Improving Skin Lesion Classification based on Fusion Multi-Learning Models

Trung-Phien Nguyen
HCMC University of Technology and Education
Ho Chi Minh city, VietNam
trungphien456@gmail.com

Van-Dung Hoang*
HCMC University of Technology and Education
Ho Chi Minh city, VietNam
Corresponding to dunghv@hcmute.edu.vn

Minh-Luan Su
HCMC University of Technology and Education
Ho Chi Minh city, VietNam
sm.luan2003@gmail.com

Thi-Nhat-Vy Nguyen
HCMC University of Technology and Education
Ho Chi Minh city, VietNam
vynguyen17033@gmail.com

Trong-Tri Bui
HCMC University of Technology and Education
Ho Chi Minh city, VietNam
buitriquangngai@gmail.com

Ba-Duy Nguyen
Binh Duong Economics and Technology University
Binh Duong, VietNam
nguyenbaduy@ktkt.edu.vn

***Abstract*—Skin cancer is a leading malignant disease with rising incidence rates, emphasizing the need for early and accurate diagnosis. This paper introduces a new fusion method for classifying multiple classes of skin lesions by combining the Vision Transformer (ViT) and Vision Permutator (ViP) models. The proposed method leverages the global attention mechanism of ViT and the spatial encoding capabilities of ViP to enhance classification performance. Additionally, various data augmentation techniques, such as random zoom, flip, shift, and range adjustments are applied to tackle the issue of class imbalance. The proposed method is evaluated and analyzed with the ISIC2019 dataset. The models have trained on ISIC2019 dataset without being pretrained on a large dataset, e.g. ImageNet. The experimental results demonstrated that the fusion models, particularly Fusion_cat and Fusion_max, achieved superior performance compared to individual ViT and ViP models. Specifically, Fusion_max attained accuracy of 80.86%, while Fusion_cat reaches in 77.96% mean recall, 76.81% mean precision, and 77.38% F1-score. These findings suggest that our proposed models can significantly enhance automated skin lesion classification, contributing to the early diagnosis of skin cancer.**

***Index Terms*—Deep learning, Medical image processing, fusion of multiple models**

I. Introduction

Skin cancer has been a common malignant disease worldwide, experiencing a notable rise in incidence rates since the late 20^{th} century [1]. In the year 2020s, skin cancer was the most common group of cancers, with over 1.5 million new cases reported, and projections indicate this number could rise to over 2.3 million in the near future. The primary cause is ultraviolet (UV) radiation, due to changes in the ozone layer and climate change at the tropospheric level. According to a report by the World Health Organization (WHO), skin cancer results from the uncontrolled proliferation of abnormal skin cells, leading to the formation of lesions, swelling, or tumors [2]. Skin cancer is mainly basal cell carcinoma (BCC) and squamous cell carcinoma (SCC), which are relatively less dangerous. Melanoma is the least common but most dangerous form of skin cancer, with the fastest increasing incidence rate among all skin cancers. Although melanoma only accounts for 4% of all skin cancer cases, it leads to 80% of skin cancer deaths [3]. However, early diagnosis of malignant melanoma (stage 1) has a relatively high 5-year survival rate of over 95%, significantly higher than the 8% to 25% for late-stage malignant melanoma. Therefore, early diagnosis's impact on improving this condition's prognosis is substantial [4], [5]. The traditional approach to dermatological diagnosis primarily relies on visual inspection of skin characteristics and subjective assessment based on experience [6], and even dermatologists are not immune to misdiagnosis due to the lack of precise, objective, and quantitative criteria. But along with the developments of science, the efforts to detect malignant tumors have also seen breakthrough developments over time. Starting from the whole-body skin examination methods without assistance, such as the ABCDE rule (asymmetry, border irregularity, color unevenness, diameter greater than 6mm) to assess suspicious areas [7], to the assisted examination methods using optical devices (dermoscopy) since 1987 which increased the sensitivity from 76% to 92% with a fixed specificity of 80% compared to naked-eye examination [8], to the development of devices like MoleMax and MelaFind in the 1990s [9], [10], the emergence of multispectral and integrated algorithmic technologies in the 2000s with real-time imaging techniques like ultrasound, confocal microscopy, and bioimpedance since 2005 [11], [12], [13]. the above methods also cannot avoid certain limitations, including the fact that even with assisted

examination methods, the accuracy still heavily depends on the training expertise of the diagnosing person [14], the persistent challenges of inaccuracies influenced by human biases, the high costs of the devices, and the inconsistent compliance of patients [10].

Recently, Deep Learning (DL) models have been applied for the classification of skin lesions [17]– [21]. Due to their ability to learn and extract features for disease classification based on images, DL models provide more accurate diagnostic results than the traditional diagnostic techniques mentioned above. Some studies have shown that DL models can classify malignant or benign skin lesions on par with or outperform dermatologists [15]. However, there are still some issues with DL, such as data limitations and class imbalance, which lead to bias between majority and minority classes in classification results. Therefore, in this paper, we focus on addressing the data imbalance issue in the ISIC2019 dataset and building a model that combines state-of-the-art DL architectures to improve the accuracy of skin lesion classification.

II. Related Works

Nowadays, image-based artificial intelligence (AI) technology has become a promising method for the diagnosis of skin diseases. The most commonly used AI algorithms include machine learning (ML) and deep learning (DL). Initially, ML models like Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Naive Bayes, and Adaboost gained attention in the field of skin disease classification, with Adaboost emerging as the optimal classifier in distinguishing between benign and malignant lesions [6]. In the years 2015, an innovative method for automated skin disease classification was developed, integrating KNN and SVM for classifying skin conditions. The KNN model alone achieved a classification accuracy of 76%, and the SVM model achieved 78%. When these two classifiers were combined, the accuracy increased to approximately 85% [16]. This novel method highlights the groundbreaking effort to merge classifiers, demonstrating the improved performance of the integrated model.

The use of deep learning models, especially Convolutional Neural Networks (CNNs), for skin cancer detection and diagnosis has garnered significant attention in recent years. The study [22] introduced a new type of neural network utilizing a modified pre-trained CNN that can better classify skin lesions by looking at images in different resolutions simultaneously. The paper [23] proposed an automated method for melanoma recognition in dermoscopic images using very deep residual networks for segmentation and CNNs for classification. A DCNN framework that incorporates a regularized Fisher discriminant analysis to detect skin lesions effectively in [24]. The study [25] used a multi-ResNet ensemble approach to analyze skin lesions, where each ResNet was subjected to different pre-processing techniques while retaining consistent labeling across the models. The paper [26] presents a novel automated method for lesion border detection in clinical images using CNN. This approach eliminates the need for pre-processing steps such as hair removal and illumination correction, demonstrating high accuracy in detecting lesion borders for melanoma detection. A hybrid deep neural network framework for skin lesion classification, combining pre-trained CNN with ensemble learning to enhance diagnostic accuracy was proposed in [27]. The research [28] introduced a method for skin lesion classification using an ensemble of DCNN to enhance the accuracy of individual networks when trained on a limited dataset. Experimental results demonstrate that creating an ensemble of different CNNs is an effective approach, as each applied fusion strategy outperforms the individual networks in classification accuracy.

Moreover, ViT [35], an attention-based model, is an advanced method in the field of deep learning that applies the transformer architecture to image processing. Instead of using traditional convolutional neural networks (CNNs), ViT divides the image into small patches and processes them like words in a sequence, similar to how transformer models handle natural language [29]. ViT has proven its superior capabilities in various computer vision tasks, including image classification and object detection [30]. The strength of ViT lies in its ability to learn global features from images, helping to improve the accuracy of detecting important characteristics. Rather than relying on the local features captured by CNNs, ViT processes the image as a sequence of patches, allowing it to learn long-range dependencies and global representations. This mechanism enables ViT to outperform CNN-based models in many visual tasks, as it can better capture the holistic and contextual information in the images. Recently, the study [31] introduced a novel approach by substituting the attention module with spatial MLPs for token mixing, demonstrating that this MLP-like model can achieve competitive results on image classification benchmarks. Further research [32], [33], [36] has enhanced these MLP-like models through efficient training techniques and specialized MLP module designs, progressively reducing the performance gap with ViT and challenging the dominance of attention mechanisms for token mixing.

The ISIC2019 Challenge [34] is one of the most significant competitions in the field of skin lesion classification, attracting the participation of top research groups worldwide. The TOP 1-3 teams in this challenge employed ensemble methods, combining multiple DL models to achieve the highest accuracy. These leading teams have demonstrated that combining multiple DL models can significantly enhance the effectiveness of skin lesion classification. Inspired by this, in our study, we propose a novel model that combines ViT [35] and ViP [36] – a MLP-like model, aiming to leverage the advantages of both approaches: the Attention mechanism of ViT to capture global information across different regions in the image and the ViP's ability to encode spatially separated information helps maintain precise of positional information, thereby improving the accuracy of skin lesion classification. Additionally, various data augmentation methods are applied to address the problem of class imbalance in ISIC2019 dataset for improving the classified performance.

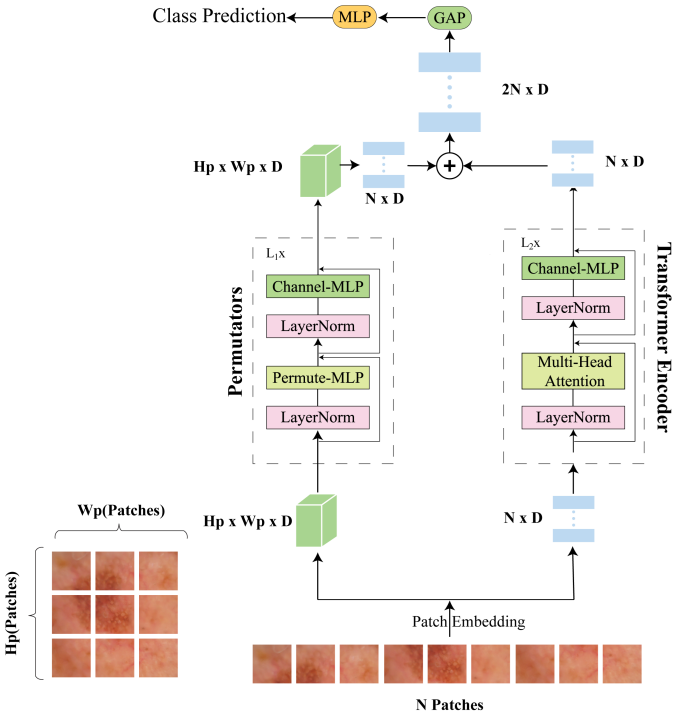

Fig. 1. Overview of the Hybrid Architecture of ViT and ViP approaches.

III. Proposed Method

This study proposes a hybrid model combining Vision Transformer [35] and Vision Permutator [36] for the image classification task by leveraging the Attention mechanism [37] to extract comprehensive information across image regions in ViT and the capability to encode spatial information independently along the height and width dimensions to maintain positional information in ViP. The input image $X \in \mathbb{R}^{H \times W \times C}$ with a spatial resolution $H \times W$ and C channels is fed into two branches: the Permutator and Transformer Encoder is solved in parallel processing. The final outputs are combined to produce the prediction result. The overall architecture of the method is demonstrated in Fig. 1.

A. Vision Transformer

Image Patching: Divide the input image $\times$ into patches of equal resolution $P \times P$, where $N = \frac{H.W}{P^2}$ is the number of patches. Next, flatten these patches into vectors $X_i \in \mathbb{R}^{P^2.C}$, for $i = 1, 2, \ldots, N$.

Patch Embedding: Map the vectors X_i above to patch embedding vectors using a linear projection method through a Dense layer. To encode the spatial information of each patch, we add the learned position embedding vectors to the patch embedding vectors as follows:

$$Z0 = [x_1 E;\ x_2 E;\ \ldots;\ x_N E] + E_{pos} \tag{1}$$

With $E \in \mathbb{R}^{(P^2 \cdot C)\ X\ D}$ denotes the linear projection matrix, $E_{pos} \in \mathbb{R}^{N\ X\ D}$ is the matrix of position embedding vectors.

Transformer Encoder: Consists of L layers, each comprising Multihead Self-Attention (MSA) and Multi-Layer Perceptron (MLP) blocks. Hence, the output of the l^{th} layer can be expressed as [35]:

$$Z'_l = MSA(LN(Z_{l-1})) + Z_{l-1}, \tag{2}$$

$$Z_l = MLP(LN(Z'_l)) + Z'_l \tag{3}$$

Where LN(.) refers to the Norm layer.

B. Vision Permutator

The ViP also divides the initial image into patches like ViT and performs the patch embedding process. However, ViP does not use positional encoding for the patch embedding vectors as this technique adversely affects the performance of the ViP model [36]. The patch embedding vectors (tokens) are then passed through a series of permutator blocks to encode spatial and channel information, see also Fig. 1.

Permute-MLP: Let $X \in \mathbb{R}^{H_p \times W_p \times D}$ denotes the tensor containing the input tokens with dimensions D, where H_p and W_p corresponds to the number of tokens along the vertical and horizontal axes of the original image, respectively. This tensor is processed in parallel through three branches to encode information along the height, width, and channel dimensions [36]. By separately encoding tokens along the spatial dimensions of the original image, ViP maintains precise positional information of the tokens.

Channel-MLP: Comprises two Dense layers with a GELU activation function in between. This module has a structure similar to the MLP module of the transformer encoder [35].

Permutator: Contains two main components: Permute-MLP and Channel-MLP. With input $X \in \mathbb{R}^{H_p \times W_p \times D}$, the operations of the permutator are defined as follows [36]:

$$Y = Permute - MLP(LN(X)) + X, \tag{4}$$

$$Z = Channel - MLP(LN(Y)) + Y \tag{5}$$

The value Z is then fed into the next permutator layer, which continues until the last layer.

C. Fusion of Transformer Encoder and Permutators

From the original input image, after performing patch embedding and processing through two parallel branches, the Transformer Encoder and Permutators, we obtain the following results:

The output of the Permutators is a 3D tensor $X \in \mathbb{R}^{H_p \times W_p \times D}$ containing tokens of size D. We then reshape X to a 2D tensor $X \in \mathbb{R}^{N \times D}$, where $N = H_p\ .\ W_p$.

The output of the Transformer Encoder is a 2D tensor $X' \in \mathbb{R}^{N \times D}$ containing tokens of size D. We concatenate the two tensors X and X' , then apply Global Average Pooling (GAP), and subsequently feed the result through the Multi-Layer Perceptron (MLP) layer to generate the classification output. The formula is described as follows:

$$Y = Concate(X, X'), \tag{6}$$

$$Z = GAP(Y), \tag{7}$$

$$ClassPrediction = MLP(Z) \tag{8}$$

IV. Experiment and Evaluation

A. Datasets and Metrics

This study uses the training dataset from the ISIC2019 competition, which contains 25,331 dermoscopic images with 8 disease categories [34], including melanoma (MEL), melanocytic nevus (NV), basal cell carcinoma (BCC), actinic keratosis (AK), benign keratosis (BKL), dermatofibroma (DF), vascular lesion (VASC), and squamous cell carcinoma (SCC).

To evaluate the performance of the model for the skin lesion classification task, we use four metrics: Accuracy (ACC), mean Recall (mREC), mean Precision (mPRE) and F1-scores. The formulas for these metrics are as follows:

$$\text{ACC} = \frac{\sum TP}{N_s} \tag{9}$$

$$\text{REC} = \frac{TP}{TP + FN} \tag{10}$$

$$\text{PRE} = \frac{TP}{TP + FP} \tag{11}$$

$$\text{F1} = \frac{2 \times mREC \times mPRE}{mREC + mPRE} \tag{12}$$

In these equations, recall and precision are calculated for each class and then averaged to obtain the mean recall and mean precision, which represent the overall performance metrics. N_s represents the total number of samples in the dataset. The values TP, FP, TN, FN correspond to the counts of true positives, false positives, true negatives, and false negatives for each class.

TABLE I
THE DISTRIBUTION OF IMAGE QUANTITIES ACROSS EACH CLASS IN THE ISIC2019 DATASET.

	MEL	NV	BCC	AK	BKL	DF	VASC	SCC	Total
Train	4,070	11,588	2,991	781	2,362	216	228	566	23,077
Test	452	1287	332	86	262	23	25	62	2,254
Total	4,522	12,875	3,323	867	2,624	239	253	628	25,331

B. Data augmentation

First, the dermatological images in the initial dataset often have different sizes, while ViP and ViT models usually require input images of a fixed size. Therefore, all dermatological images were resized to 224×224.

According to the statistics in the chart as demonstrated in Fig. 2, we can recognize a significant imbalance among the disease categories in the ISIC2019 dataset. This can lead to a major bias towards the majority class in prediction results and significantly reduce the model's performance. To address this issue, we conducted several studies on input data processing methods such as using data augmentation (AU) and applying loss functions (LFs) [38].

According to our research results, both methods aim to address the imbalance in the dataset classes. However, LFs seem to be slightly less effective due to their weighting mechanism. LFs address the issue without using data augmentation; instead, we need to adjust the parameters appropriately to make the most of the information from the important classes. This adds a significant amount of cost, whereas this study aims to optimize the parameters compared to the original methods. Moreover, in some cases, LFs may not fully solve the posed problem. Therefore, we decided to use the data augmentation (AU) method to solve the problem in the quickest and most effective way.

Specifically, we use a list w to set up lists corresponding to each disease category to adjust the number of augmented samples for each class.

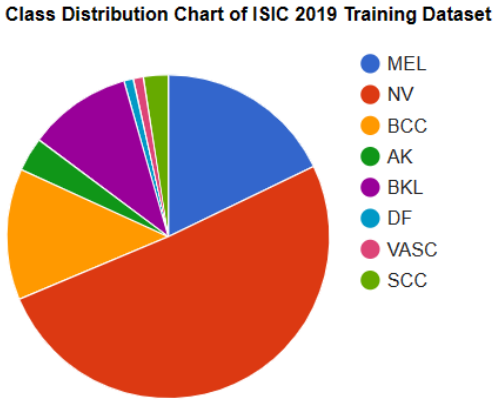

Fig. 2. The distribution of image quantities across each class in the ISIC2019 dataset.

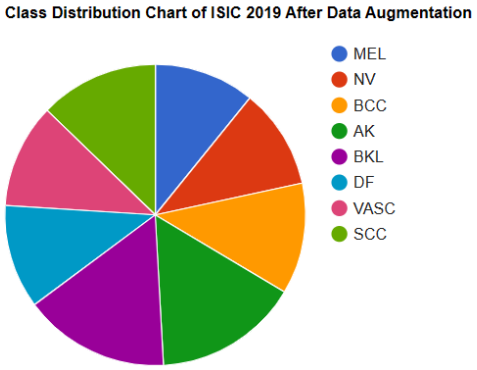

Fig. 3. The distribution of image quantities across each class in the ISIC2019 dataset after data augmentation.

TABLE II
DISTRIBUTION OF WEIGHTS FOR EACH CLASS

MEL	NV	BCC	AK	BKL	DF	VASC	SCC
1	2	2	7	2	19	18	8

Based on the weights, as shown in Table II , random data augmentation methods are applied to each image in each class

to reduce the input imbalance. Specifically, we used random zoom, flip, shift, and range methods [39]]. The output results have demonstrated the success of using AU, with the number of samples in each class becoming significantly more balanced, as shown in Table III. This balance can be seen more clearly in the distribution chart of image quantities across each class, as illustrated in Fig 3.

TABLE III
THE NUMBER OF TRAINING IMAGES ACROSS EACH CLASS BY AUGMENTATION PROCESSING.

	MEL	NV	BCC	AK	BKL	DF	VASC	SCC	Total
Original	4,345	11,588	2,991	781	2,362	216	228	566	25,331
Augmented	9,768	12,000	11,964	11,872	11,338	11,924	11,858	11,773	83,496

Since the test set of the original ISIC2019 dataset does not have ground truth labels, in this experiment, we randomly split the original ISIC2019 training set into a training set (90% of the images) and a test set (10% of the images), with a split ratio of 9:1 for each of the 8 disease types, as shown in Table I. After performing augmentation on the training set, we further split the training set into two smaller sets: a training set (80% of the images) and a validation set (20% of the images) with a split ratio of 4:1 for each disease type. The models in this paper is trained on the training set, evaluated during the training process on the validation set, and assessed for performance on real-world data using the test set.

C. Model configuration and training

The summarizing the different configurations of the ViP-ViT fusion models are demonstrated in Table IV. In this study, the "Fusion_cat" represents for the early fusion based on concatenating the outputs of each branch, meanwhile "Fusion_max" uses the maximum values of the corresponding elements in the two output tensors as representative values.

TABLE IV
CONFIGURATION DETAILS OF VIP-VIT MODEL VARIANTS.

Model		Patch size	Hidden size	#Blocks	#Parameters
Fusion_cat	ViP	14	384	18	30M
	ViT	14	384	12	7M
Fusion_max	ViP	14	384	18	30M
	ViT	14	384	12	7M

The fusion models was trained on the ISIC2019 dataset with data augmentation but without any pre-trained data to objectively evaluate the model's performance. The entire training process was conducted on a Linux server with an Intel(R) Core(TM) i9-10900X CPU@3.70GHz (20 CPUs), 128 GB RAM, and an NVIDIA Corporation TU104GL [Quadro RTX 4000] GPU. The training process utilized the AdamW [40] optimization function with default parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The models underwent training for 100 epochs, utilizing a batch size of 8, an initial learning rate of 0.0005, and a weight decay of 0.1.

D. Experimental results and analysis

Medically, skin lesions is classified into two primary categories: malignant and benign. Among them, MEL, BCC, and SCC are categorized as malignant, while the other types are considered benign. According to Fig. 5, for the Fusion_max model, The number of true positive samples for the MEL and BCC classes are 318 and 298, respectively, higher than the corresponding values of the ViP_base model (Fig. 6), which are 299 and 288, and also higher than the corresponding values in the ViT_base model (Fig. 7), which are 199 and 256. This indicates that the fusion_max model has a higher accuracy in classifying malignant classes compared to the ViP_base model and ViT_base model. Moreover, misclassifying malignant lesions as benign can lead to serious consequences as it may prevent patients from receiving timely treatment. MEL (melanoma) is a highly fatal type of malignant skin lesion that is often misclassified as NV, a benign type. In the Fusion_max model, The total number of misclassified samples between the MEL and NV classes is 189 (114 + 75), whereas for the ViP_base model, this number is 201 (124 + 77). This demonstrates that the proposed fusion model can reduce the misclassification between these two types. In summary, our proposed model makes a significant medical contribution to the problem of skin lesion classification.

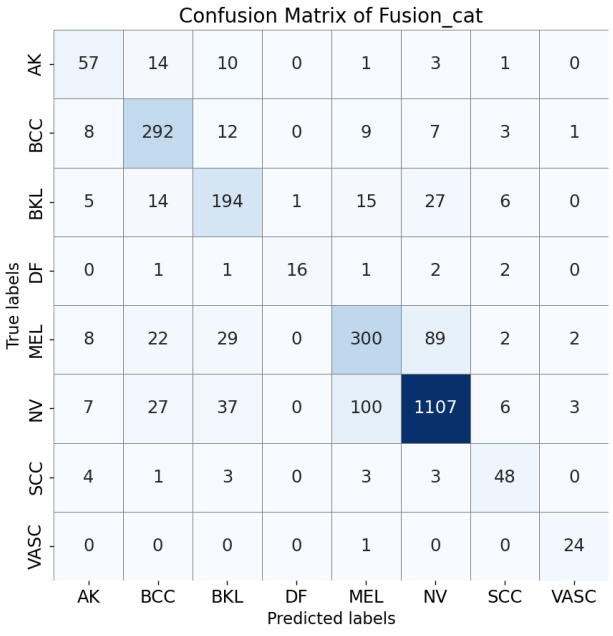

Fig. 4. Confusion matrix of classified results using Fusion_cat model

We conducted an overall performance evaluation of the models based on metrics such as Accuracy, mean recall, mean precision, and F1-score, as shown in Table V. The results indicate that the Fusion_max model achieves the highest Accuracy with 80.86%, whereas the top values for mean Recall (77.96%), mean Precision (76.81%), and F1-score (77.38%)

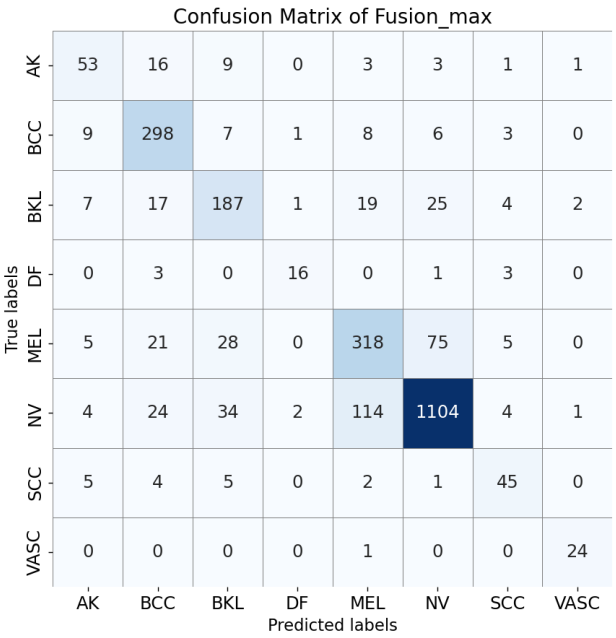

Fig. 5. Confusion matrix of classified results using Fusion_max model

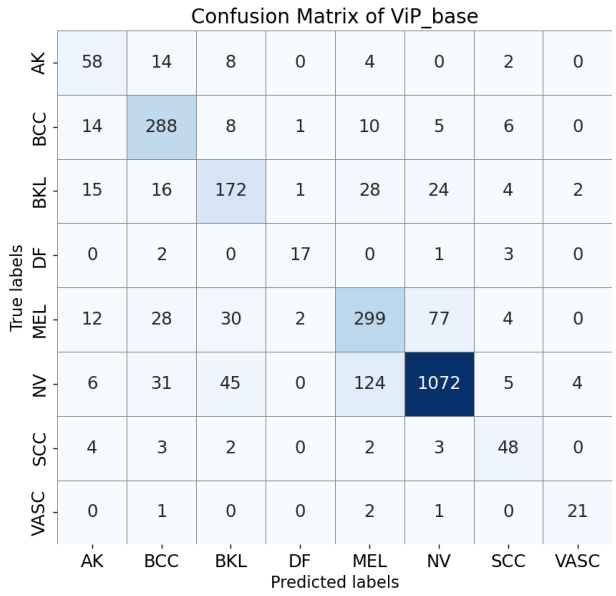

Fig. 6. Confusion matrix of classified results using ViP_base model

belong to the Fusion_cat model (see also Fig. 4). This allows us to conclude that, despite having lower accuracy, Fusion_cat exhibits better classification precision across different classes compared to Fusion_max. Therefore, Fusion_cat is the most effective model for the multi-class skin lesion classification problem among the tested models. Additionally, it is noted that the performance of the ViT_base model is relatively low, as demonstrated in Fig. 7. Because the ViT_base model was not pretrained on a large dataset such as ImageNet [41]. In the case of the model were trained with more epochs, its performance would significantly improve. However, given the relatively

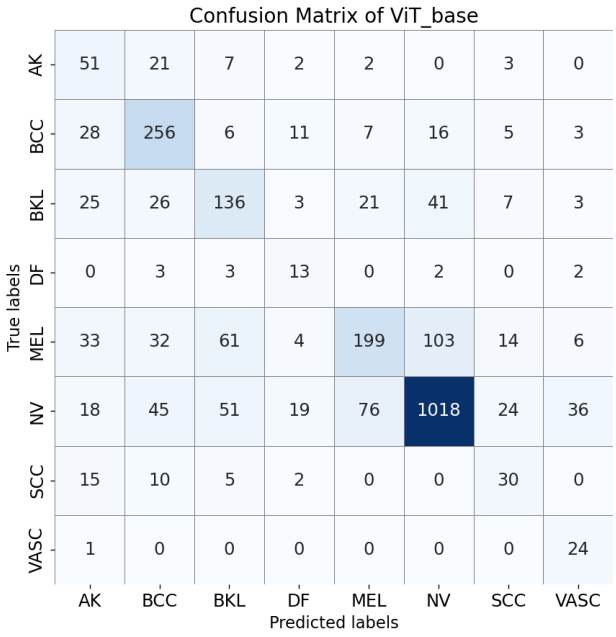

Fig. 7. Confusion matrix of classified results using ViT_base model

TABLE V
ACCURACY, MEAN RECALL, MEAN PRECISION AND F1-SCORE OF MODELS

Models	#Params	Acc(%)	mRecall(%)	mPrec(%)	F1(%)
Fusion_cat	37M	80.59	77.96	76.81	77.38
Fusion_max	37M	80.86	77.13	75.64	76.38
ViP_base	30M	78.09	75.58	71.63	73.55
ViT_base	86M	68.29	64.04	48.71	55.34

large number of parameters in ViT_base (86M), the proposed fusion models remain the optimal choice for deployment on devices with limited memory resources and computational speed while still requiring relatively high accuracy.

V. CONCLUSION

In this paper, a new approach based on fusion of multi-learning models of ViT and ViP is proposed for the task of multi-class skin lesion classification. By leveraging the strengths of both models: ViT's attention mechanism for capturing global image information and ViP's capability for preserving spatial information. We aimed to enhance the classification accuracy and robustness. To resolve the problem of data imbalance in the ISIC2019 dataset, we utilized a variety of data augmentation methods such as random zoom, flip, shift, and range adjustments. These methods significantly improved the balance of sample distribution among different disease classes. This preprocessing step was crucial in ensuring that the models trained effectively on the diverse classes without bias towards any particular class. Our experimental results demonstrated that the fusion models, particularly Fusion_cat

and Fusion_max, outperform the individual ViT and ViP models. The Fusion_cat model, despite having a slightly lower overall accuracy compared to Fusion_max, showed superior mean recall, mean precision, and F1-score, indicating better class balance and classification precision. In the future, we aim to optimize the fusion models for deployment in real-world clinical settings to ensure they are both effective and practical for use by healthcare professionals.

ACKNOWLEDGEMENTS

This work belongs to the project grant No: SV2024-194 funded by Ho Chi Minh City University of Technology and Education, Vietnam.

REFERENCES

[1] E. Chatzilakou, Y. Hu, N. Jiang, and A. K. Yetisen, "Biosensors for melanoma skin cancer diagnostics," Biosensors Bioelectronics/Biosensors Bioelectronics (Online), vol. 250, p. 116045, Apr. 2024, doi: 10.1016/j.bios.2024.116045.

[2] H. W. Rogers, M. A. Weinstock, S. R. Feldman, and B. M. Coldiron, "Incidence estimate of nonmelanoma skin cancer (Keratinocyte carcinomas) in the US population, 2012," JAMA Dermatology, vol. 151, no. 10, p. 1081, Oct. 2015, doi: 10.1001/jamadermatol.2015.1187.

[3] V. A. O. Nancy, P. Prabhavathy, M. S. Arya, and B. S. Ahamed, "Comparative study and analysis on skin cancer detection using machine learning and deep learning algorithms," Multimedia Tools and Applications, vol. 82, no. 29, pp. 45913–45957, Aug. 2023, doi: 10.1007/s11042-023-16422-6.

[4] S. Mane and S. Shinde, "A method for melanoma skin cancer detection using dermoscopy images," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Apr. 2019, doi: 10.1109/iccubea.2018.8697804.

[5] M. Phillips, J. Greenhalgh, H. Marsden, and I. Palamaras, "Detection of malignant melanoma using Artificial intelligence: An Observational study of Diagnostic Accuracy," Dermatology Practical Conceptual, p. e2020011, Dec. 2019, doi: 10.5826/dpc.1001a11.

[6] J. Zhang, F. Zhong, K. He, M. Ji, S. Li, and C. Li, "Recent Advancements and Perspectives in the diagnosis of skin diseases using machine learning and deep Learning: a review," Diagnostics, vol. 13, no. 23, p. 3506, Nov. 2023, doi: 10.3390/diagnostics13233506.

[7] B. Sreedhar, M. S. BE, and M. S. Kumar, "A comparative study of melanoma skin cancer detection in traditional and current image processing techniques," 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Oct. 2020, doi: 10.1109/i-smac49090.2020.9243501.

[8] J. Dinnes et al., "Reflectance confocal microscopy for diagnosing keratinocyte skin cancers in adults," Cochrane Library, vol. 2018, no. 12, Dec. 2018, doi: 10.1002/14651858.cd013191.

[9] D. Gutkowicz-Krusin et al., "Precision of automatic measurements of pigmented skin lesion parameters with a MelaFindTM multispectral digital dermoscope," Melanoma Research, vol. 10, no. 6, pp. 563–570, Dec. 2000, doi: 10.1097/00008390-200012000-00008.

[10] A. T. Young et al., "The role of technology in melanoma screening and diagnosis," Pigment Cell Melanoma Research, vol. 34, no. 2, pp. 288–300, Aug. 2020, doi: 10.1111/pcmr.12907.

[11] P. Aberg, I. Nicander, J. Hansson, P. Geladi, U. Holmgren, and S. Ollmar, "Skin cancer identification using Multifrequency Electrical Impedance—A potential Screening tool," IEEE Transactions on Bio-medical Engineering/IEEE Transactions on Biomedical Engineering, vol. 51, no. 12, pp. 2097–2102, Dec. 2004, doi: 10.1109/tbme.2004.836523.

[12] M. Carrara et al., "Multispectral imaging and artificial neural network: mimicking the management decision of the clinician facing pigmented skin lesions," Physics in Medicine Biology/Physics in Medicine and Biology, vol. 52, no. 9, pp. 2599–2613, Apr. 2007, doi: 10.1088/0031-9155/52/9/018.

[13] D. S. Rigel, J. Russak, and R. Friedman, "The Evolution of Melanoma Diagnosis: 25 years Beyond the ABCDs," Ca, vol. 60, no. 5, pp. 301–316, Jul. 2010, doi: 10.3322/caac.20074.

[14] A. Blum et al., "The status of dermoscopy in Germany – results of the cross-sectional Pan-Euro-Dermoscopy Study," Journal Der Deutschen Dermatologischen Gesellschaft, vol. 16, no. 2, pp. 174–181, Jan. 2018, doi: 10.1111/ddg.13431.

[15] P. Carli et al., "Pattern analysis, not simplified algorithms, is the most reliable method for teaching dermoscopy for melanoma diagnosis to residents in dermatology," British Journal of Dermatology/British Journal of Dermatology, Supplement, vol. 148, no. 5, pp. 981–984, May 2003, doi: 10.1046/j.1365-2133.2003.05023.x.

[16] R. Sumithra, M. Suhil, and D. S. Guru, "Segmentation and classification of skin lesions for disease diagnosis," Procedia Computer Science, vol. 45, pp. 76–85, Jan. 2015, doi: 10.1016/j.procs.2015.03.090.

[17] I. Papachristou and N. Bosanquet, "Improving the prevention and diagnosis of melanoma on a national scale: A comparative study of performance in the United Kingdom and Australia," Journal of Public Health Policy, vol. 41, no. 1, pp. 28–38, Sep. 2019, doi: 10.1057/s41271-019-00187-0.

[18] J. Yap, W. Yolland, and P. Tschandl, "Multimodal skin lesion classification using deep learning," Experimental Dermatology, vol. 27, no. 11, pp. 1261–1267, Sep. 2018, doi: 10.1111/exd.13777.

[19] P. Tschandl et al., "Expert-Level diagnosis of nonpigmented skin cancer by combined convolutional neural networks," JAMA Dermatology, vol. 155, no. 1, p. 58, Jan. 2019, doi: 10.1001/jamadermatol.2018.4378.

[20] T. J. Brinker et al., "Comparing artificial intelligence algorithms to 157 German dermatologists: the melanoma classification benchmark," European Journal of Cancer, vol. 111, pp. 30–37, Apr. 2019, doi: 10.1016/j.ejca.2018.12.016.

[21] H. A. Haenssle et al., "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," Annals of Oncology, vol. 29, no. 8, pp. 1836–1842, Aug. 2018, doi: 10.1093/annonc/mdy166.

[22] J. Kawahara and G. Hamarneh, "Multi-resolution-Tract CNN with Hybrid Pretrained and Skin-Lesion Trained Layers," in Lecture notes in computer science, 2016, pp. 164–171. doi: 10.1007/978-3-319-47157-0_20.

[23] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," IEEE Transactions on Medical Imaging, vol. 36, no. 4, pp. 994–1004, Apr. 2017, doi: 10.1109/tmi.2016.2642839.

[24] N. N. Sultana, B. Mandal, and N. B. Puhan, "Deep residual network with regularised fisher framework for detection of melanoma," IET Computer Vision, vol. 12, no. 8, pp. 1096–1104, Jul. 2018, doi: 10.1049/iet-cvi.2018.5238.

[25] S. Guo and Z. Yang, "Multi-Channel-ResNet: An integration framework towards skin lesion analysis," Informatics in Medicine Unlocked, vol. 12, pp. 67–74, Jan. 2018, doi: 10.1016/j.imu.2018.06.006.

[26] P. Sabouri and H. GholamHosseini, "Lesion border detection using deep learning," 2016 IEEE Congress on Evolutionary Computation (CEC), pp. 1416–1421, Jul. 2016, doi: 10.1109/cec.2016.7743955.

[27] A. Mahbod, G. Schaefer, C. Wang, R. Ecker, and I. Ellinge, "Skin Lesion Classification Using Hybrid Deep Neural Networks," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1229–1233, May 2019, doi: 10.1109/icassp.2019.8683352.

[28] B. Harangi, "Skin lesion classification with ensembles of deep convolutional neural networks," Journal of Biomedical Informatics, vol. 86, pp. 25–32, Oct. 2018, doi: 10.1016/j.jbi.2018.08.006.

[29] A. Dosovitskiy and T. Brox, "Inverting Visual Representations with Convolutional Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 9, pp. 2365–2378, Jun. 2016, doi: 10.1109/cvpr.2016.522.

[30] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers distillation through attention," PMLR, Jul. 01, 2021. https://proceedings.mlr.press/v139/touvron21a

[31] I. Tolstikhin et al., "MLP-Mixer: an all-MLP architecture for vision," arXiv (Cornell University), vol. 34, Dec. 2021, [Online]. Available: https://arxiv.org/pdf/2105.01601

[32] H. Liu, Z. Dai, D. R. So, and Q. V. Le, "Pay attention to MLPs," arXiv (Cornell University), vol. 34, Dec. 2021, [Online]. Available: https://arxiv.org/abs/2105.08050

[33] H. Touvron et al., "RESMLP: Feedforward Networks for Image Classification with Data-Efficient Training," IEEE Transactions on Pat-

tern Analysis and Machine Intelligence, pp. 1–9, Jan. 2022, doi: 10.1109/tpami.2022.3206148

[34] Y. Li and L. Shen, “Skin Lesion Analysis towards Melanoma Detection Using Deep Learning Network,” Sensors, vol. 18, no. 2, p. 556, Feb. 2018, doi: 10.3390/s18020556.

[35] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” arXiv.org, Oct. 22, 2020. https://arxiv.org/abs/2010.11929.

[36] Q. Hou, Z. Jiang, L. Yuan, M.-M. Cheng, S. Yan, and J. Feng, “Vision Permutator: a permutable MLP-Like architecture for visual recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 1, pp. 1328–1334, Jan. 2023, doi: 10.1109/tpami.2022.3145427.

[37] A. Vaswani et al., “Attention is All you Need,” 2017. https://proceedings.neurips.cc/paper/7181-attention-is-all

[38] T.-A. Pham and V.-D. Hoang, “Combination of deep learning and ambiguity rejection for improving Image-Based Disease diagnosis,” in Lecture notes in computer science, 2023, pp. 147–160. doi: 10.1007/978-981-99-5834-4_12.

[39] A. Mumuni and F. Mumuni, “Data augmentation: A comprehensive survey of modern approaches,” Array, vol. 16, p. 100258, Dec. 2022, doi: 10.1016/j.array.2022.100258.

[40] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” arXiv.org, Nov. 14, 2017. https://arxiv.org/abs/1711.05101

[41] J. Deng, W. Dong, R. Socher, L.-J. Li, N. K. Li, and N. L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” 2009 IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2009, doi: 10.1109/cvpr.2009.5206848.