm} * P_{derm}^i + W_{fusion} * P_{fusion}^i, \quad i \in [1, 2, 3, 4, 5, 6, 7, 8], \quad (6.5)$$

where  $W_{clin}$ ,  $W_{derm}$  and  $W_{fusion}$  are the search weights from Algorithm 1.

## 6.2.4 Fusion Scheme 2

Fusion Scheme 2 further incorporates patient’s meta-data with decision information from two-modality image data. There is only one main component in the second stage: Fusion Scheme 2, which consists of two key steps.

**Step 1:** First, we use the one-hot function from Keras [23] to encode the patient’s meta-data into a feature vector  $F_{meta} \in \mathbb{R}^{N_{meta}}$ , as shown in Fig. 4.4. Next, we concatenate the meta-data vector and the flattened normalized multi-label predictive vector  $F_1 \in \mathbb{R}^{N_1}$  to construct a combined feature vector  $F_{comb} \in \mathbb{R}^{N_{comb}}$ , which is subsequently fed into the SVM cluster. Here  $N_{meta}$ ,  $N_1$  and  $N_{comb}$  equals 20, 24, and 44, which are the lengths of  $F_{meta}$ ,  $F_1$ , and  $F_{comb}$ , respectively.

The SVM cluster is comprised of eight SVMs, which are trained for eight-label classification respectively. The eight predictions generated by all the SVMs are denoted as  $P_2$ . Note that we do a performance comparison using different classifier-based clusters in Table 6.7, including logistic regression, SVM, and multi-layer perceptron, and the SVM cluster performs best.

**Step 2:** We adopt a decision-level fusion based on eq. (6.6), which uses a 2-predictions weighted average scheme to fuse the prediction  $P_1$  at the first stage and the prediction  $P_2$  of step 1 at the second stage to form the final prediction  $P_3 = [P_3^1, P_3^2, P_3^3, P_3^4, P_3^5, P_3^6, P_3^7, P_3^8]$ .

$$P_3^i = P_1^i * W_1 + P_2^i * W_2, \quad i \in [1, 2, 3, 4, 5, 6, 7, 8], \quad (6.6)$$

where  $W_1$  and  $W_2$  are the search weights in Algorithm 2. The pseudo-code of the 2-predictions weights searching algorithm is shown in Algorithm 2.

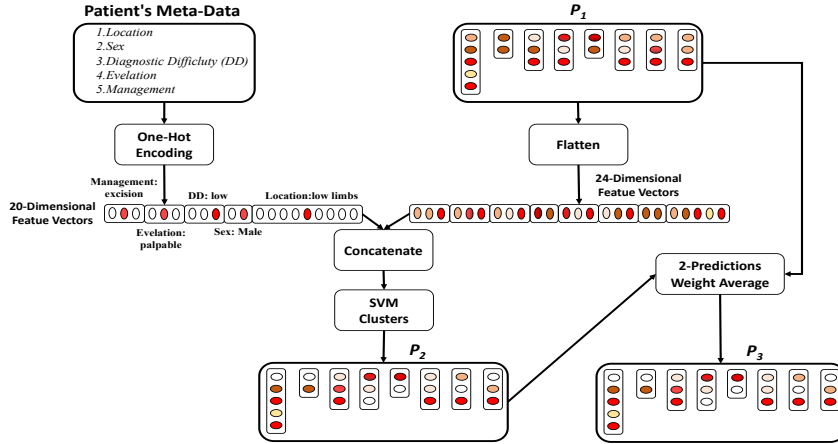


Figure 6.4: The description of the second stage of Fusion-M4Net.

## 6.3 Experiments and Results

In this section, we will first introduce the seven-point checklist dataset [60], which is composed of multi-modality data and eight labels.

Then, to investigate the performance of our FusionM4Net, we will display the performance comparison between single- and cross-modality in Table 6.3, performance comparison of all the predictions of FusionM4Net in Table 6.4. To explore how to select weights, clusters and meta-fusion schemes, we will show the classification results of the 2-predictions weighted average in Table 6.5, the classification results of 3-predictions weighted average in Table 4.8, the classification results with the different classifier-based clusters in Table 6.7, and the classification results with different meta-fusion schemes in Table 6.8.

Finally, the performance comparison with other current advanced multi-label SDR algorithms is shown in Table 6.9 and Table 6.10.

### 6.3.1 Dataset

Our method was evaluated on the Seven-Point Checklist (*SPC*) dataset [60]. The *SPC* dataset contains 413 training cases, 203 validation cases, and 395 testing cases. Each case comprises a dermoscopy image, a clinical image, the *SPC* labels, and a diagnostic label, as seen in Fig. 6.1. The *SPC* labels are (I) Pigment Network ( $P_n$ ), (II) Streak ( $Str$ ), (III) Pigment ( $Pig$ ), (IV) Regression Structures ( $Rs$ ), (V) Dots and Globules ( $Dag$ ), (VI) Vascular Structures ( $VII$ ) and (7) black whitish Veil ( $Bwv$ ). Each label has different types, including (1) Present (Pre), (2) Absent (Asb), (3)

---

**Algorithm 2** 2-predictions weights searching algorithm

---

**Input:** The predictions  $P_1^i$  and  $P_2^i$  and corresponding ground truth  $y$  on validation dataset

**Output:** The weights ( $W_1$  and  $W_2$ ) with best accuracy on validation dataset

```
1: Initialize  $ACC_{best} \leftarrow 0$  and the array  $B_2$  (100 equally spaced entries).
2: for each  $i \in B_2$  do
3:   for each  $i \in [1, 2, 3, 4, 5, 6, 7, 8]$  do
4:      $P_{temp}^i = i \times P_1^i + (1 - i) \times P_2^i$ 
5:   end for
6:    $P_{temp} = [P_{temp}^1, P_{temp}^2, P_{temp}^3, P_{temp}^4, P_{temp}^5, P_{temp}^6, P_{temp}^7, P_{temp}^8]$ 
7:    $ACC_{temp} = \text{Compare}(P_{temp}, GT)$ 
8:   if  $ACC_{temp} > ACC_{best}$  then
9:      $ACC_{best} \leftarrow ACC_{temp}$ 
10:     $W_1 \leftarrow i$ .
11:     $W_2 \leftarrow 1 - j$ ;
12:   end if
13: end for
return  $W_1, W_2$ 
```

---

Typical (Typ), (4) Atypical (Atp), (5) Regular (Reg), and (6) Irregular (Irg). The diagnosis (*Diag*) label is divided into five types: (1) Melanoma (Mel), (2) Nevus (Nev), (3) Seborrheic Keratosis (SK), (4) Basal Cell Carcinoma (Bcc), and (5) Miscellaneous (Misc). The author of [60] defined the dividing standard of SPC and diagnosis types. For example, arborizing, comma, hairpin, within regression, and wreath are divided into the type of Reg in the label of  $Vs$ . More details of the dividing standard can be seen in the Table I of [58].

### 6.3.2 Evaluated Metrics

Following several publications [60, 136], the evaluation metrics include average accuracy, the area under the receiver operating characteristic curve (AUC), sensitivity (SEN), specificity (SPE), and precision (PRE), to compare the performance of our FusionM4Net with other multi-modal learning methods. The average accuracy (AVG) is the main metric for the comparison.

### 6.3.3 Implementation Details

In our FusionM4Net, the FusionNet and the SVM cluster are independently trained.

The FusionNet is co-trained by  $x_{clin}$  and  $x_{derm}$  images in pairs and corresponding ground truth  $y$ . During the training of FusionNet, the batch size is 32, and the optimizer is Adam [62] with the initial learning rate (LR)  $3e-5$ . The size of the input image is  $224 \times 224 \times 3$ , which are the values of  $W$ ,  $H$ , and  $C$  respectively. The length  $L$  of the encoded patient’s metadata feature equals 20.

The total epochs of training are 250, and the LR changes with a CosineAnnealingLR schedule in PyTorch. The training stops after 250 epochs and the weight with the best accuracy on the validation dataset is saved for testing. The two CNN backbones are initialized by the pre-trained weights of ImageNet [30]. Data augmentations, including random vertical and horizontal flips, shifts, and distortions, are conducted to expand the dataset. Before training and testing, all images are resized to  $224 \times 224 \times 3$ , which is the default image size of ResNet-50. Since the dataset has been divided into the training images and validation images by the creator, we do not need to divide them again.

In the second stage, we divide the training dataset into two equally sized parts: sub-training and sub-testing dataset. During the SVM cluster training, the FusionNet is first trained by a sub-training dataset and then adopted to obtain the  $P_1$  from a sub-testing dataset. Next, the encoded  $F_1$  and  $F_{meta}$  from the sub-testing dataset are concatenated to form  $F_{comb}$ , which is used with the corresponding label  $y$  to optimize the weights of the SVM-cluster. The parameters and configuration of the SVM cluster’s training are the default ones in Sklearn, except the kernel, which is set to rbf. In Algorithm 1, we didn’t use an array with 100 components because it needs to take about 120 hours to search for the weights per model. In this case, we used the array  $B_1$  with 50 components as the candidate weight list; it just needs to take about 1 hour for searching. All the models are trained five times independently to get the mean value and standard deviation in our experiments, except for the results of Table 6.9, Table 6.10 and Fig. 6.7, which are from a single training run.

The workflow of model building, training, testing, result analysis, and plotting is constructed using several python libraries, mainly including Keras [23] (one-hot encoding), PyTorch [85] (building the FusionNet), Numpy [44] (array computation), Sklearn [86] (building the SVM-cluster and providing the metric functions), and albumentation [18] (data augmentation).

## 6.4 Experiments and Results

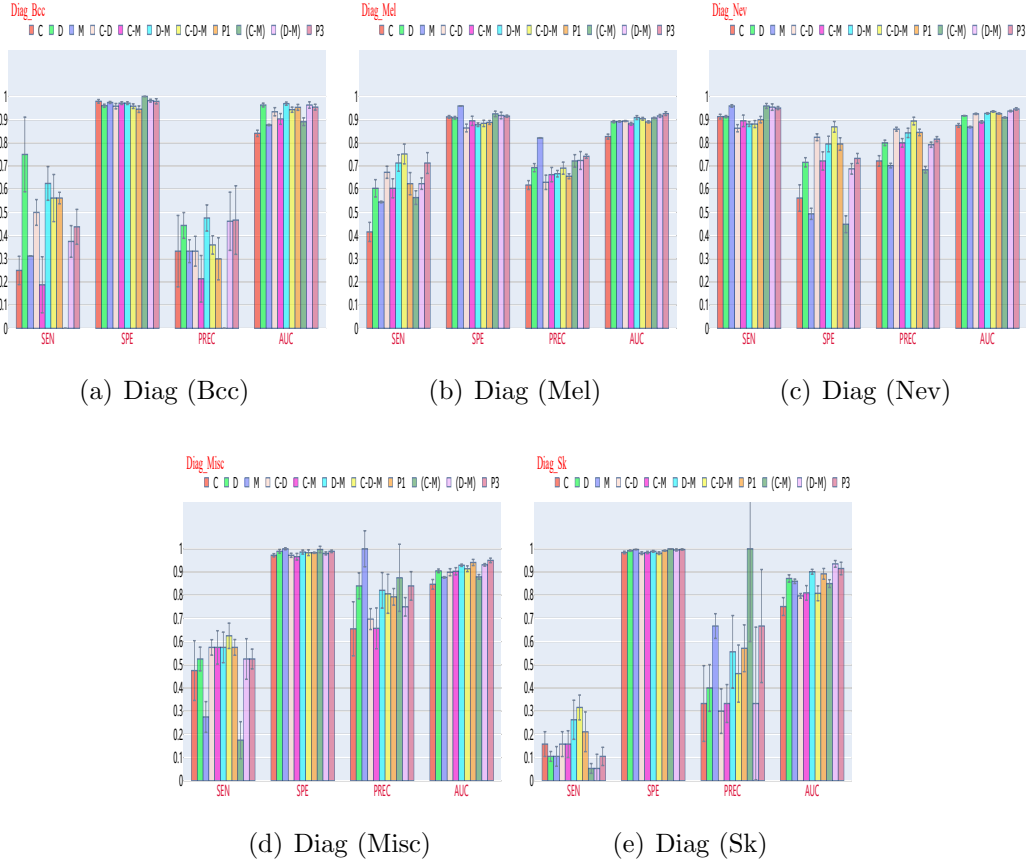
In this section, we will first introduce the seven-point checklist dataset [60], which is composed of multi-modality data and eight labels.

**Table 6.3:** The classification results of single-modality and multi-modality (average accuracy (AVG) in %), see Sec. 3.4.1 and Sec. 4.1.1. C, D, and M are the predictions of the models trained by clinical images, dermoscopy images, and patient meta-data, respectively. C-D is the prediction of the model, as shown in Fig.2(a), trained by clinical and dermoscopy images. C-M is the prediction of the model trained by clinical images and patient metadata. D-M is the prediction of the model trained by dermoscopy images and patient metadata. C-D-M is the prediction of the model trained by clinical and dermoscopy images and patient meta-data at one stage, as shown in Fig. 2(a). For more details, you can see Fig. 3 of the supplementary material.  $P_1$  and  $P_3$  are the predictions from our method.  $P_1$  is the counterpart of C-D and  $P_3$  is the counterpart of C-D-M. The flowcharts of all the models in this table are shown in the supplementary material to clarify the difference. The bold black number highlights the highest AVG accuracy in each column. *average* is the average value of all the labels in the other column.

	Model	<i>Buv</i>	<i>Dag</i>	<i>Pig</i>	<i>Pn</i>	<i>Rs</i>	<i>Str</i>	<i>Vs</i>	<i>Diag</i>	<i>average</i>
Single-Modality	C	83.7±0.8	53.3±2.0	59.3±1.4	57.5±1.9	74.5±1.0	66.3±1.0	80.8±0.2	67.0±1.5	67.8±0.5
	D	87.2±0.4	60.0±1.5	68.3±1.0	69.0±1.0	80.1±1.6	73.7±0.9	81.4±0.3	74.7±1.2	74.3±0.5
	M	84.6±0.0	59.9±0.1	59.2±0.2	61.1±0.2	72.5±0.2	73.1±0.1	79.2±0.0	72.4±0.8	70.2±0.1
Multi-Modality (One-stage)	C-D	86.9±0.7	60.8±1.9	69.6±0.6	67.5±1.1	78.8±1.0	72.6±1.0	81.2±0.7	75.4±1.0	74.1±0.2
	C-M	84.1±0.1	59.1±0.8	59.0±0.9	62.9±0.8	74.4±0.9	69.8±0.8	79.4±0.8	72.7±0.8	70.2±0.3
	D-M	87.7±1.0	63.2±0.7	69.9±1.0	<b>69.4±1.0</b>	80.8±0.7	74.2±0.8	81.5±1.1	76.2±1.5	75.4±0.4
	C-D-M	87.1±0.5	62.3±1.2	69.9±0.5	69.0±1.3	78.9±1.3	73.1±1.0	80.6±0.9	76.6±1.7	74.7±0.6
Multi-Modality (Ours, Two-stage)	P1 (C-D)	87.9±0.7	60.2±0.5	<b>71.5±0.7</b>	67.6±1.8	<b>81.9±1.0</b>	73.8±0.6	<b>82.1±0.5</b>	75.6±1.2	75.1±0.4
	(C-M)	84.8±0.6	62.1±0.6	60.2±1.0	64.7±1.3	73.9±0.3	71.9±0.7	79.3±0.1	70.5±0.8	70.9±0.3
	(D-M)	87.6±0.4	<b>66.2±1.9</b>	69.7±1.8	69.3±1.5	80.2±0.6	75.0±0.8	80.6±0.4	76.6±1.6	75.6±0.5
	P3 (C-D-M)	<b>88.5±0.4</b>	64.4±1.3	71.3±1.3	69.2±1.5	81.4±0.8	<b>76.1±1.1</b>	81.6±0.5	<b>77.6±1.5</b>	<b>76.3±0.7</b>

Then, to investigate the performance of our FusionM4Net, we will display the performance comparison between single- and cross-modality in Table 6.3, performance comparison of all the predictions of FusionM4Net in Table 6.4. To explore how to select weights, clusters and meta-fusion schemes, we will show the classification results of the 2-predictions weighted average in Table 6.5, the classification results of 3-predictions weighted average in Table 6.6, the classification results with the different classifier-based clusters in Table 6.7, and the classification results with different meta-fusion schemes in Table 6.8.

Finally, the performance comparison with other current advanced multi-label SDR algorithms is shown in Table 6.9 and Table 6.10.



**Figure 6.5:** The detailed metrics of *Diag* labels of the results in Table. 6.3.

## 6.4.1 Evaluation of FusionM4Net

### 6.4.1.1 Comparisons between single- and multi-modality learning

In Table 6.3, we show the classification results of different models that are trained by single-modality (clinical images, dermoscopy images, or patient’s meta-data) or multi-modality data. Fig. 6.5 displays the detailed metric of *Diag* label of the results in Table 6.3. These results are to illustrate the improvement of multi-modal learning (MML) for SLC compared with single-modal learning and our two-stage MML algorithm’s advantage compared with the one-stage MML algorithm.

In this experiment, we first adopt ResNet-50 as the model for the training of clinical images and dermoscopy images to get the predictions C and D, respectively, and the SVM cluster for the training of patient metadata to obtain the prediction M.

**Table 6.4:** The classification results comparison of the stage of our FusionM4Net (*AVG* accuracy in %). The black bold number is the highest value for each column. *average* is the average value of all the labels in the other column.

	<i>Bwv</i>	<i>Dag</i>	<i>Pig</i>	<i>Pn</i>	<i>Rs</i>	<i>Str</i>	<i>Vs</i>	<i>Diag</i>	<i>average</i>
$P_{clin}$	84.8±0.7	53.3±1.5	56.9±2.8	57.8±0.6	73.8±0.7	67.7±1.1	78.8±0.6	68.2±0.6	67.7±0.3
$P_{derm}$	87.6±0.7	59.4±1.6	70.8±0.6	66.6±1.6	80.9±1.0	74.0±0.7	81.1±0.7	74.6±0.7	74.4±0.4
$P_{fusion}$	87.0±0.6	60.4±1.2	70.2±0.9	67.0±1.6	80.5±1.6	74.0±0.6	81.8±0.5	75.7±1.1	74.6±0.5
$P_1$	87.9±0.7	60.2±0.5	<b>71.5±0.7</b>	67.6±1.8	<b>81.9±1.0</b>	73.8±0.6	<b>82.1±0.5</b>	75.6±1.2	75.1±0.4
$P_2$	86.7±0.6	<b>65.4±0.4</b>	65.7±1.1	<b>70.4±1.2</b>	79.0±0.9	75.8±0.3	79.2±0.0	75.4±0.7	74.7±0.2
$P_3$	<b>88.5±0.4</b>	64.4±1.3	71.3±1.3	69.2±1.5	81.4±0.8	<b>76.1±1.1</b>	81.6±0.5	<b>77.6±1.5</b>	<b>76.3±0.7</b>

**Table 6.5:** The classification results of 3-predictions weighted average scheme (*AVG* accuracy in %),  $P_1$  (averaged) is the prediction that is obtained by averaging the  $P_{clin}$ ,  $P_{derm}$  and  $P_{fusion}$ , while  $P_1$  (ours) is obtained by our Fusion Scheme 1. *average* is the average value of all the labels in the other column. The black bold number is the highest value for each column.

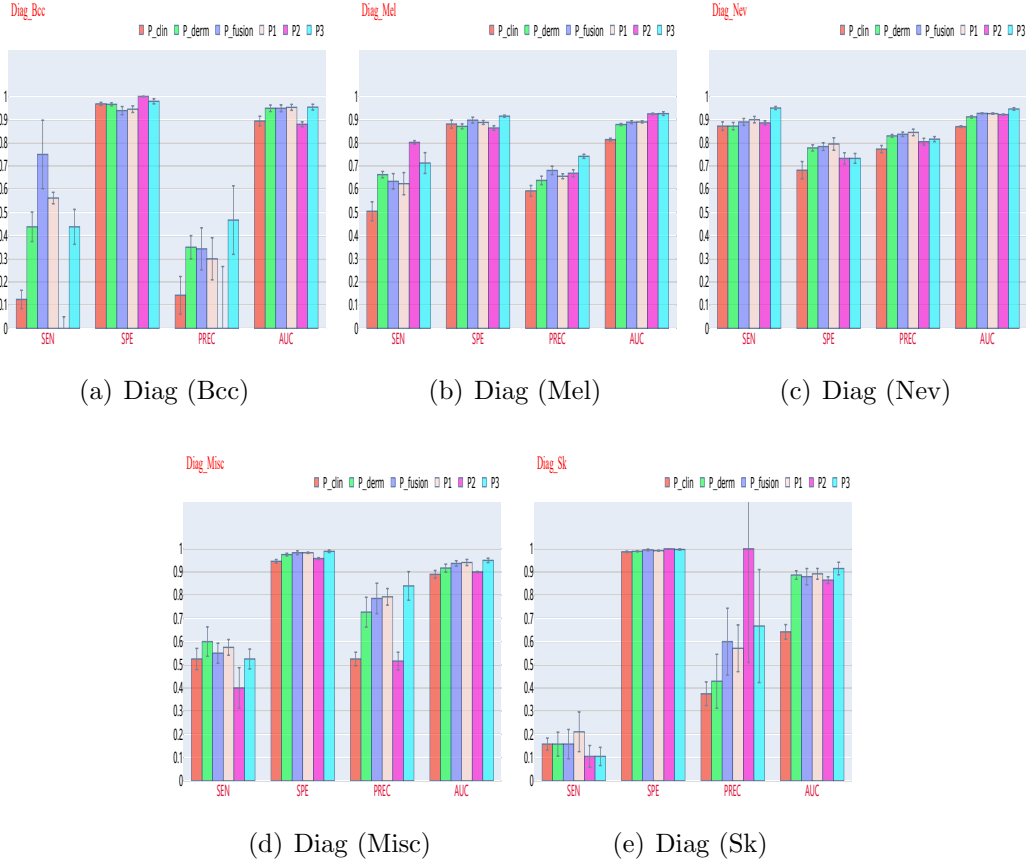
	$W_{clin}$	$W_{derm}$	$W_{fusion}$	<i>Bwv</i>	<i>Dag</i>	<i>Pig</i>	<i>Pn</i>	<i>Rs</i>	<i>Str</i>	<i>Vs</i>	<i>Diag</i>	<i>average</i>
$P_{clin}$	100	0	0	84.8±0.7	53.3±1.5	56.9±2.8	57.8±0.6	73.8±0.7	67.7±1.1	78.8±0.6	68.2±0.6	67.7±0.3
$P_{derm}$	0	100	0	87.6±0.7	59.4±1.6	70.8±0.6	66.6±1.6	80.9±1.0	74.0±0.7	81.1±0.7	74.6±0.7	74.4±0.4
$P_{fusion}$	0	0	100	87.0±0.6	<b>60.4±1.2</b>	70.2±0.9	67.0±1.6	80.5±1.6	74.0±0.6	81.8±0.5	<b>75.7±1.1</b>	74.6±0.5
$P_1$ (averaged)	33.3	33.3	33.3	87.5±0.5	60.3±1.0	69.5±0.8	66.8±1.6	81.0±0.6	<b>74.2±0.9</b>	81.7±0.5	75.4±1.5	74.6±0.4
$P_1$ (ours)	9.63±3.43	60.9±20.3	29.5±20.2	<b>87.9±0.7</b>	60.2±0.5	<b>71.5±0.7</b>	<b>67.6±1.8</b>	<b>81.9±1.0</b>	73.8±0.6	<b>82.1±0.5</b>	75.6±1.2	<b>75.1±0.4</b>

**Table 6.6:** The classification results of 2-predictions weighted average scheme (*AVG* accuracy in %),  $P_3$  (averaged) is the prediction obtained by averaging the  $P_1$  and  $P_2$ , while  $P_3$  (ours) is obtained by our Fusion Scheme 2. *average* is the average value of all the labels in the other column. The black bold number is the highest value for each column.

	$W_1$	$W_2$	<i>Bwv</i>	<i>Dag</i>	<i>Pig</i>	<i>Pn</i>	<i>Rs</i>	<i>Str</i>	<i>Vs</i>	<i>Diag</i>	<i>average</i>
$P_1$	100	0	87.9±0.7	60.2±0.5	<b>71.5±0.7</b>	67.6±1.8	<b>81.9±1.0</b>	73.8±0.6	<b>82.1±0.5</b>	75.6±1.2	75.1±0.4
$P_2$	0	100	86.7±0.6	65.4±0.4	65.7±1.1	<b>70.4±1.2</b>	79.0±0.9	75.8±0.3	79.2±0.0	75.4±0.7	74.7±0.2
$P_3$ (averaged)	50	50	88.2±0.5	<b>66.3±0.6</b>	69.5±1.5	70.2±0.8	80.9±0.9	<b>76.7±0.9</b>	80.3±0.2	75.9±0.9	76.0±0.3
$P_3$ (ours)	73.5±6.43	26.5±6.43	<b>88.5±0.4</b>	64.4±1.3	71.3±1.3	69.2±1.5	81.4±0.8	76.1±1.1	81.6±0.5	<b>77.6±1.5</b>	<b>76.3±0.7</b>

Then, we follow the structure in Fig. 4.2(a), where ResNet-50 is selected as CNN, to get C-D-M and remove the concatenation of the non-image modality to obtain C-D. Training of the models for C, D, C-D, and C-D-M is the same as that of our FusionNet. Training of the model for M is the same as that of the SVM cluster except that the input is  $F_{meta}$ . Finally,  $P_1$  and  $P_3$  are from our proposed FusionM4Net.





**Figure 6.6:** The detailed metrics of *Diag* labels of the results in Table. 6.4.

### 6.4.1.2 Performance of the predictions at different stages in our algorithm

To explore the effectiveness of different stages in our algorithm, we display the classification results of all the predictions at the two stages in FusionM4Net in Table 6.4. Fig. 6.6 displays the detailed metric of *Diag* label of the results in Table 6.3. All the predictions in Table 6.4 are from our FusionM4Net.  $P_{clin}$ ,  $P_{derm}$  and  $P_{fusion}$  are the predictions from three FCL branches  $B_{clin}$ ,  $B_{derm}$  and  $B_{fusion}$ , respectively.  $P_1$  is obtained by fusing the above-mentioned three predictions in the 3-Predictions Weighted Average scheme.  $P_2$  is the prediction from the SVM clusters that are trained by the patient’s metadata and  $P_1$ .  $P_3$  is obtained by integrating  $P_1$  and  $P_2$  in the 2-Predictions Weighted Average scheme.

**Table 6.7:** The classification results with different classifier-based clusters in the second stage of FusionM4Net (*AVG* accuracy in %). LR is the logistic regression, MLP is the multi-layer perceptron, and SVM is our chosen SVM cluster. *average* is the averaged value of all the labels in the other column. The black bold number is the highest value for each column.

	Clusters	<i>Bwv</i>	<i>Dag</i>	<i>Pig</i>	<i>Pn</i>	<i>Rs</i>	<i>Str</i>	<i>Vs</i>	<i>Diag</i>	<i>average</i>
$P_3$	LR	87.0±0.8	<b>64.8±1.0</b>	70.8±0.6	<b>70.9±1.7</b>	<b>81.7±0.7</b>	74.2±0.6	81.5±0.4	76.1±0.5	75.9±0.4
$P_3$	MLP	<b>89.1±0.7</b>	64.1±1.3	71.0±0.7	70.0±1.6	81.5±0.9	75.1±0.8	81.1±0.3	77.2±1.5	76.1±0.4
$P_3$	SVM	88.5±0.4	64.4±1.3	<b>71.3±1.3</b>	69.2±1.5	81.4±0.8	<b>76.1±1.1</b>	<b>81.6±0.5</b>	<b>77.6±1.5</b>	<b>76.3±0.7</b>

**Table 6.8:** The classification results with different meta-fusion schemes in the ISIC 2019 challenge, including Concat, Metanet [69] and Metablock [82], in the second stage of FusionM4Net (*AVG* accuracy in %). *average* is the average value of all the labels in the other column. The black bold number is the highest value for each column.

	Fusion stage	<i>Bwv</i>	<i>Dag</i>	<i>Pig</i>	<i>Pn</i>	<i>Rs</i>	<i>Str</i>	<i>Vs</i>	<i>Diag</i>	<i>average</i>
Concat	One-Stage	87.1±0.5	62.3±1.2	69.9±0.5	69.0±1.3	78.9±1.3	73.1±1.0	80.6±0.9	76.6±1.7	74.7±0.6
Metablock		86.2±1.0	60.6±1.1	59.9±1.4	64.6±1.2	73.6±1.4	71.8±0.8	77.7±1.2	75.0±1.0	71.2±0.4
Metanet		87.1±0.8	61.7±0.7	70.6±0.8	68.0±0.9	79.6±0.7	72.4±0.8	<b>81.9±0.9</b>	75.0±0.7	74.6±0.2
Ours	Two-Stage	<b>88.5±0.4</b>	<b>64.4±1.3</b>	<b>71.3±1.3</b>	<b>69.2±1.5</b>	<b>81.4±0.8</b>	<b>76.1±1.1</b>	81.6±0.5	<b>77.6±1.5</b>	<b>76.3±0.7</b>

#### 6.4.1.3 Effectiveness of the 3- and 2-predictions weighted average schemes

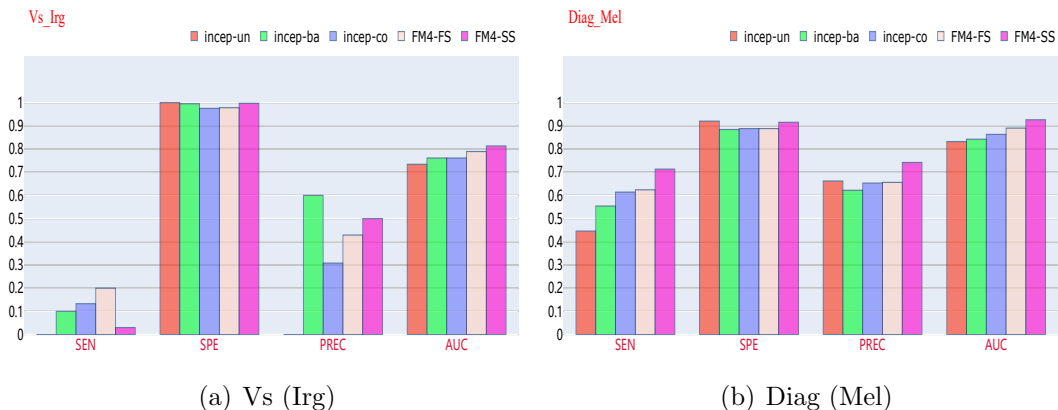
To show the effectiveness of the 3-predictions and 2-predictions weighted average schemes, we compared our proposed weighted average schemes with the normal average scheme in Tables 6.5 and 6.6. In Table 4.7,  $W_{clin}$ ,  $W_{derm}$  and  $W_{fusion}$  are the weights of  $P_{clin}$ ,  $P_{derm}$ , and  $P_{fusion}$ , respectively, and obtained by the 3-predictions weights searching scheme, as written in Algorithm 1. In Table 4.8,  $W_1$ ,  $W_2 = 1 - W_1$  are the corresponding weighted average of  $P_1$  and  $P_2$  and obtained by Algorithm 2.

#### 6.4.1.4 Comparison between different clusters at the second stage

To explain why we select a SVM cluster for the training at the second stage of our algorithm, we compare different clusters at the second stage, including logistic regression (LR), multi-layer perceptron (MLP) and SVM. The performance comparisons are shown in Table 6.7.

**Table 6.9:** Performance comparison between our algorithm and other state-of-the-art methods (*AVG* accuracy in %). The inception-based methods [60], Mma [81], TripleNet [39], EmbeddingNet [129], HcCNN [11] are compared with our methods.. All the results of the methods mentioned above, except our proposed method, are obtained from the papers of [60] and [11]. *average* is the average value of all the labels in the other column. The black bold number is the highest value for each column.

	<i>Bwv</i>	<i>Dag</i>	<i>Pig</i>	<i>Pn</i>	<i>Rs</i>	<i>Str</i>	<i>Vs</i>	<i>Diag</i>	<i>average</i>
Inception-unbalanced	87.6	56.7	65.6	68.1	78.2	75.9	81.3	68.4	72.7
Inception-balanced	87.3	60.3	64.8	68.9	78.2	75.7	81.5	70.8	73.4
Inception-combine	87.1	60	66.1	70.9	77.2	74.2	79.7	74.2	73.7
MmA	83	59.2	61.3	65.6	73.9	69.4	75.7	70.6	69.8
EmbeddingNet	84.3	57.5	64.3	65.1	78.0	73.4	82.5	68.6	71.7
TripleNet	87.9	61.3	67.3	63.3	76.0	74.4	83.0	68.6	72.7
HcCNN	87.1	65.6	68.6	70.6	80.8	71.6	<b>84.8</b>	69.9	74.9
FusionM4Net-FS	86.8	61.0	<b>72.4</b>	70.9	<b>83.0</b>	74.4	81.8	74.9	75.7
FusionM4Net-SS	<b>88.1</b>	<b>66.1</b>	70.1	<b>71.1</b>	81.5	<b>78.0</b>	81.8	<b>78.5</b>	<b>77.0</b>



**Figure 6.7:** The detailed metrics of *Diag* labels in Table. 6.10.

#### 6.4.1.5 Comparison between different meta-fusion schemes at the second stage

To show the effectiveness of our meta-fusion scheme, we compared our proposed Fusion Scheme 2 with other currently advanced meta-fusion methods, including C-D-M (concatenation), Metanet[69] and Metablock [82], which have been evaluated by [82] on the classification dataset of the International Skin Imaging Collaboration (ISIC) 2019 challenge [26, 115, 27]. The whole structure of C-D-M is the same as that of Metanet and Metablock, except for the fusion operation. C-D-M uses a

**Table 6.10:** The detailed seven-point checklist label and diagnosis label classification results of the Top 5 methods are in Table 6.9 (*AVG* accuracy in %). The HcCNN did not give detailed results, and the highlighted number is the highest values of SEN, SPE, PRE, and AUC for each column. Incep-un: Inception-unbalanced, Incep-ba: Inception-balanced, Incep-co: Inception-combined, FM4-FS: the prediction of the first stage in FusionM4Net, FM4-SS: the prediction of the first stage in FusionM4Net. The highest values OF SEN, SPE, PRE, and AUC are highlighted in black, black, green, and black, respectively. The results of Inception-based methods are from the corresponding paper [60].

Model	Met	Pn			Str			Pig			Rs		Vs		Buw		Dag			Diag					
		Asb	Typ	Aty	Asb	Reg	Irg	Asb	Reg	Irg	Asb	Prs	Asb	Reg	Irg	Asb	Prs	Asb	Reg	Irg	Nev	Bcc	Mel	Misc	Sk
Incep-un	SEN	<b>78.8</b>	77.4	35.5	<b>98.1</b>	36.4	34	83	6.2	57.3	<b>95.5</b>	31.1	<b>98.7</b>	23.1	0	<b>96.6</b>	49.3	34	59.3	67.8	94.1	25.0	44.6	35.0	5.3
	SPE	80.8	75.5	<b>93.7</b>	47.8	<b>98.6</b>	<b>94</b>	53.5	<b>99.4</b>	80.1	31.1	<b>95.5</b>	22	<b>97.1</b>	<b>100</b>	49.3	<b>96.6</b>	92.2	72.2	67.4	50.6	<b>98.4</b>	<b>92.2</b>	98.0	99.5
	PRE	72.8	64.9	<b>63.5</b>	77.8	<b>76.2</b>	<b>64</b>	69.8	60	56.8	79.1	71.7	82.8	<b>54.3</b>	0	89	<b>77.1</b>	59.6	47.6	62.8	70.3	40.0	66.2	66.7	33.0
	AUC	87.8	83.6	78.6	84.2	87.8	78.3	77.4	67.2	78.1	79.9	79.9	82.1	81.8	73.4	87	87	72.3	72.6	76.4	87.7	92.2	83.2	86.6	88.4
Incep-ba	SEN	78.2	76.0	41.9	90.7	43.2	50	73.5	16.7	<b>67.7</b>	84.1	62.3	86.8	30.8	10	72.5	65.3	43	66.1	66.1	91.3	25.0	55.4	42.5	15.8
	SPE	81.6	77.9	92.1	63.8	97.4	87.7	64.5	98.3	73.4	62.3	84.1	31.7	95.6	99.5	65.3	92.5	89.8	75.1	73.4	62.5	98.9	88.4	97.2	99.7
	PRE	73.5	66.9	61.9	82.3	67.9	56	72.9	57.1	53.8	58.8	58.9	84.4	51.6	<b>60</b>	91.9	67.1	58.9	53.1	66.9	75.2	50.0	62.2	63.0	75.0
	AUC	88.6	83.6	78.9	84.9	87.1	78.7	78.8	75.2	79.4	<b>83.5</b>	<b>83.5</b>	85.0	84.0	76.1	87.5	87.5	73	76.5	78.0	88.1	89.2	84.2	86.8	90.4
Incep-co	SEN	77.6	78.1	48.4	86.0	<b>54.5</b>	51.1	77.6	29.2	59.7	81.3	<b>66</b>	92.3	<b>42.3</b>	13.3	89.4	<b>77.3</b>	<b>47</b>	67.8	62.1	88.6	<b>62.5</b>	61.4	47.5	<b>42.1</b>
	SPE	<b>85.8</b>	78.7	90.7	<b>67.4</b>	85.7	85.7	<b>65.1</b>	94.2	80.1	<b>66.0</b>	81.3	45.1	92.4	97.5	<b>77.3</b>	89.4	87.8	72.6	<b>78.9</b>	71.6	97.9	88.8	97.5	99.5
	PRE	<b>78.1</b>	68.3	61.6	83.1	52.7	52.7	74.2	41.2	57.8	<b>86.7</b>	56.5	86.5	45.8	30.8	<b>94.4</b>	63.0	56.6	51.3	70.5	79.5	<b>55.6</b>	65.3	67.9	80.0
	AUC	<b>89.9</b>	84.2	79.9	86.1	78.9	78.9	79.0	74.9	79.0	82.9	82.9	86.2	85.5	76.1	89.2	89.2	74.1	76.5	79.0	89.7	92.9	86.3	88.3	91.0
FM4-FS	SEN	76.3	78.8	<b>49.5</b>	87.9	52.3	47.9	87.0	<b>39.6</b>	58.9	96.2	47.1	95.5	34.6	<b>20</b>	91.6	66.7	41.0	66.9	68.4	90.0	56.2	62.4	<b>57.5</b>	21.1
	SPE	83.3	<b>81.9</b>	90.1	64.5	95.2	88.4	62.8	97.1	87.1	47.2	96.2	<b>48.8</b>	93.6	97.8	66.7	91.6	89.5	76.9	72.9	<b>79.5</b>	94.5	88.8	98.3	99.2
	PRE	74.8	<b>71.9</b>	60.5	82.2	57.5	56.2	<b>75.2</b>	65.5	67.6	83.2	82.0	<b>87.7</b>	45	42.9	92.1	64.9	56.9	55.2	67.2	<b>84.5</b>	30.0	65.6	79.3	57.1
	AUC	86.9	85.9	83.9	85.8	87.9	81.4	<b>83.1</b>	80.9	83.5	81.7	81.7	89.4	87.8	78.9	90.6	90.6	73.9	79.1	80.1	92.6	95.3	89.0	94.1	89.2
FM4-SS	SEN	76.3	<b>79.5</b>	<b>49.5</b>	91.4	47.7	<b>55.3</b>	<b>89.7</b>	18.8	54.8	95.2	44.3	98.1	28.8	0.3	94.4	64.0	40.0	<b>76.3</b>	<b>76.3</b>	<b>95.0</b>	43.8	<b>71.3</b>	52.5	10.5
	SPE	83.3	81.5	90.7	66.7	97.2	89.7	50.0	98.8	<b>89.7</b>	44.3	95.2	29.3	96.2	99.7	64.0	94.4	<b>92.5</b>	<b>78.0</b>	76.6	73.3	97.9	91.5	<b>98.9</b>	<b>99.7</b>
	PRE	74.8	71.6	62.2	<b>83.6</b>	67.7	62.7	69.9	<b>69.2</b>	<b>70.8</b>	82.3	<b>77.0</b>	84.1	53.6	50.0	91.8	72.7	<b>64.5</b>	<b>58.5</b>	<b>72.6</b>	81.6	46.7	<b>74.2</b>	<b>84.0</b>	<b>66.7</b>
	AUC	88.1	<b>87.3</b>	<b>85.7</b>	<b>87.2</b>	<b>89.6</b>	<b>84.6</b>	<b>83.1</b>	<b>82.5</b>	<b>84.9</b>	83.0	83.0	<b>89.5</b>	<b>87.9</b>	<b>81.3</b>	<b>92.5</b>	<b>92.5</b>	<b>78.2</b>	<b>82.5</b>	<b>84.5</b>	<b>94.6</b>	<b>95.4</b>	<b>92.6</b>	<b>95.0</b>	<b>91.5</b>

concatenation operation to combine the encoded metadata vector with the image vector extracted from the CNNs (see Fig. 2(a)). Metanet and Metablock employed metadata information as attention maps to highlight the most relevant features of image vectors. We follow the released code from [82] to build these three models. The performance comparisons are shown in Table 6.8.

## 6.4.2 Comparison between our algorithm and other state-of-the-art methods

To show the advantage of the proposed algorithm, in Table 6.9, we compare our FusionM4NET with current state-of-the-art methods, including the inception-based

methods [60], Mma [81], TripleNet [39], EmbeddingNet [129], HcCNN [11] based on the value of average accuracy. All the results of the methods mentioned above, except our proposed method, are obtained from the respective paper and from [11]. Note that all of the cited results are from a single experiment, so they are presumed to be the result with the highest average accuracy value. We show FusionM4Net-FS and FusionM4Net-SS, the first and second stage of FusionM4Net, the corresponding outputs of which are  $P_1$  and  $P_3$ . Table 6.10 presents the full detailed results of the top-performing multi-modal skin lesion classification methods, including the value of SEN, SPE, PRE, and AUC.

## 6.5 Discussion

### 6.5.1 Evaluation of FusionM4Net

#### 6.5.1.1 Comparisons between single- and multi-modality learning

Table 6.3 displays that in the comparisons with single-modality models, the model trained by dermoscopy images outperforms that trained by clinical images and patient metadata. This corresponds with medical doctor’s diagnosis, as in dermatologists’ actual diagnosis, the results according to dermoscopy images are more accurate than that by naked eyes and patient’s metadata.

In the comparisons between our two-stage multi-modality models and common one-stage multi-modality (COSMM) models, all the models trained by our method have higher averaged accuracy than their counterparts trained by COSMM learning. These results reflected that the proposed two-stage method is more suitable than the one-stage method to fuse multi-modality data for multi-label skin lesion classification tasks.

It can also be seen that the C-D model trained by the COSMM method has a lower value in *AVG* accuracy but a higher value in *Diag* accuracy compared to the single-modality model of D. A similar situation happened in the comparison between C-D-M and D-M models trained by the COSMM method.

We attribute this to the initial design of a seven-point checklist, which is used to find visible features under dermoscopy images [60]. Therefore, integrating clinical images may not improve the accuracy of SPC labels and even negatively affect the co-training of multi-label classification.

Furthermore, we observed that our  $P_3$  is only about 2% higher than the single-modality model D but over 6% higher than single-modality models C and M. The similar phenomenon also happened in the multi-modality model comparison. For

example, in comparing our two-stage multi-modality models, the  $P_3$  only gets 0.7% – 1.2% higher Avg accuracy than the models  $P_1$ (C-D) and D-M that consist of the modality of dermoscopy image, and get over about 6% higher Avg accuracy than the model C-M. These results also reflect that dermoscopy image is the most important modality for the multi-label skin lesion classification task, and clinical images and patient metadata play a supplementary role.

### 6.5.1.2 Performance of the predictions at different stages in our algorithm

From Table 6.3 and Table 6.4, we can observe that  $P_{clin}$  and  $P_{derm}$  obtained by our multi-modality learning method achieve similar average accuracy to the predictions from model C and D obtained by single modality training. These results demonstrate our method can learn similar single-modality-specific feature representations. However, the prediction from fusion branch  $P_{fusion}$  gets a 0.5% higher average accuracy than the prediction from model C-D, which only focuses on the optimization of cross-modality common representations from two-modality images. It proves that while optimizing cross-modality feature representations, optimizing single-modality feature representations can get a better common feature from these two-modality images for the multi-label classification task (see Fig. 6.3 as well as Fig. 3 in our supplementary materials to see the difference between our method and models C, D and C-D).

Table 6.4 also shows that, fusing the prediction  $P_2$  from SVM clusters and  $P_1$  into  $P_3$  can boost the average accuracy from 75.1% to 76.3% and the *Diag* accuracy from 75.6% to 77.6%, compared to  $P_1$ . These results show  $P_2$ , which is trained by patient’s metadata and multi-label predictive information can provide useful supplementary information to  $P_1$ , which is trained by clinical and dermoscopy images for the multi-label classification task.

### 6.5.1.3 Effectiveness of 3- and 2- predictions weighted average scheme

Table 6.5 shows that  $P_1$  (ours) gets the best average accuracy 75.1%, while  $P_1$  (averaged) has a same value in average accuracy 74.6% , compared to the 74.6% of  $P_{fusion}$ . These results demonstrate that the normal average scheme can not get higher accuracy and may even yield a worse multi-label classification result than  $P_{fusion}$ , while our Fusion Scheme 1 obtains an increase of about 0.5%. These results, in turn, show the importance of information fusion at the decision level. It also can be seen that the 3-predictions weighted average scheme reaches the best accuracy on the validation dataset when  $W_{clin} = 9.6\% \pm 3.4\%$ ,  $W_{derm} = 60.9\% \pm 20.3\%$ , and

$W_{fusion} = 29.5\% \pm 20.2\%$ , which illustrates that the integrated results  $P_1$  is mainly formed by dermoscopy  $P_{derm}$  and fusion branch  $P_{fusion}$ .

From Table 6.6, we can see that no matter which fusion scheme is based on, the  $P_3$ s, including  $P_3$  (ours) at 76.3% and  $P_3$  (averaged) at 76.0%, are about 1% higher than either  $P_1$  at 75.1% or  $P_2$  at 74.7%. We attribute this to the complementarity between  $P_1$  obtained from the patient’s metadata and multi-label predictive information and  $P_2$  obtained from two-modality images. Furthermore,  $P_3$  (ours) outperforms  $P_3$  (averaged) in both *AVG* accuracy and *Diag* accuracy, which illustrates the 2-predictions weight average scheme can more effectively integrate the  $P_1$  and  $P_2$  than the normal averaging scheme. Furthermore, Table 6.6 presents that the 2-predictions weighted average scheme achieves the best accuracy on the validation dataset when  $W_1 = 73.5\% \pm 6.43\%$  and  $W_2 = 26.5\% \pm 6.4\%$ . These results reflect that the captured information from the image-modalities play the key role in the final classification, and that the patient’s metadata are the supplementary materials. What’s more, in Table 6.5, we can see that the improvement from  $P_{derm}$  and  $P_{fusion}$  to  $P_1$  (averaged) and  $P_1$  (ours) is about 0.5%, which is not significant. We also attribute this to the above-mentioned reason that the dermoscopy image plays the most crucial role in the multi-label skin lesion classification task. Therefore, the performance gap between all the models trained with dermoscopy images or their combination will not be significant. We think the weighted fusion scheme can only make the clinical image and patient’s metadata play a better supplementary role than the common averaged scheme. So, the performance gaps between  $P_1$  (ours) and  $P_1$  (averaged) (and  $P_3$  (ours) and  $P_3$  (averaged) in Table 6.5 ) is not too obvious.

#### 6.5.1.4 Comparison between different clusters at the second stage

The classification results of the different clusters at the second stage are presented in Table 6.7. From this table, it can be seen that all the clusters boost the classification performance based on  $P_1$ .  $P_3$  obtained by MLP and SVM has a comparable classification performance and outperforms that obtained by LR. The prediction based on LR has a slight decrease in average accuracy compared with that based on MLP and SVM in Table 6.7. We believe this is because the clusters based on a more advanced classifier can extract more useful information from the multi-label predictive vector and the patient’s meta-data vector.

Also, in Table 6.7, we can see that no matter which classifier is based on, the  $P_3$  can get about 1% higher *Avg* accuracy than  $P_1$ . These results illustrate the improvement is from the benefits of our two-stage fusion scheme and not from the classifiers. Furthermore, in our experiments, the SVM gets the highest *Avg* accuracy

of 76.3%, so we chose it as the classifier cluster in the second stage of our method. However, the MLP also achieves comparable performance compared with the SVM, so it is also recommended that the MLP be tested in other cases.

#### 6.5.1.5 Comparison between different meta-fusion scheme at the second stage

It can be observed in Table 6.8 that our proposed FusionM4Net achieves the best *AVG* accuracy of 76.4%, which is about 1.5% higher than that of C-D-M and Metanet and significantly higher than that of Metablock. We attribute the improvement of our method to the effectiveness of our two-stage fusion method, which makes full use of patient metadata and the correlation of multi-label predictive information.

Also, we can see that C-D-M and Metanet outperform Metablock by a large margin in terms of the value of *AVG* accuracy. Actually, according to the results reported by [82], Metablock can achieve better performance than C-D-M and Metanet on a single skin disease recognition task.

#### 6.5.2 Comparison between our algorithm and other state-of-the-art methods

According to the experimental results in Table 6.9, MmA obtains the lowest value among all the methods. The explanation is that MmA learns the feature from a single shared layer, which is not capable of capturing the essential spatial information from the image. TripleNet gets a 1% higher average accuracy than EmbeddingNet, which is attributed to the effectiveness of the extra subnetwork and the co-optimization of single- and cross-modality in TripleNet. Also, the inception-balanced method obtains a 0.7% increase in average accuracy compared with the inception-unbalanced method because it makes the less-frequent labels be trained by the model more times. It should also be noted that the Inception-combined method has higher accuracy than that of HcCNN on the *Diag* label classification but lower accuracy in other label classifications, including *Bwv*, *Pig*, *Vs* and *Rs*. Therefore, the Inception-combined method finally gets a 1.2% lower average accuracy in comparison with HcCNN, which resulted from HcCNN’s extra multi-scale network; the hybrid network can fuse feature information from early-stage to late-stage of the CNN backbone.

Our FusionM4Net achieves the highest average accuracy and diagnosis accuracy on the multi-modal multi-label SLC dataset, of which the average accuracy is 1.6% - 3.8% higher than Inception-based methods and HcCNN (see Table 6.9). First, in comparison with the HcCNN in Table 6.9, the first stage of our FusionM4Net



(FusionM4Net-FS) gets an increase of 1% in average accuracy and about 5% accuracy on *Diag* label compared to HcCNN, under the training of only two-modality images. This is because our FusionM4Net-FS not only focuses on feature-level fusion by single-modality and cross-modality optimization together but also decision-level fusion via the 3-predictions weighted average scheme.

Second, compared with the Inception-combined method, our FusionM4Net has a higher value in all the label classification tasks (see Table 6.9) and a significant increase of 2.8% in average accuracy, under the training of three-modality data (clinical image, dermoscopy image, and patient’s metadata). These results are attributed to the superiority of our two-stage multi-modal learning algorithm, which effectively extracts and fuses information from two-modality images at the first stage and takes full use of the information from the patient’s meta-data at the second stage.

Table 6.10 gives detailed information on the top-performing methods, including the Inception-based methods and our algorithm. The second stage of FusionM4Net (FusionM4Net-SS) get the highest AUC value in all the label types except the Pn (Asb) and Str (Irg). These results correspond to the result in Table 6.9 that the second stage of FusionM4Net (FusionNet-SS) outperforms the inception-based methods in all labels regarding the *AVG* accuracy. In the SPC label, it can be observed that FusionM4Net-SS gets the value of 0 in sensitivity of *Vs* Irg, while Inception-balanced and Inception-combined methods obtain 10% and 13.3%, respectively (see Fig. 6.7 (a)). This increase in the Inception-combined method is attributed to the effectiveness of balanced-sampling techniques of Inception-balanced and Inception-combined methods. However, regarding the precision PRE and AUC, FusionM4Net-SS obtains the highest value in most of the labels. In the *Diag* label, the Inception-combined method gets the highest AUC value in *Diag* Sk, while our FusionM4net achieves the best AUC value in the remaining four *Diag* labels. Notably, in the classification of the skin cancer melanoma, the FusionM4Net-SS improves all the metrics compared with inception-combined methods, from SEN 61.4%, SPE 88.8%, PRE 65.3% and AUC 86.3% to 71.3%, 91.5%, 74.2%, and 92.6%, respectively (see Fig. 6.7 (b)).

### 6.5.3 Potential future work

In this work, we only focus on the better fusion scheme to combine clinical image, dermoscopy image, and patient’s metadata for multi-label SLC and ignore making use of the label dependencies, which is important for multi-label SLC task. Therefore, in the future, we will employ the currently advanced graph convolutional network (GCN) [22] to capture and explore the useful label dependencies to further improve the performance of multi-label SLC tasks.

## 6.6 Conclusion

We propose a multi-stage multi-modal learning algorithm for multi-label skin diseases classification, named FusionM4Net, which involves two stages: first learning feature information from clinical and dermoscopy, and second further integrating patient's meta-data and decision information from the two-modality images. Extensive experiments and comparisons on a publicly available dataset show that FusionM4Net has higher accuracy than any other current state-of-the-art methods.

# Concluding Remarks

This dissertation bridges the gap in multi-modal-based skin lesion classification by exploring the application of multi-modal fusion methods in various scenarios, including clinical-dermoscopy images, single images, and patient metadata. The primary objective of this doctoral project is to enhance the efficiency of dermatologists, and it was conducted in collaboration with medical professionals in dermatology.

After showcasing the advantages of multi-modal deep learning for dermatologist’s diagnosis in Chapter 1, the subsequent chapters delve into the background of deep learning and the application of single-modal skin lesion classification in Chapters 2 and 3, respectively. Following this, four innovative fusion methods were proposed to enhance classification performance, which includes accuracy, computational cost, or both, in different scenarios. These methods are one for single image and metadata fusion (Chapter 4), two for clinical-dermoscopy image fusion (Chapter 5), and one for the fusion of all three modalities (Chapter 6).

As discussed in the previous chapter, current deep-learning models have the potential to alleviate doctors’ burdens and enhance the efficiency of dermatologists. The results of this dissertation highlight the potential of designing a more accurate and parameter-efficient decision support system based on multi-modal data. I hope that my work can inspire future research in the field of multi-modal-based skin lesion analysis, and some parts of this work can be used in clinics or teledermatology.

## 7.1 Outlook

Despite the progress made, there are areas for further improvements in the field of skin lesion classification, such as the lack of high-quality, large-scale datasets, the need to process dermatologists’ clinical reports, and the integration of large language models (LLMs) into the patient care pipeline.

### 7.1.1 Large Scale Dataset

In the field of skin lesion classification, high-quality public datasets are limited in terms of the number of cases. For example, the PH2 dataset [76] includes dermoscopic images, corresponding diagnosis labels, segmentation masks, and annotations of dermoscopic features, but it only has two samples. Similarly, the SPC dataset [60] consists of data from three modalities, diagnosis labels, and seven-point checklist features, but it only contains about 1,000 cases. Large-scale and high-quality datasets are not only used for more robust evaluation of algorithms. More important is that it is the basis for building General AI for skin lesion analysis like ChatGpt [89, 90, 17, 1]. Semi-supervised methods could be utilized to reduce the annotation of the dataset.

### 7.1.2 Extracting information from clinical report

In the clinic, the patient's information and diagnosis are not well-structured and suitable for direct training in deep learning methods. They are usually included in a report. So, to reduce the amount of time needed for data annotation, natural language processing systems are required to extract structured labels and patient metadata from free-text dermatology reports.

### 7.1.3 Large Language Models

In the clinic, it takes lots of time for the dermatologist to explain the skin disease and give suggestions to the patients. Especially in rural areas, there is a shortage of dermatologists. With the success of ChatGpt, [89, 90, 17, 1]., the interactive skin diseases analysis platform based on LLMs has the potential to address this issue.

# Appendices



# List of Publications

## A.1 Journal Publications

- **Tang, P.**, Yan, X., Nan, Y., Xiang, S., Krammer, S., and Lasser, T. (2022). FusionM4Net: A multi-stage multi-modal learning algorithm for multi-label skin lesion classification. *Medical Image Analysis*, 76, 102307.  
**Tang, P.**, Yan, X., Nan, Y., Hu, X., Krammer, B. H. M., and Lasser, T. (2024). Joint-Individual Fusion Structure with Fusion Attention Module for Multi-Modal Skin Cancer Classification. *Pattern Recognition*, 154, 110604.
- Nan, Y., **Tang, P.**, Zhang, G., Zeng, C., Liu, Z., Gao, Z., ... and Yang, G. (2022). Unsupervised tissue segmentation via deep constrained gaussian network. *IEEE Transactions on Medical Imaging*, 41(12), 3799-3811.
- Nan, Y., Li, F., **Tang, P.**, Zhang, G., Zeng, C., Xie, G., ... and Yang, G. (2022). Automatic fine-grained glomerular lesion recognition in kidney pathology. *Pattern Recognition*, 127, 108648.
- Krammer, S., Li, Y., Jakob, N., Boehm, A. S., Wolff, H., **Tang, P.**, ... and Hartmann, D. (2022). Deep learning-based classification of dermatological lesions given a limited amount of labelled data. *Journal of the European Academy of Dermatology and Venereology*, 36(12), 2516-2524.

## A.2 Conference Publications

- **Tang, P.**, Xu, Z., Zhou, C., Wei, P., Han, P., Cao, X., and Lasser, T. (2024, March). Prior and Prediction Inverse Kernel Transformer for Single Image Defocus Deblurring. *In Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 6, pp. 5145-5153).

### A.3 Submitted Journal Publications

- **Tang, P.**, and Lasser, T. (2024). Unsupervised Anomaly Detection in Medical Images Using Encoder-Attention-2Decoders Reconstruction (Submitted to *IEEE Transactions on Medical Imaging*, Major Revision)
- **Tang, P.**, and Lasser, T. (2024). Pay Less On Clinical Images: Asymmetric Multi-Modal Fusion Method For Efficient Multi-Label Skin Lesion Classification. (Submitted to *Medical Image Analysis*, Under Review)
- **Tang, P.**, and Lasser, T. (2024). Single-Shared Network with Prior-Inspired Loss for Parameter-Efficient Multi-Modal Imaging Skin Lesion Classification. (Submitted to *Experts Systems with Applications*, Under Review)
- **Tang, P.**, Nan, Y., and Lasser, T. (2023). Graph-Ensemble Learning Model for Multi-label Skin Lesion Classification using Dermoscopy and Clinical Images. arXiv preprint (Submitted to *IEEE Journal of Biomedical and Health Informatics*).
- **Tang, P.**, Xu, Z., Wei, P., Hu, X., Zhao, P., Cao, X., ... and Lasser, T. (2023). SR-R<sup>2</sup> KAC: Improving Single Image Defocus Deblurring. (Submitted to *IEEE Transactions on Cybernetics*, Under Review)





## Bibliography

- [1] Josh Achiam et al. “Gpt-4 technical report.” In: *arXiv preprint arXiv:2303.08774* (2023).
- [2] Shubhra Aich et al. “Multi-scale weight sharing network for image recognition.” In: *Pattern Recognition Letters* 131 (2020), pp. 348–354.
- [3] Giuseppe Argenziano et al. “Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: comparison of the ABCD rule of dermoscopy and a new 7-point checklist based on pattern analysis.” In: *Archives of dermatology* 134.12 (1998), pp. 1563–1570.
- [4] Giuseppe Argenziano et al. “Seven-point checklist of dermoscopy revisited.” In: *British Journal of Dermatology* 164.4 (2011), pp. 785–790.
- [5] Vaswani Ashish. “Attention is all you need.” In: *Advances in neural information processing systems* 30 (2017), p. I.
- [6] Pradeep K Atrey et al. “Multimodal fusion for multimedia analysis: a survey.” In: *Multimedia systems* 16.6 (2010), pp. 345–379.
- [7] Charles M Balch et al. “Final version of 2009 AJCC melanoma staging and classification.” In: *Journal of clinical oncology* 27.36 (2009), p. 6199.
- [8] C. Barata, M. E. Celebi, and J. S. Marques. “Improving Dermoscopy Image Classification Using Color Constancy.” In: *IEEE Journal of Biomedical and Health Informatics* 19.3 (2015), pp. 1146–1152.
- [9] Catarina Barata, M. Emre Celebi, and Jorge S. Marques. “Development of a clinically oriented system for melanoma diagnosis.” In: *Pattern Recognition* 69 (2017), pp. 270–285.
- [10] Marin Benčević et al. “Understanding skin color bias in deep learning-based skin lesion segmentation.” In: *Computer methods and programs in biomedicine* 245 (2024), p. 108044.

- [11] Lei Bi et al. “Multi-label classification of multi-modality skin lesion via hyper-connected convolutional neural network.” In: *Pattern Recognition* 107 (2020), p. 107502.
- [12] Yequan Bie, Luyang Luo, and Hao Chen. “MICA: Towards Explainable Skin Lesion Diagnosis via Multi-Level Image-Concept Alignment.” In: *arXiv preprint arXiv:2401.08527* (2024).
- [13] Michael Binder et al. “Epiluminescence microscopy: a useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists.” In: *Archives of dermatology* 131.3 (1995), pp. 286–291.
- [14] Alexander BörVE et al. “Smartphone teledermoscopy referrals: a novel process for improved triage of skin cancer patients.” In: *Acta dermato-venereologica* 95.2 (2015), pp. 186–190.
- [15] Ralph Peter Braun et al. “Dermoscopy of pigmented skin lesions.” In: *Journal of the American Academy of Dermatology* 52.1 (2005), pp. 109–121.
- [16] Freddie Bray et al. “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.” In: *CA: a cancer journal for clinicians* 68.6 (2018), pp. 394–424.
- [17] Tom Brown et al. “Language models are few-shot learners.” In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.
- [18] Alexander Buslaev et al. “Albumentations: fast and flexible image augmentations.” In: *Information* 11.2 (2020), p. 125.
- [19] Gan Cai et al. “A multimodal transformer to fuse images and metadata for skin disease classification.” In: *The Visual Computer* (2022), pp. 1–13.
- [20] M Emre Celebi, Noel Codella, and Allan Halpern. “Dermoscopy image analysis: overview and future directions.” In: *IEEE journal of biomedical and health informatics* 23.2 (2019), pp. 474–478.
- [21] Tirtha Chanda et al. “Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma.” In: *Nature Communications* 15.1 (2024), p. 524.
- [22] Zhao-Min Chen et al. “Multi-Label Image Recognition With Graph Convolutional Networks.” In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 5172–5181.
- [23] François Chollet et al. “Keras.” In: (2015). URL: <https://github.com/fchollet/keras>.

- [24] François Chollet. “Xception: Deep learning with depthwise separable convolutions.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1251–1258.
- [25] Noel Codella et al. “Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic).” In: *arXiv preprint arXiv:1902.03368* (2019).
- [26] Noel CF Codella et al. “Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic).” In: *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE. 2018, pp. 168–172.
- [27] Marc Combalia et al. “Bcn20000: Dermoscopic lesions in the wild.” In: *arXiv preprint arXiv:1908.02288* (2019).
- [28] Jifeng Dai et al. “Deformable convolutional networks.” In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 764–773.
- [29] A Dascalu et al. “Non-melanoma skin cancer diagnosis: a comparison between dermoscopic and smartphone images by unified visual and sonification deep learning algorithms.” In: *Journal of cancer research and clinical oncology* (2022), pp. 1–9.
- [30] Jia Deng et al. “Imagenet: A large-scale hierarchical image database.” In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.
- [31] Sadia Din, Omar Mourad, and Erchin Serpedin. “LSCS-Net: A lightweight skin cancer segmentation network with densely connected multi-rate atrous convolution.” In: *Computers in Biology and Medicine* (2024), p. 108303.
- [32] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale.” In: *arXiv preprint arXiv:2010.11929* (2020).
- [33] Andre Esteva et al. “Dermatologist-level classification of skin cancer with deep neural networks.” In: *nature* 542.7639 (2017), pp. 115–118.
- [34] Daniel G Federman, John Concato, and Robert S Kirsner. “Comparison of dermatologic diagnoses by primary care practitioners and dermatologists: a review of the literature.” In: *Archives of family medicine* 8.2 (1999), p. 170.
- [35] Hao Feng et al. “Comparison of dermatologist density between urban and rural counties in the United States.” In: *JAMA dermatology* 154.11 (2018), pp. 1265–1271.

- [36] Xiaohang Fu et al. “Graph-based intercategory and intermodality network for multilabel classification and melanoma diagnosis of skin lesions in dermoscopy and clinical images.” In: *IEEE Transactions on Medical Imaging* 41.11 (2022), pp. 3266–3277.
- [37] Geng Gao et al. “Multi-view compression and collaboration for skin disease diagnosis.” In: *Expert Systems with Applications* (2024), p. 123395.
- [38] Jibin Gao et al. “An asymmetric modeling for action assessment.” In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer. 2020, pp. 222–238.
- [39] Zongyuan Ge et al. “Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images.” In: *Medical Image Computing and Computer Assisted Intervention– MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11–13, 2017, Proceedings, Part III 20*. Springer. 2017, pp. 250–258.
- [40] Ian Goodfellow et al. *Deep learning*. Vol. 1. MIT press Cambridge, 2016.
- [41] Yanyang Gu et al. “Progressive transfer learning and adversarial domain adaptation for cross-domain skin disease classification.” In: *IEEE journal of biomedical and health informatics* 24.5 (2019), pp. 1379–1393.
- [42] David Gutman et al. “Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC).” In: *arXiv preprint arXiv:1605.01397* (2016).
- [43] Balazs Harangi. “Skin lesion classification with ensembles of deep convolutional neural networks.” In: *Journal of biomedical informatics* 86 (2018), pp. 25–32.
- [44] Charles R Harris et al. “Array programming with NumPy.” In: *Nature* 585.7825 (2020), pp. 357–362.
- [45] Kaiming He et al. “Deep residual learning for image recognition.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [46] Xiaoyu He et al. “Co-attention fusion network for multimodal skin cancer diagnosis.” In: *Pattern Recognition* 133 (2023), p. 108990.
- [47] Xingxin He et al. “Multi-Modal Retinal Image Classification With Modality-Specific Attention Network.” In: *IEEE Transactions on Medical Imaging* 40.6 (2021), pp. 1591–1602.

- [48] Tamara Salam Housman et al. “Skin cancer is among the most costly of all cancers to treat for the Medicare population.” In: *Journal of the American Academy of Dermatology* 48.3 (2003), pp. 425–429.
- [49] Andrew Howard et al. “Searching for mobilenetv3.” In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 1314–1324.
- [50] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. “Sharable and Individual Multi-View Metric Learning.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.9 (2018), pp. 2281–2288.
- [51] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. “Sharable and individual multi-view metric learning.” In: *IEEE transactions on pattern analysis and machine intelligence* 40.9 (2017), pp. 2281–2288.
- [52] Gao Huang et al. “Densely connected convolutional networks.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.
- [53] Lang Huang et al. “Learning where to learn in cross-view self-supervised learning.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 14451–14460.
- [54] Shih-Cheng Huang et al. “Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines.” In: *NPJ digital medicine* 3.1 (2020), pp. 1–9.
- [55] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift.” In: *International conference on machine learning*. pmlr. 2015, pp. 448–456.
- [56] Pavel Izmailov et al. “Averaging weights leads to wider optima and better generalization.” In: *arXiv preprint arXiv:1803.05407* (2018).
- [57] Shruti Jadon. “A survey of loss functions for semantic segmentation.” In: *arXiv preprint arXiv:2006.14822* (2020).
- [58] Jean Kanitakis. “Anatomy, histology and immunohistochemistry of normal human skin.” In: *European journal of dermatology* 12.4 (2002), pp. 390–401.
- [59] J. Kawahara, A. BenTaieb, and G. Hamarneh. “Deep features to classify skin lesions.” In: *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. 2016, pp. 1397–1400.
- [60] Jeremy Kawahara et al. “Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets.” In: *IEEE Journal of Biomedical and Health Informatics* 23.2 (2019), pp. 538–546.

- [61] Hoel Kervadec et al. “Boundary loss for highly unbalanced segmentation.” In: *International conference on medical imaging with deep learning*. 2019, pp. 285–296.
- [62] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization.” In: *arXiv preprint arXiv:1412.6980* (2014).
- [63] Paul AJ Kolarsick, Maria Ann Kolarsick, and Carolyn Goodwin. “Anatomy and physiology of the skin.” In: *Journal of the Dermatology Nurses’ Association* 3.4 (2011), pp. 203–213.
- [64] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks.” In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.
- [65] Yann LeCun et al. “Generalization and network design strategies.” In: *Connectionism in perspective* 19.143-155 (1989), p. 18.
- [66] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning.” In: *nature* 521.7553 (2015), p. 436.
- [67] Weipeng Li et al. “Fusing Metadata and Dermoscopy Images for Skin Disease Diagnosis.” In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. 2020, pp. 1996–2000.
- [68] Tsung-Yi Lin et al. “Focal loss for dense object detection.” In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.
- [69] Yuan Liu et al. “A deep learning system for differential diagnosis of skin diseases.” In: *Nature medicine* 26.6 (2020), pp. 900–908.
- [70] Ze Liu et al. “Swin transformer: Hierarchical vision transformer using shifted windows.” In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022.
- [71] Ze Liu et al. “Video swin transformer.” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 3202–3211.
- [72] Zhuang Liu et al. “A convnet for the 2020s.” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 11976–11986.
- [73] Zihao Liu, Ruiqin Xiong, and Tingting Jiang. “CI-net: clinical-inspired network for automated skin lesion recognition.” In: *IEEE Transactions on Medical Imaging* 42.3 (2022), pp. 619–632.
- [74] Li Ma and Richard C. Staunton. “Analysis of the contour structural irregularity of skin lesions using wavelet decomposition.” In: *Pattern Recognition* 46.1 (2013), pp. 98–106. ISSN: 0031-3203.

- [75] Kazuhisa Matsunaga et al. “Image Classification of Melanoma, Nevus and Seborrheic Keratosis by Deep Neural Network Ensemble.” In: *CoRR* abs/1703.03108 (2017). arXiv: [1703.03108](https://arxiv.org/abs/1703.03108).
- [76] Teresa Mendonça et al. “PH 2-A dermoscopic image database for research and benchmarking.” In: *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE. 2013, pp. 5437–5440.
- [77] Afonso Menegola et al. “Knowledge transfer for melanoma screening with deep learning.” In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. 2017, pp. 297–300.
- [78] Michael Migden et al. “Burden and treatment patterns of advanced basal cell carcinoma among commercially insured patients in a United States database from 2010 to 2014.” In: *Journal of the American Academy of Dermatology* 77.1 (2017), pp. 55–62.
- [79] Gilberto Moreno et al. “Prospective study to assess general practitioners’ dermatological diagnostic skills in a referral setting.” In: *Australasian journal of dermatology* 48.2 (2007), pp. 77–82.
- [80] Franz Nachbar et al. “The ABCD rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions.” In: *Journal of the American Academy of Dermatology* 30.4 (1994), pp. 551–559.
- [81] Jiquan Ngiam et al. “Multimodal Deep Learning.” In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. ICML’11. Bellevue, Washington, USA: Omnipress, 2011, pp. 689–696. ISBN: 9781450306195.
- [82] Andre G. C. Pacheco and Renato A. Krohling. “An Attention-Based Mechanism to Combine Images and Metadata in Deep Learning Models Applied to Skin Cancer Classification.” In: *IEEE Journal of Biomedical and Health Informatics* 25.9 (2021), pp. 3554–3563.
- [83] Andre G.C. Pacheco et al. “PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones.” In: *Data in Brief* 32 (2020), p. 106221. ISSN: 2352-3409.
- [84] Andre GC Pacheco and Renato A Krohling. “The impact of patient clinical information on automated skin cancer detection.” In: *Computers in biology and medicine* 116 (2020), p. 103545.

- [85] Adam Paszke et al. “Pytorch: An imperative style, high-performance deep learning library.” In: *Advances in neural information processing systems* 32 (2019), pp. 8026–8037.
- [86] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python.” In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [87] Michael Person et al. “Multimodal fusion object detection system for autonomous vehicles.” In: *Journal of Dynamic Systems, Measurement, and Control* 141.7 (2019).
- [88] Shangran Qiu et al. “Fusion of deep learning models of MRI scans, Mini-Mental State Examination, and logical memory test enhances diagnosis of mild cognitive impairment.” In: *Alzheimer’s Dementia: Diagnosis, Assessment Disease Monitoring* 10 (2018), pp. 737–749.
- [89] Alec Radford et al. “Improving language understanding by generative pre-training.” In: (2018).
- [90] Alec Radford et al. “Language models are unsupervised multitask learners.” In: *OpenAI blog* 1.8 (2019), p. 9.
- [91] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation.” In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.
- [92] Veronica Rotemberg et al. “A patient-centric dataset of images and metadata for identifying melanomas using clinical context.” In: *Scientific data* 8.1 (2021), p. 34.
- [93] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. “Learning representations by back-propagating errors.” In: *nature* 323.6088 (1986), pp. 533–536.
- [94] A. Saez, C. Serrano, and B. Acha. “Model-Based Classification Methods of Global Patterns in Dermoscopic Images.” In: *IEEE Transactions on Medical Imaging* 33.5 (2014), pp. 1137–1147.
- [95] Mark Sandler et al. “Mobilenetv2: Inverted residuals and linear bottlenecks.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4510–4520.



- [96] Konstantin Schürholt, Dimche Kostadinov, and Damian Borth. “Self-supervised representation learning on neural network weights for model characteristic prediction.” In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 16481–16493.
- [97] Divya Shanmugam et al. “Better aggregation in test-time augmentation.” In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 1214–1223.
- [98] Dinggang Shen, Guorong Wu, and Heung-Il Suk. “Deep learning in medical image analysis.” In: *Annual review of biomedical engineering* 19 (2017), pp. 221–248.
- [99] Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. “Text data augmentation for deep learning.” In: *Journal of big Data* 8.1 (2021), p. 101.
- [100] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition.” In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: <http://arxiv.org/abs/1409.1556>.
- [101] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition.” In: *arXiv preprint arXiv:1409.1556* (2014).
- [102] Razia Sulthana et al. “A novel end-to-end deep convolutional neural network based skin lesion classification framework.” In: *Expert Systems with Applications* 246 (2024), p. 123056.
- [103] Bo Sun et al. “Combining feature-level and decision-level fusion in a hierarchical classifier for emotion recognition in the wild.” In: *Journal on Multimodal User Interfaces* 10 (2016), pp. 125–137.
- [104] Xiaoxiao Sun et al. “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. 2016, pp. 206–222.
- [105] Christian Szegedy et al. “Rethinking the Inception Architecture for Computer Vision.” In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 2818–2826.
- [106] Christian Szegedy et al. “Rethinking the inception architecture for computer vision.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826.

- [107] Mingxing Tan and Quoc Le. “Efficientnet: Rethinking model scaling for convolutional neural networks.” In: *International Conference on Machine Learning*. PMLR. 2019, pp. 6105–6114.
- [108] Mingxing Tan and Quoc Le. “Efficientnetv2: Smaller models and faster training.” In: *International conference on machine learning*. PMLR. 2021, pp. 10096–10106.
- [109] P. Tang et al. “GP-CNN-DTEL: Global-Part CNN Model With Data-Transformed Ensemble Learning for Skin Lesion Classification.” In: *IEEE Journal of Biomedical and Health Informatics* 24.10 (2020), pp. 2870–2882.
- [110] Peng Tang et al. “FusionM4Net: A multi-stage multi-modal learning algorithm for multi-label skin lesion classification.” In: *Medical Image Analysis* 76 (2022), p. 102307.
- [111] Peng Tang et al. “GP-CNN-DTEL: Global-part CNN model with data-transformed ensemble learning for skin lesion classification.” In: *IEEE journal of biomedical and health informatics* 24.10 (2020), pp. 2870–2882.
- [112] Chenxin Tao et al. “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*. 2022, pp. 14431–14440.
- [113] Yaei Togawa, Yosuke Yamamoto, and Hiroyuki Matsue. “Comparison of images obtained using four dermoscope imaging devices: An observational study.” In: *JEADV Clinical Practice* 2.4 (2023), pp. 888–892.
- [114] Tomasz Trzcinski. “Multimodal social media video classification with deep neural networks.” In: *Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2018*. Vol. 10808. International Society for Optics and Photonics. 2018, 108082U.
- [115] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. “The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions.” In: *Scientific data* 5.1 (2018), pp. 1–9.
- [116] Philipp Tschandl et al. “Human–computer collaboration for skin cancer recognition.” In: *Nature Medicine* 26.8 (2020), pp. 1229–1234.
- [117] ME Vestergaard et al. “Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting.” In: *British Journal of Dermatology* 159.3 (2008), pp. 669–676.

- [118] Hongyi Wang et al. “Adaptive decomposition and shared weight volumetric transformer blocks for efficient patch-free 3d medical image segmentation.” In: *IEEE Journal of Biomedical and Health Informatics* (2023).
- [119] Yan Wang et al. “Adversarial multimodal fusion with attention mechanism for skin lesion classification using clinical and dermoscopic images.” In: *Medical Image Analysis* 81 (2022), p. 102535.
- [120] Yikai Wang et al. “Learning deep multimodal feature representation with asymmetric multi-layer fusion.” In: *Proceedings of the 28th ACM International Conference on Multimedia*. 2020, pp. 3902–3910.
- [121] Maxine E Whitton et al. “Interventions for vitiligo.” In: *Cochrane Database of Systematic Reviews* 2 (2015).
- [122] Hui Wu et al. “Asymmetric Feature Fusion for Image Retrieval.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 11082–11092.
- [123] Guang Yang, Suhuai Luo, and Peter Greer. “A Novel Vision Transformer Model for Skin Cancer Classification.” In: *Neural Processing Letters* (2023), pp. 1–17.
- [124] Hao Yang et al. “Asymmetric 3d convolutional neural networks for action recognition.” In: *Pattern recognition* 85 (2019), pp. 1–12.
- [125] Jiancheng Yang et al. “Asymmetric 3d context fusion for universal lesion detection.” In: *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*. Springer. 2021, pp. 571–580.
- [126] Jufeng Yang et al. “Clinical skin lesion diagnosis using representations inspired by dermatologist criteria.” In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 1258–1266.
- [127] Jufeng Yang et al. “Self-paced balance learning for clinical skin disease recognition.” In: *IEEE transactions on neural networks and learning systems* 31.8 (2019), pp. 2832–2846.
- [128] Peng Yao et al. “Single model deep learning on imbalanced small datasets for skin lesion classification.” In: *IEEE transactions on medical imaging* 41.5 (2021), pp. 1242–1254.
- [129] Jordan Yap, William Yolland, and Philipp Tschandl. “Multimodal skin lesion classification using deep learning.” In: *Experimental dermatology* 27.11 (2018), pp. 1261–1267.

- [130] Fisher Yu and Vladlen Koltun. “Multi-scale context aggregation by dilated convolutions.” In: *arXiv preprint arXiv:1511.07122* (2015).
- [131] Lequan Yu et al. “Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Networks.” In: *IEEE Transactions on Medical Imaging* 36.4 (2017), pp. 994–1004.
- [132] Zhen Yu et al. “Melanoma Recognition in Dermoscopy Images via Aggregated Deep Convolutional Features.” In: *IEEE Transactions on Biomedical Engineering* 66.4 (2019), pp. 1006–1016.
- [133] Heng Zhang, Vishal M. Patel, and Rama Chellappa. “Hierarchical Multimodal Metric Learning for Multimodal Classification.” In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 2925–2933.
- [134] Jianpeng Zhang et al. “Attention Residual Learning for Skin Lesion Classification.” In: *IEEE Transactions on Medical Imaging* 38.9 (2019), pp. 2092–2103.
- [135] Ru Zhang et al. “Depth-wise separable convolutions and multi-level pooling for an efficient spatial CNN-based steganalysis.” In: *IEEE Transactions on Information Forensics and Security* 15 (2019), pp. 1138–1150.
- [136] Yilan Zhang, Fengying Xie, and Jianqi Chen. “TFormer: A throughout fusion transformer for multi-modal skin lesion diagnosis.” In: *Computers in Biology and Medicine* 157 (2023), p. 106712.
- [137] S Kevin Zhou, Hayit Greenspan, and Dinggang Shen. *Deep learning for medical image analysis*. Academic Press, 2017.
- [138] Zhen Zhu et al. “Asymmetric non-local neural networks for semantic segmentation.” In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 593–602.